UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Químicas e Ingeniería

TÍTULO: A Novel Network Science and Similarity-Searching-Based Approach for Discovering Potential Antiviral Peptides

Trabajo de integración curricular presentado como requisito para la obtención del título de Química

Autor:

Daniela de Llano García

Tutor:

Hortensia Rodriguez, Ph.D.

Co-Tutor:

Yovani Marrero-Ponce, Ph.D.

Urcuquí, Diciembre - 2023

Autoría

Yo, **Daniela de Llano García**, con cédula de identidad 095947954-4, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el autor del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Diciembre - 2023.

Daniela de Llano García

CI: 095947954-4

Autorización de publicación

Yo, **Daniela de Llano García**, con cédula de identidad 095947954-4, cedo a la Universidad de Investigación de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación

Urcuquí, Diciembre - 2023.

Firmado electrónicamente por:
DANIELA DE LLANO GARCIA

Daniela de Llano García

CI: 095947954-4

Dedication

Dedicated to my determined sister, who insisted on her presence in this dedication, and to my incredible parents, my ultimate role models, who have supported and cared for me all these years.

Daniela de Llano García

Acknowledgment

First and foremost, I extend my deepest gratitude to my parents, Alen and Viviana, who have been my role models since I was a child, and have always lovingly nurtured my passion for science and curiosity.

I am immensely grateful to my advisors, Dr. Hortensia Rodríguez and Dr. Yovani Marrero-Ponce, who have been the best possible mentors. Their unmeasurable patience, expertise, and dedication have been instrumental in shaping my journey and teaching me so much. Their guidance has not only enriched my knowledge but also shown me the rewarding excitement of research.

To my closest friends, who have made my 5-year journey all the more enjoyable. Special thanks to Danny, Ana and Maria Fernanda, from whom I learn every day how to become a better person and a more accomplished chemist. Your friendship and inspiration have been invaluable to me.

Lastly, I give all my gratitude to my incredible boyfriend, Argenis, my library partner and a constant source of support. He has been there to lift me up when I felt I couldn't keep going. Without his encouragement and assistance, I wouldn't have been able to complete this.

Daniela de Llano García

Resumen

Los péptidos antivirales (AVP, por sus siglas en inglés) poseen un gran potencial como fármacos contra las infecciones virales. Sin embargo, la cantidad de información disponible supera la capacidad de interpretación de los investigadores. Para abordar este desafío, se utilizó la minería interactiva de datos y las facilidades de las redes de espacio proximal a través del software StarPep para explorar el espacio químico de los AVPs. Adicionalmente se utilizó el algoritmo de clustering de Louvain para crear un perfil basado en las comunidades obtenidas, revelando así características biológicas, patrones y diferentes relaciones entre los péptidos. Esta exploración fue extendida a través de las redes de metadato (MN) que aportó importante información sobre "bases de datos", "función" ,"origen" y "objetivo". Este análisis permitió detectar nuevas interconexiones, enriqueciendo la comprensión de los AVPs y sus características. Para crear representaciones simplificadas del espacio químico se realizó un proceso de extracción de scaffold que resultó en cuatro subconjuntos definidos que conservan las características centrales mientras simplifican la complejidad de la red. Como resultado, reportamos 33 potenciales motivos antivirales, de los cuales 23 son completamente novedosos para el campo. Además se desarrollaron cinco Modelos de Búsqueda por Similitud Múltiple que fueron comparados y superaron 14 diferentes predictores disponibles en la literatura. Sobre estos hallazgos mencionados se encontraron 46 potenciales secuencias antivirales derivadas de diferentes bases de datos que juntas contenían más de 100,000 secuencias. Este trabajo no solo proporciona información valiosa sobre las características de los AVP sino que también sienta las bases para el desarrollo de péptidos con uso terapéutico.

Palabras Clave:

Péptidos Antivirales, Espacio Químico,Redes de Espacio Proximal, Motivos, Búsqueda por Similitud Multi-Referencia, Clivado Virtual

Abstract

Antiviral peptides (AVPs) hold substantial promise as therapeutic agents against viral infections. However, the sheer volume of available data surpasses researchers' capacity for interpretation. To address this challenge, interactive data mining and the Half-Space Proximal Network (HSPN) technique within the StarPep toolbox were utilized to address this challenge and explore the chemical space of AVPs. Louvain Clustering was employed to conduct community-based chemical profiling, revealing intricate biological patterns and relationships among peptides. Exploration extended to Metadata Networks (MNs), shedding light on the broader AVP landscape with attributes like "database," "function," "origin," and "target." This exposed interconnections and associations, enriching the understanding of AVPs and their attributes. Scaffold extraction was applied to streamline the representation of the AVP chemical space, yielding four well-defined subsets that retained core characteristics while simplifying network complexity. Moreover, the study identified 33 potential antiviral motifs via an alignment-free de novo approach, including 23 entirely novel motifs. Furthermore, five Multi-Query Similarity Search Models (MQSSMs) were developed, outperforming several state-of-the-art predictors. Building upon these findings, the research yielded 46 potential Antiviral Sequences derived from three diverse databases, encompassing over 100,000 sequences. This work provides valuable insights into AVP characteristics and lays the foundation for novel antiviral therapies rooted in their distinctive chemical properties and interactions.

Keywords:
Antiviral peptide, Chemical space, Half-Space Proximal Network, Interactive mining, StarPep toolbox, Motif Discovery, Muli-Query Similarity Search, Virtual Cleavage

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Viruses encompass an extensive group of pathogens responsible for numerous critical and infectious medical events throughout human history. From smallpox to the most recent SARS-CoV-2 global pandemic, viral diseases have been a focal point for scientific, agricultural, and medical research [1]. One of the most remarkable abilities exhibited by certain viruses is their capacity to adapt to new hosts and environments through mutations in relatively short periods [2]. This characteristic leads to a constant emergence of viral diseases worldwide, necessitating the development of therapeutics to address this threat [3].

Given the perpetual risk of new infectious viral diseases, antiviral therapeutics development has been an enduring scientific endeavor. What is more, viruses have a broad range of action mechanisms for replication and infection. Thus, it is highly complex to design therapeutics that target various viruses, and each pathogen requires dedicated investment and effort [4]. Antiviral drugs and vaccines are the most employed strategies to combat viral infections. Historically, vaccines have played a central role in containing viral outbreaks, but they often require substantial resources and time and are not always effective [5]. Conversely, antiviral drugs focus on treating infections once they have already commenced. For antiviral drugs to be effective, they must exhibit both safety and potency. Over the past few decades, numerous antiviral agents have been designed to target viral proteins or host factors [6].

Antiviral drugs can be categorized as either small molecules or peptide-based molecules. Pharmaceutical companies have traditionally favored the development of small molecules due to their relatively more straightforward development process than peptides [7]. However, in recent years, peptides as pharmaceuticals have regained attention as technological advancements have addressed their main drawbacks [8]. Peptides, defined as polypeptides consisting of up to 50 connected amino acids (AAs), exhibit a wide range of biological activities and play critical roles in human physiology. As therapeutics, peptides offer a closer resemblance to biological entities, making them considered safer, less toxic, and highly effective while also possibly scaling up the production from mg to kg levels [9]. Nonetheless, they face challenges such as biological instability and poor membrane permeability and stability [10].

Despite these limitations, peptides have significantly impacted the modern pharmaceutical industry, and contributed to advancements in both chemical and biological sciences [11]. Some strategies for overcoming these setbacks include modifying the molecules and their delivery, stability, and application in preclinical stages. For example, D-peptides, resistant to natural proteases, have longer half-lives and can be absorbed orally, making them more suitable for therapeutic use than L-peptides. [12].

As of 2022, over 60 peptide drugs have been approved for commercial use in the United States, Europe, and Japan [13]. Additionally, more than 400 peptides are currently undergoing clinical trials, with 150 active clinical development and 260 completed human clinical trials [13]. Consequently, developing antiviral peptides (AVPs) as therapeutics assumes a crucial role in the fight against viral diseases. Notable examples of AVPs target various viruses including HIV [14], SARS-CoV-2 [15], Influenza [16], Herpes Simplex [17], Dengue [18], Tobacco Mosaic Virus [19], HSV [20] and Zika virus [21].

1.2 Problem statement

Undoubtedly, the quantity of information amassed in databases concerning AVPs has grown exponentially with rising interest, making it challenging to analyze each entry individually. This vast and ever-expanding dataset can overwhelm researchers' capacity for interpretation [22]. Experimentally scrutinizing tens of thousands of sequences is both

resource-intensive and time-consuming. Consequently, traditional approaches rely on manual scrutiny and experimental validation, which becomes increasingly impractical given the surge in available information.

To address this challenge, scientists have developed techniques to bring order to this wealth of information, and computational tools and methodologies have emerged as a solution. Some of these tools, rooted in data mining and network analysis, offer hope in this complex landscape. They not only expedite the research process but also enhance accuracy and efficiency. These methodologies provide researchers with a robust starting point for navigating the chemical space of biologically active peptides, streamlining their focus on the most promising candidates.

Moreover, these computational approaches provide a comprehensive overview of the available data, empowering researchers to make informed decisions. This paves the way for targeted and effective research. In the pursuit of developing peptide therapeutics, these tools are invaluable. They bridge the gap between the vast amount of data available and actionable insights, ensuring that the journey from sequence discovery to therapeutic development remains efficient, effective, and well-informed.

1.3 Objectives

1.3.1 General Objective

To comprehensively discover and understand the potential of Antiviral Peptides (AVPs) using advanced computational methodologies, including interactive mining and network science

1.3.2 Specific Objectives

- To thoroughly explore the chemical space of AVPs found in StarPepDB
- To perform a community analysis using biological and chemical information
- To conduct an advanced similarity search using the MQSS method to identify potential AVPs.

- Design and refine a representative AVPs model from starPepDB.
- To evaluate and refine models through various calibration stages
- To discover new motifs with potential antiviral activity

Chapter 2

Theoretical Framework

2.1 Chemoinformatics and Chemical Space

The enormity of the ever-expanding data overwhelms researchers' interpretation capabilities. In response to this challenge, scientists have developed techniques to bring coherence to this wealth of information [22]. The exponential growth of data stored in public databases has led to the concept of "chemical space," akin to the vast expanse of the cosmic universe filled with compounds. In this context, two important concepts are similarity and diversity[23]. However, Medina-Franco et al. have identified at least eight definitions for "chemical space" [24]. The systematic study and exploration of chemical space are commonly called "chemoinformatics"[25].

Chemoinformatics focuses on manipulating information about chemical structures using informatic methods. Since its proper definition in 2006, it has played a pivotal role in analyzing and mining chemical information, contributing to a better understanding of structure-property relations.

Chemical space and drug discovery are closely intertwined. This information has laid the foundation for many biologically and medicinally relevant structures[26]. The concept of chemogenomics precisely links the prediction and validation of the intersection between chemical and biological space [23]. The exploration of chemical space serves two main purposes. Firstly, it allows for the comparison of compound datasets from different sources, particularly libraries. Secondly, it involves the classification of bioactive compounds, illu-

minating the "biological active space" based on the principle that similar compounds share similar activity[27].

2.2 Graph-based Interactive Mining

Analyzing the chemical space involves dealing with high-complexity multivariate data. Proper visualization techniques have been developed to simplify this complexity, enabling dimensionality reduction and facilitating human brain analysis [25]. Dimensionality reduction is crucial for summarizing information while preserving its essence. As a result, various techniques for visualizing structure-activity relationships have emerged. Chemical space visualizations require a set of compounds and a set of molecular descriptors. In its simplest form, spaces are generated by plotting the coordinates of a pure descriptor. In more advanced scenarios, spaces are created based on the descriptor values [28]. Over the last two decades, different approaches have been developed, typically consisting of two major parts: clustering or organizing the chemical structure information, and visualization to project activity data onto complex information [29].

Graph-based representation is one of the popular approaches for visualizing chemical space. In this representation, each compound becomes a part of a grid of nodes. Constructing similarity-based graphs offers a visual depiction of the chemical space and is a powerful tool for extracting information, enabling intricate analyses of connectivity. These networks provide an intuitive portrayal of the chemical space, displaying a distinct structure delineating the distances between different structures [30].

In graph theory, graphs are mathematical representations between multiple objects, with each object represented as a node, and the relationships between nodes are represented as edges Each molecule is considered as its entity, and (dis)similarity metrics are used to compare them with other nodes on the grid [28], [31]. The following mathematical definition is presented to provide a more detailed description of graphs [32]:

Definition 1 (Graph). *A labeled Graph is 4-tuple, G=(V,E,L,l)*

V is a set of vertices

$E \subseteq V \times V$ *is a set of edges*

L is a set of Labels

$l : V \cup E \rightarrow L$, is a function assigning labels to the vertices and edges

Advances in network science have paved the way for novel applications, exploring methods to study global and local patterns and structures. Utilizing networks (graphs) offers three major advantages. Firstly, networks provide a more natural representation of the chemical space and its discrete structure. Secondly, they offer a concise framework for statistical analysis. Lastly, networks excel in managing large volumes of information with diverse features [30]. However, the use of networks in chemistry is a field that requires further exploration [33] Moreover, despite their high importance, systematic studies for peptides chemical space exploration and diversity are still lacking [34].

While visual representation and exploration of the chemical space are key factors in cheminformatics, the ultimate goal is not solely to extract information from the representation but to generate knowledge. This knowledge includes discovering patterns and establishing coherence and meaning within the information itself [35]. Thus, the interaction between the researcher and the information representation becomes crucial, as the human mind is the key in drawing conclusions beyond simple description. This process of turning information into knowledge is often referred to as "data mining" [36, 37]. Graph-based methods play a significant role in the realm of interactive advanced data mining and pattern discovery [36, 22].

2.2.1 Important Notions on Network Science

Chemical similarity measures described in the literature can be calculated using various methods, including: (a) Molecular graphs, (b) Descriptor vectors, (c) Molecular fields, (d) Kernels,(e) unsupervised modeling, and (f) Supervised modeling studies. Among these, similarity measures based on fixed-sized descriptor vectors are the most popular. They involve different types of distances, such as Euclidean, Manhattan, Mahalanobis, and Minkowski, to measure molecular dissimilarity [38]. Of these, the Euclidean distance is the most commonly used (dis)similarity metric. This is primarily due to its widespread application in calculating distances in physical spaces, making it more understandable and manageable for human interpreters.

Euclidean (dis) similarity metric: The techniques used for measuring distance are particularly important in data mining. The most widely known distance is the Euclidean, which works well with compact clusters, is easy to compute, and is sensitive to outliers [39].

$$d_{euc} = [\sum_{i=1}^{n}(x_i - y_i)^2]^{\frac{1}{2}} \tag{2.1}$$

Bipartite Graph: These kinds of graphs have different and independent sets of vertices (V1 and V2), and every edge of the graph connects one node of V1 to a node of V2 [40]. Examples of these kinds of graph are the metadata networks introduced in later sections.

How a network is connected is known as network topology, and it is directly influenced by the threshold value selected. Different topologies reveal distinct activity clusters within the entire space. Consequently, topology is a crucial parameter to consider when characterizing a network at both global and local levels [30]. The characterization of topology is often achieved through specific network properties, including node degree, clustering coefficient, shortest path length, community structures, and network density [33]

Degree: The degree of a node in a graph is the number of edges connected to it [41]

Path: In network science, a path is defined as any sequence of vertices such that every consecutive pair of nodes in the sequence is connected by one edge. The shortest path is a connection between two nodes in a manner that no other possible path is shorter [41].

Graph density: This property reflects the ratio between the existing edges and the maximum of edges that could exist in said graph. Said maximum, where n is the number of nodes , can be calculated as follows [41]

$$\binom{n}{x} = \frac{1}{2}n(n-1) \tag{2.2}$$

Modularity: As defined by Newman, modularity refers to the difference between the number of edges within groups in a complex network and the expected number of edges in a random network with similar characteristics. The value of modularity indicates the existence or absence of community structure within the network. Maximizing modularity suggests the presence of distinct communities within the graph [42].

The method of optimal modularity can be explained as follows [43]. Suppose a network with n number of nodes that is up to be divided into two subgroups for simplicity. Assign 1 to s_i if the node is found in group 1 and -1 if it's assigned to group 2. Additionally, let the number of edges in the network be A_{ij} which is also referred to as the adjacent matrix that can take values of 1 or 0. Also, the expected number of edges if i and j are placed at random is $\frac{k_i k_j}{2m}$: where k_i and k_j are the degrees of each node and $m = \sum_i k_i$ (total number of edges in the network. Hence, the modularity (Q) is given by the sum, of $A_{ij} - k_i k_j/2m$ over all pair of vertices placed in the same group. Stating that the quantity $\frac{1}{2}(s_i s_j + 1)$ is 1 if i and j are in different groups and 0 if not. Summarizing all this, modularity is expressed as :

$$Q = \frac{1}{4m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})(s_i s_j + 1) = \frac{1}{4m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})s_i s_j \tag{2.3}$$

Average Clustering Coefficient (ACC): The clustering coefficient (CC) is measured for each node, in a way that relates the number of edges that connect the node to its respective neighbors and the total possible number of neighbors that could be connected. This total possible number can be defined as when a node i, with j_i neighbors; in a way that every neighbor is also connected to the other neighbors. The CC gives information about a singular node, whereas ACC gives a more general idea of the topology of the network and the "small world" effect [44, 45]. "Small world" networks are highly clustered and have small characteristic path lengths. Consequently, these networks have specialized nodes or regions [46].

The mathematical representation goes as follows [47]. First, the clustering coefficient $CC(i) \in \mathbb{Q}\ [0,1]$ of a node i

$$CC(i) = \left\{ \begin{array}{ll} 0 & deg(i) \in 0,1 \\ \frac{2|N|}{deg(i)(deg(i)-1)} & deg(i) > 1 \end{array} \right\} \tag{2.4}$$

The clustering coefficient takes only values for one node, contrary to that, the average clustering coefficient $ACC \in \mathbb{R}[0,1]$ is a global parameter about the general topology of a graph. Let $X \subseteq V | \forall x \in X : deg(x) > 0$, denote the subset of such nodes [47].

$$.ACC(G) \left\{ \begin{array}{ll} 0 & |X| = 0 \\ \frac{1}{|X|} \sum_{i=1}^{|X|} CC(x_i) & |X| > 0 \end{array} \right\} \tag{2.5}$$

Clustering is the process of classifying objects into different groups or subsets based on shared critical traits. It is a common technique used in various statistical analyses, including machine learning, data mining, pattern recognition, bioinformatics, and chemoinformatics [48]. Community detection is essential because it helps identify clusters and their boundaries, enabling the classification of nodes in a network. Nodes with a central position within a specific community may play a vital role in controlling and stabilizing the group, while more external nodes may mediate exchanges and communication between communities [49]. Modularity has been employed as a measure to assess the quality of partitions in a graph, but its significance lies more importantly in its function as a means of optimization [50]. When clustering is done using modularity optimization, it ensures that each cluster consists of a connected subgraph, guaranteeing the validity and cohesiveness of the clusters [51].

Louvain Clustering Algorithm: This algorithm is quite simple and optimizes the modularity quality function; this process is carried out in two phases: (1) local moving of nodes; and (2) aggregation of the network [52]. The algorithm starts by assigning a different community to each of the nodes. Then for each node i, each of the neighbours j is considered, and the gain in modularity is evaluated if the node i is removed from its community and placed in j's community instead. Then, the node i is left in the community in which the positive gain is the highest; if this is not possible the node i stays in its

community. This process is recursive until there's a local maximum in modularity, ending the first phase. The second phase builds a new network using the priorly designated communities as nodes by establishing the weight of the links as the sum of the weight of the links within the two communities [50].

The concept of *centrality* plays a crucial role in understanding the structural attributes of networks, and it is closely related to many other significant group properties and processes [53]. The most commonly used centrality measures are degree, betweenness, closeness, and eigenvector centrality [54]. These measures quantify a node's ability to influence or be influenced by other nodes based on the network's topology, thus identifying important and central nodes often referred to as hubs [55]. The ranking of a node with a specific centrality measure depends on the dynamic process assumed to be taking place. Some centralities focus on shortest-path behaviors, while others prioritize interactions within the local community. As a result, there are over 200 centrality measures reported to date, reflecting the diverse and dynamic variations in network analysis [56].

Community Hub-Bridge Centrality (HB): a metric that assumes each node can function as a bridge or a hub. Its purpose is to identify nodes that have balanced connections within their community and between communities. This measure assigns weight to the intra-community link based on the size of the community, and weight to the inter-community link based on the neighboring communities [57][58]. To compute the centrality measure, we define the internal strength of each node i as k_i^{int} and the external strength as k_i^{ext}. The formula for the Community Hub-Bridge Centrality is as follows (Eq.2.6), where $card(c_i)$ represents the size of the community to which node i belongs and *nnc(i)* denotes the number of neighboring communities [59].

$$C_{HB} = k_i^{int} * card(c_i) + k_i^{ext} * nnc(i) \tag{2.6}$$

Harmonic Centrality Measure (HC): This centrality measure is global and obtained through the shortest path since the connectivity length of the graph is the harmonic mean. Defined as follows, where 1/d(x,y) is the shortest path from x to y [60][61].

$$C_{HC} = \sum_{x \neq y} \frac{1}{d(x,y)} \tag{2.7}$$

Betweenness Centrality: It is a measure of the degree to which a node is needed by others when connecting along the shortest paths, in comparison with centrality measures based on closeness, in betweenness centralities, it must be computed the number of shortest paths between pairs. This centrality helps measure how "popular" a node is based on the multiple shortest paths strategy [51].

2.2.2 Half Space Proximal Network

As stated, the chemical space can be visualized using networks, these networks are co-ordinated free representations called Chemical Space Networks (CSN). These networks are conceived as weighted graphs, that is for a given threshold, the similarity matrix $\mathcal{S}_{\mathcal{M}} = [s_{ij}]_{n \times n}$ stores the values computed by the similarity function between 0 and 1, for every set of nodes and become the adjacent matrix $\mathcal{A} = [a_{ij}]_{n \times n}$ whose values are given by [59]

$$a_{ij} + \left\{ \begin{array}{cc} s_{ij} if i \neq j, & s_{ij} \geq t \\ 0 & otherwise \end{array} \right\} \tag{2.8}$$

However, this kind of representation, especially with t = 0, is not recommended for large data sets. To build a complex network for large data sets, a large amount of RAM is required to store the corresponding (dis)similarity matrix. With a representation of lower density, the computational resources required decrease exponentially. The *Half-space proximal network* (HSPN) reduces the number of links between nodes while maintaining a metric space's (dis)similarity properties. [59] The principle for this complex network, is the Half-Space Proximal Test. This process generates a far less dense graph than a regular chemical space network [62].

HSP test [63]

Input : a vertex u of a geometric graph and a lit L_1 of edges incident with v

Output : A list of directed edges L_2 which are retained for the $H\vec{S}P(G)$ graph

1. Set the forbidden area F(u) to be $\varnothing$
2. Repeat the following while L_1 is not empty

- Remove from L_1 the shortest edge, say [u, v], and insert L_2 directed edge (u, v)with u being the initial vertex
- Add to F(u) the open half plane determined by the line perpendicular to the edge [u,v] in the middle of the edge and containing the vertex v, the point of the line does not belong to the forbidden area
- Scan the List L_i and remove from it any edge whose en vertix is in F(u)

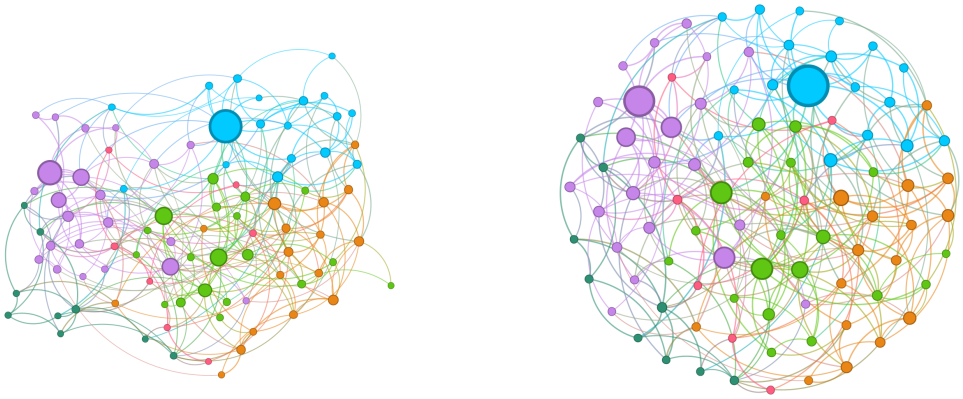
Layouts

Network visualization is achieved through various applications of a force-directed layout, a graph layout algorithm that models edges as springs, attracting nodes closer together while setting repulsion values to avoid overlap. This process resembles energy minimization in computational chemistry, which involves minimizing potential functions[64]. Also, it has been proved that force-directed layouts can help optimize modularity, which is key for community formation [65]. Nevertheless, these representations can be limited by the amount of nodes, where extensive networks can end up looking la hairballs. Thus, it is important to choose a layout that will help understand the topological structure of a network.

Fruchterman Reingold layout : Fruchterman [66] explains the principle for graph drawing using the following analogy: *"the vertices behave as atomic particles or celestial bodies, exerting attractive and repulsive forces on one another, this forces induced a movement. The algorithm will resemble a molecular or a planetary simulation "*. Certainly, this type of layout will always try to display the network as a sphere, where important parameters to consider are the surface area and the repulsion forces.

ForceAtlas 2: This is the default layout of Gephi [67]. Similar to the Fruchterman-Reingold layout, it is also based on the principle that edges attract and act as springs,

while nodes repel each other. However, unlike Fruchterman-Reingold, this layout has been more recently optimized and offers more tunable variables. The scaling and "prevent overlap mode" are particularly useful for handling large datasets. Other parameters, such as repulsion, gravity, and edge weight, can also be adjusted to achieve better results.



ForceAtlas 2 Fruchterman Reingold

Figure 2.1: Comparison of Different Layouts

2.3 Pairwise Alignment Algorithms

Pairwise Sequence Alignment is one of the keystone operations in bioinformatics. It is widely used to determine if two different sequences are structurally or functionally related. The operation aligns the two sequences to achieve maximal levers of identity to measure the similarity and the possibility of homology [68].

Smith-Waterman Alignment Algorithm: Often referred to as local alignment, this algorithm is employed to find the best subsequence between a target and a query sequence. The algorithm consists of two phases. In the first phase, an alignment matrix is computed for the two distinct sequences. The second phase involves extracting the best subsequence from the alignment matrix. Sequences expected to be quite dissimilar can be compared using this algorithm, which finds local regions with high levels of similarity [69].

Needleman-Wunsch Alignment Algorithm: It is commonly referred to as global alignment because is optimal to cover the entire length of two sequences. It is appropriate when both sequences have a similar length and demonstrate a considerable similarity throughout. Starting with an AA pair, a comparison is made between corresponding AAs

sequence. All potential pairs are organized in a two-dimensional array, and path represent all possible comparisons through this array [70].

2.4 Molecular Descriptors

Molecular descriptors, fundamental to contemporary computer-assisted toxicological and chemical applications, are numerical representations derived from molecular characteristics. They enable a mathematical treatment of molecules, signifying a pivotal step in converting molecular features into quantifiable data. These descriptors, defined as mathematical representations obtained through specific algorithms or experimental protocols, encapsulate distinct aspects of a molecule's chemical information [71].

Aliphatic Index: Aliphatic AAs are responsible for the thermal stability of peptides. Therefore are a good indicator of thermostability in general. This index is computed using the relative volume occupied by AAs with aliphatic side chain: alanine, valine, isoleucine and leucine [72].

Boman Index: this function was introduced in 2003 based on the aminoacidic composition of a sequence and measures the potential protein interaction, primarily based on the solubility of the AAs present as an estimation ability of a peptide to bind to membranes to receptors is a good indicative [73].

Hydrophobicity: This parameter is an important criterion to consider for the stabilization of the peptide. However, this interaction can change depending on the solvent in which the sequence is found[74].

Isoelectric Point: The isoelectric point is the value of pH at which a molecule, in this case a peptide, carries no net electrical charge, this parameter may affect the solubility of the of the compound depending of the pH of the medium [75].

Charge: The overall charge of a peptide is the sum of the charges of every group in a peptide that can be ionized [75].

GRAVY: The Grand average of hydropathicity index serves as a representation of a peptide's hydrophobicity. It is computed by summing the hydropathy values of all the amino acids within the sequence and then dividing this sum by the sequence length [76].

2.5 Multi Query Similarity Search

A fundamental principle in Medicinal Chemistry is that similar structures often exhibit similar biological activities, with the degree of structural similarity correlating to the degree of biological activity [77]. The Multi-Query Similarity Search (MQSS) method is grounded in this axiom. This approach leverages a known set of sequences as a reference for the biologically active domain. This set, known as the query set, must be carefully selected by researchers to represent the breadth of the active space. Subsequently, the target or unknown dataset is compared to this query set, and based on a predefined similarity threshold, it is determined whether a sequence belongs to the biologically active space or not. This process is illustrated in the schematic diagram below Figure.2.2.

Recent studies have showcased this non-trained supervised technique for predicting peptide bioactivities, including Hemolysis [78], Tumor-Homing [79], and Antiparasitic [80], with impressive results. This method trumps conventional ML methods in several ways: it's user-friendly, does not rely on web server availability, consumes fewer computational resources, and processes sequences with non-standard amino acids or varying lengths. Remarkably, MQSS models function without extensive training, relying instead on fine-tuning certain parameters like sequence alignment type – the similarity cutoff value. They can be developed without needing a negative dataset, a significant advantage given the scarcity of validated negative sequences, ensuring the learning phase is not skewed by data imbalances.

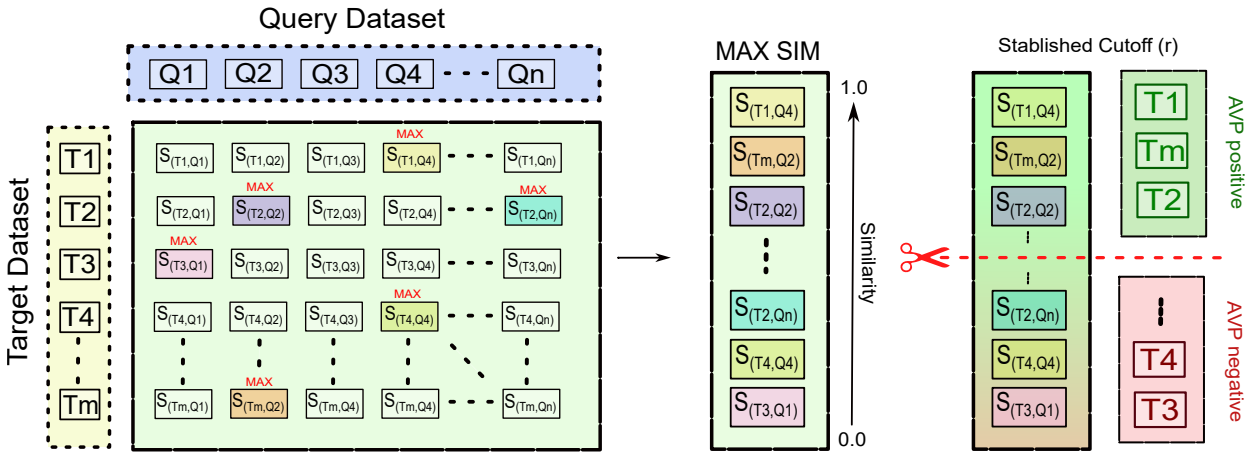


Figure 2.2: Muli-Query Similarity Search Principle.

2.5.1 Metrics

Friedman Test:

The interpretation of statistical information usually relies on variance analysis to propose ordinal rankings. However, this tool is unsuited for data that does not follow a normal distribution. In this case, The Friedman test is used to detect differences in treatments across multiple test attempts [81].

Definition 2 (Test Statistic)**.**

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \tag{2.9}$$

where k is the number of test attempts, n is the number of subjects and R_j is the sum of the ranks for the jth group

Mathews Correlation Coefficient

A binary classification task involves categorizing data into two distinct groups, making the evaluation of binary classifiers crucial in various research areas. Typically, in these classifiers, a positive response is associated with '1,' and a negative response with '0.' To assess performance, the standard practice is to create a confusion matrix, a 2×2 table that groups correctly and incorrectly classified instances into true positives, false positives, true negatives, and false negatives[82]. From the confusion matrix, fundamental ratios are calculated to gauge the predictor's performance. These ratios encompass the true positive rate (sensitivity), true negative rate (specificity), positive predictive value (precision), and negative predictive value[83].

The Mathews Correlation Coefficient (MCC), originally introduced in 1975 for comparing chemical structures, found new purpose in the 2000s as a performance metric for machine learning. One of its advantages is its resilience to imbalanced datasets. The MCC utilizes a contingency matrix approach, derived from the Pearson product-moment correlation coefficient. What sets MCC apart is its ability to yield high scores when the binary predictor accurately predicts both the majority of positive instances and the majority of negative instances in the data[84]. Additionally, it has already been established its supe-

riority in comparison with ROC AUC, F1 Score, and Balanced Accuracy, which are other common metrics derived from the basic ratios of the confusion matrix [82].

Chapter 3

State of the Art

3.1 Databases

Biological databases are pivotal in bioinformatics, serving as repositories of organized and relevant data [85]. As interest in Antiviral Peptides (AVPs) continues to grow within the scientific community, numerous research endeavors have been dedicated to investigating their structures, and mechanisms of action, and aggregating this knowledge into comprehensive databases [86]. AVPs are often classified as a subset of Antimicrobial Peptides (AMPs) in various categorization systems. This classification is particularly significant, as many existing databases incorporate AVPs within the broader framework of AMPs. Notably, the landscape of tools and databases available for AMP studies has been extensively surveyed by Ramazi et al. [87], encompassing a range of general AMP databases. Several of these databases are also covered here, while additional focus is directed toward specific antiviral databases, see Table.3.1.

Furthermore, this review highlights the scope of coverage of antiviral sequences within these general databases. It is worth noting that while some specialized peptide databases might not explicitly categorize their entries as "Antiviral," databases such as Defensins [88] and Cybase [89] still holds substantial importance. They contribute invaluable insights into the mechanisms of action and structural characteristics of therapeutic peptides, thus enhancing our understanding of potential antiviral candidate peptides.

Table 3.1: A summary of existing antimicrobial and antiviral peptide databases

Database Name	Covering Class	Aproximate Size	Year	Reference
dbAMP 2.0	AMPs	1803 AVPs 187 Anti SARS-CoV	2022	[90]
DBAASP	AMPs	1454 AVPs 53 Anti HIV 80 Anti SARS-CoV	2021	[91]
LAMP	AMPs	4320 AVPs	2020	[92]
DRAMP	AMPs	2219 AVPs	2019	[93]
InverPep	AMPs	10 AVPs	2017	[94]
CAMP R4	AMPs	117 AVPs	2022	[95]
APD3	AMPs	172 AVPs	2015	[96]
AVPdb	AVPs	2683 seq	2014	[97]
HIPdb	Anti HIV AVPs	981 seq	2013	[98]
ACovPepDB	Anti SARS-CoV	214 seq	2022	[99]
AntiCoV_DB	Anti SARS-CoV	34 Anti COVID-19 104 anti SARS-CoV-2	2023	[100]
DRAVP	AVPs	1986 AVPs. 46 Specific Virus Classification	2023	[101]

3.1.1 StarPepDB and StarPep Toolbox

StarPepDB functions as an integrated graph database housing peptide sequences and corresponding metadata, organized in interconnected nodes. This repository boasts a substantial collection, comprising 71,310 nodes and 348,505 connections [102]. To facilitate easy access and utilization of the wealth of information within StarPepDB, the complementary software StarPep toolbox was developed [59]. This toolbox offers a range of visual analytic processes, encompassing peptide queries, filtering, 3D structure visualization, network construction, characterization, and more. Both StarPepDB and StarPep toolbox are accessible at http://mobiosd-hub.com/starpep/

Distinguished as the largest compiled database to date, StarPepDB uniquely integrates various individual databases [103, 104]. This affords a notable edge over previously established peptide databases. The toolkit provided by StarPep toolbox and its graph-based approach extends beyond the realm of antiviral peptides, serving diverse research endeav-

ors. For instance, it has been instrumental in proposing 54 leads for tumor-homing peptides [79], identifying motifs for anti-biofilm peptides [105], repurposing 95 AMPs as potential anti-parasitic peptide hits [80], and identifying 47 potential hemolytic motifs [78].

3.2 Prediction Models

3.2.1 Encodings

With the increasing application of machine learning-based models in bioinformatics, one key factor for successfully using these models is translating of peptide sequences into numeric vectors. Although this is a crucial consideration when designing these methods, it's important to note that most encoding methods were developed before the advent of deep learning. The most common encoding methods include one-hot encoding, BLOck SUbstitution Matrix (Blosum), and physicochemical character-based encodings[106].

In a more general approach, encoding can be subdivided into sequence-based encodings and structure-based encodings. Sequence-based encoding methods include Sparse, Amino Acid Composition (AAC), distance frequency, quantitative Matrix, CTD (Composition, Transition, and Distribution), Pseudo-amino Acid (PseAAC), AAindex, Physicochemical properties, Substitution Scoring Matrix, and BLOMAP. On the other hand, some structure-derived encodings include QSAR, general structure, electrostatic hull, distance distribution, and more [107].

Substitution Score Matrix

Substitution matrices, such as **BLOSUM62**, represent accepted mutations between amino acid pairs in sequence alterations[107]. Their primary purpose is to determine whether a pair of sequences are homologous by assigning an alignment score to them. In this context, it is assumed that the positioning of amino acid pair residues in alignments is statistically independent of others. Thus, there's a probability for pairing amino acids 'a' and 'b' residues in homologous sequences. Hence, it is possible to calculate the likelihood that these two residues are uncorrelated and occurring independently[108].

So, if this specific residue pair is found more often than expected by random chance,

a positive score is assigned, leading to a conservative substitution. In **BLOSUM62**, the scores assigned to some pairs of amino acids are not the same. This variation in scores reflects that in the homologous alignment data on which **BLOSUM62** was trained, some pairs of amino acids appeared more commonly than others, resulting in a different assignment of scores for rarer alignments[109].

3.2.2 Machine Learning Based Models

The current pipeline for managing big data stored in databases involves machine learning (ML), which efficiently analyzes extensive multidimensional information [110]. Traditional ML algorithms, such as Support Vector Machine (SVM), k-nearest neighbor (kNN), random forest (RF), single neural network (NN), and deep learning algorithms (DL), have proven to be efficient methods for recognizing patterns within peptide sequences and exploring the potential of new sequences [111]. The available prediction models are presented in Table.3.2 The typical workflow of these methods begins with input encoding, followed by model construction using either traditional ML or DL algorithms. One notable advantage of DL methods is their reduced dependency on prior knowledge and well-engineered input features. These DL techniques encompass Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Recursive Neural Networks (RNN)[112].

However, a significant point of controversy within the field revolves around the justification of using DL models for predicting AMPs in general. Most DL methods necessitate extensive datasets of experimentally validated peptide sequences. In contrast to other fields where deep learning is widely employed, the available data in this context is often insufficient. One approach to address this insufficiency is "data augmentation," as employed by Lin et al. [113] for negative sequences. Nevertheless, this approach has not been fully explored.As noted by Garcia-Jacas et al. [114], there is no significant improvement in using DNN over traditional ML methods, and the chemical space produced by these algorithms often overlaps. Additional challenges related to available datasets include the overrepresentation of certain sequences and imbalanced data distribution, which can lead to difficulties when evaluating performance solely based on accuracy. Furthermore, issues concerning the lack of result reproducibility persist, as not all researchers share their

source code or datasets, hindering the broader adoption of these methods within the scientific community. Consequently, while machine learning-based methods offer promising tools for predicting biologically active peptides, there is ongoing work required to refine these models and obtain valuable results [111, 112].

Table 3.2: Available Web Servers for AVP Prediction

Name	Year	Algorithm	Encoding	Implementation	Ref
iACVP	2022	RF	word-embedding word2vec	http://kurata35.bio.kyutech.ac.jp/iACVP/	[115]
AI4AVP	2022	CNN	PC6, ENNAVIA descriptor , AAC, PseACC, AA index, DPC	https://axp.iis.sinica.edu.tw/AI4AVP/	[113]
PTPAMP	2022	SVM	AAC, DPC	http://www.nipgr.ac.in/PTPAMP/	[116]
Deep-AVPpred	2022	ANN	Pretained Embeddings	https:// deep- avppr ed. anvil. app/	[117]
AVPIden	2021	Sharpley Value	Phyc, PAAC, DPC, AAC, CKSAAGP	https://awi.cuhk.edu.cn/AVPIden/#/	[118]
ENNAVIA	2021	NN	AAC, DPC, TPC, Reduced AAC, GDC, GTC, CTD, PseACC, Phyc, AA index , steric hindrance , bulkiness, secondary strcuture propensities, side chain interactions, membrane buried preference parameters	https://research.timmons.eu/ennavia	[119]

Table 3.3: Available Tools for AVPs prediction

Name	Year	Algorithm	Encoding	Implementation	Ref
ProtDcal	2021	RF, RNN	GTPC,APseAAC, PseACC, GAAC, ACC, PHYS	https://biocom-ampdiscover.cicese.mx/	[103]
AMPfun	2020	RF	AAC-bas,Phyc, and word frequency-based features	http://fdblab.csie.ncu.edu.tw/AMPfun/index.html	[120]
Meta-iAVP	2019	RF	AAC, Am-PseAAC	http://codes.bio/meta-iavp/	[121]
PEPred-Suite	2019	RF	AAC, ASDC,CTD, 188D, GGAP,BIT20, BIT21,OLP,IT,DC	http://server.malab.cn/PEPred-Suite.	[122]
MLAMP	2016	RF, ML-SMOTE	PseAAC	http://www.jci-bioinfo.cn/MLAMP	[123]
AVP-IC50Pred	2015	SVM, RF, IBk, kStar	IC50	http:// crdd.osdd.net/servers/ ic50avp/	[124]
iAMP-2L	2013	FKNN	PseAAC	http://www.jci-bioinfo.cn/iAMP-2L.	[125]
ClassAMP*	2012	RF, SVM	AAC,Phyc, BLOSUM 50 MATRIX, normalized van der Waals volume, secondary structure propensity, DPC, TPC, CTD	http://www.bicnirrh.res.in/classamp/	[126]
AVPpred	2012	SVM	PseAAC	http://crdd.osdd.net/servers/avppred/	[127]

Table 3.4: Available Tools for AVPs prediction

Name	Year	Algorithm	Encoding	Implementation	Ref
LSTM_pep	2023	LSTM	-	https://github.com/haiping1010/New_peptide_iteration	[128]
Pep-CNN	2022	CNN	AAC,Phys, evolutionary derived	https://github.com/alivelxj/Pep-CNN	[129]
seqpros	2022	MLP,LSTM	Phys	https://github.com/eotovic/seqprops_therapeutic.	[130]
TransImbAMP	2022	Transformer	-	https://github.com/ BiOmicsLab/TransImbAMP	[131]
PreAntiCoV	2021	RF	AAC,DPC, K-spaced AA, PseAAC,Phys	https://github.com/poncey/PreAntiCoV	[132]
iAMP-CA2L	2021	CNN-BiLSTM-SVM	cellular automata image	https://github.com/liujin66/iAMP-CA2L/tree/main	[133]
FIRM-AVP	2020	RF, SVM, DL	Aac, DC, APseAAC, CTD, secondary structure	https://github.com/pmartR/FIRM-AVP	[134]
AntiVPP	2019	RF	Phys, AAC	https://github.com/bio-coding/AntiVPP	[135]

Chapter 4

Methodology

4.1 Half-Space Proximal Network

The methodology employed for constructing the HSPN followed a similar approach as described in [79, 78, 80]and [105] when successfully studying hemolytic, tumor-homing, antiparasitic and antibiofilm activities in peptides. However, the graph-based method has not been applied to antiviral peptides. This methodology allows us to leverage a network representation for data mining. The initial StarPepDB consisted of 45120 sequences, out of which 4663 were classified as antiviral based on their function. To ensure a more representative subset for the HSPNs, the Smith-Waterman alignment algorithm removed sequences with a redundancy higher than 95%. As a result, the number of sequences was reduced to 3494 AVPs. The Euclidean metric and Min-Max normalization were employed to calculate pairwise similarity between peptides. The molecular descriptors used for feature extraction, which are found in Starpep toolbox, included Peptide Length, Net Charge, Isoelectric Peptide, Molecular Weight, Boman Index, Hydrophobic Moment, Average Hydrophilicity, Hydrophobic Periodicity, Aliphatic Index, Instability Index, and Indices based on Aggregation Operators.

Various values of t ranging from 0.3 to 0.9 were tested to investigate the impact of the similarity threshold (t) in the HSPNs. Additionally, a network with $t = 0$ (no similarity cut-off is used, that is, a parameter-free complex network) was included. Further refinement of each HSPN was performed by applying the Louvain algorithm for clustering. The centrality

characterization of most HSPNs was carried out using the HB centrality measure. However, with smaller networks, HC was also used. All these steps were carried out using the StarPep Toolbox (http://mobiosd-hub.com/starpep/).

4.2 Metadata Complex Network

Another notable feature of the StarPep toolbox is its ability to generate Metadata complex Networks (MNs). These networks are constructed as bipartite graphs, leveraging the 3494-sequence subset of AVPs as the first set of nodes. For the second set of nodes, metadata information categorized by the StarPep toolbox as "Database", "Origin", "Function", and "Target" was utilized. To fully appreciate the bipartite graph structure provided by the MNs, edges linking peptides within the same set were removed based on the initial similarity network. This removal highlights the hierarchical relationships within the MNs. For instance, each "peptide" node is connected to an "origin" node, and different "peptide" nodes can be connected to the same "origin" node, but not to each other, resulting in a visible hierarchy. In contrast to the previous networks discussed, the centrality of the MNs was measured using the Betweenness centrality.

4.3 Network visualization and Characterization

All HSPNs were visualized using Gephi 0.10 ([136], https://gephi.org/), utilizing the Fruchterman-Reingold Layout [66] with an Area set to 1×10^8 and a speed of 20. The HSPNs were color-coded to represent different clusters, and the size of nodes was scaled according to their respective centrality measure (HB centrality) to enhance visual clarity. Using the Statistical Features integrated into Gephi 0.10, various parameters were calculated for each HSPN with varying cut-offs (t = 0.3-0.9). The reported parameters included Average Degree, Network Density, Modularity, ACC, and Average Path Length. The number of singletons (atypical sequences) in the network was estimated using the "Giant Component subgraph" and/or vertex degree equal to 0. These parameters aided in selecting the HSPN with the optimal t value. The selected HSPN with optimal t value was taken as the main reference to compare with the HSPN without cut-off value; to assess the

effect of said parameter in the representation of AVPs.

The five most central AVP nodes were extracted from each HSPN to gain further insight into the selected complex networks. Some molecular properties of these sequences were calculated using the 'Peptides' package for R ([137], https://cran.r-project.org) . The molecular descriptors used for this characterization included Aliphatic Index, Boman Index, Hydrophobicity, Isoelectric Point, Charge, and peptide Length. Additionally, some biological activity information was incorporated through cross-referencing the metadata provided by StarPepDB . The peptides were visualized as a ten-node HSPN , and their similarity overlap was measured using Dover Analyzer [138].

4.4 Exploration of Scaffold and selection of most representative Subset

Starting from the selected HSPNs (t = 0.75 and t = 0) a scaffold extraction was performed using the data mining tools provided by the StarPep toolbox. This process involves modifying several variables, including the centrality measure, the pairwise sequence alignment algorithm, and pairwise sequence % identity to obtain the best-reduced representation of the chemical space. Both the Harmonic and Community Hub-Bridge Centralities were utilized for the centrality measure. The alignment algorithms used were Needleman-Wunsch and Smith-Waterman . Additionally, the pairwise sequence's % identity was adjusted from 90% to 50%. These modifications resulted in the generation of 20 distinct scaffolds from each of the selected HSPNs (t = 0.75 and t = 0) Subsequently, the generated subsets were compared using the Dover Analyzer. To assess the impact of the centrality measure, alignment algorithm, and HSPN cut-off on the scaffold extraction process, subsets obtained with the same sequence's % identity were grouped for analysis. The comparison was based on the percentage of identical and similarity overlap between sequences within each scaffold. These comparisons identified the best and least redundant subsets for each sequence's % identity. These subsets were compiled into a consolidated dataset, representing the optimal and most representative choices for each percentage of allowed similarity. Finally, these scaffolds were visualized in Gephi using the same methodology described earlier.

4.5 Motif Discovery

The motif discovery was conducted using the alignment-free method called STREME ([139], https://meme-suite.org/meme/tools/streme) , which is part of the MEME Suite 5.5.2 [140]. The 8 communities identified by the clustering algorithm in the HSPN (t = 0) from StarPep toolbox were used to obtain these motifs. Each community was converted and separated into individual .fasta files, which served as input for the motif extraction process using STREME. The motif width was set to a minimum of 3 and a maximum of 6 AAs. The search for motifs was limited to those with a p-value threshold of 0.05.

Parallel to the motif discovery, an analysis of the communities was conducted. Initially, several molecular descriptors were calculated for the three most central nodes of each of the 8 clusters, following a similar chemical characterization approach as explained previously. Subsequently, a literature search was performed to expand the characterization of these peptides and gain a deeper biological insight into their properties. Additionally, based on the sequence of these peptides there were searches for the occurrence of the motifs discovered within the same cluster.

4.6 Alignment Free Motif Enrichment

The motifs extracted by STREME were further validated using the Sequence Enrichment Analysis [141] tool from the MEME Suite 5.5.2 . This validation involved assessing the relative enrichment of these motifs in external databases [103]. The following external datasets were used for validation:

- B-TS-StarPepAVP (272 positives)
- Ex-StarPepAVP (1230 positives)
- TR-StarPepAVP (622 positives)
- TS-StarPepAVP (622 positives)

To avoid redundancy between the datasets, the similarity among them was also studied using Dover Analyzer. In the validation process using SEA, the E-value threshold was set

to be lower than or equal to 10. The enrichment E-value of a motif is calculated as the adjusted p-value multiplied by the number of motifs in the input. The adjusted p-value represents the probability of the motif distinguishing the primary sequences from the control sequences.

A negative dataset was selected for "inverse-validation" to eliminate the possibility that the identified motifs have the same probability of occurring in both positive and negative sequences. The negative dataset was constructed by combining the negative sequences from the external datasets mentioned previously. Redundant sequences were removed, resulting in a 13715 unique negative sequences dataset. After validating the motifs, they were searched within the most central sequences of each source cluster. This analysis aimed to explore the presence and significance of the motifs within the central sequences of each cluster. A literature search was also conducted to investigate any similarities or connections to other prediction reports, further enhancing the understanding of the motifs and their potential biological relevance. By conducting these analyses, it was possible to validate and explore the motifs concerning both positive and negative sequences, and to gain insights from existing literature and prediction studies.

4.7 Multi-Query Similarity Search

The Multi-Query Similarity Search (MQSS) method was previously introduced in the Theoretical Fundamentals section (see Chapter 2). For executing the MQSS, 15 distinct "Query" datasets were employed. These reference datasets were created using the scaffold extraction algorithm provided by the StarPep toolbox. The selection of these scaffolds was made based on their representativeness in comparison to similar ones. Furthermore, five of the "Query" datasets were formed by merging different scaffolds with similar sequence identity percentages, enhancing the sequence representativeness of these datasets. More comprehensive information about the "Query" datasets can be found in Table 4.1.

In conjunction with the "Query" datasets, several critical variables must be considered for the MQSS process. The first involves determining the similarity cutoff (r), which was varied within the range of 0.3 to 0.9. The computation of similarity was based on the BLOSUM-62 substitution matrix. The second variable involves selecting the type of

alignment algorithm, whether global or local. These parameters collectively resulted in a total of 210 models being tested during the initial phase of the MQSS process.

Table 4.1: Description of the used "Query" Datasets

Name	Size	Description	Name	Size	Description
Md1	1,872 seq	Merged of scaffolds with 50% from scaffold extraction	SG4	2,562 seq	Constructed using Global Alignment, HB Similarity Measure and 80% sequence identity
Md2	2,152 seq	Merged of scaffolds with 60% from scaffold extraction	SG5	3,119 seq	Constructed using Global Alignment, HB Similarity Measure and 90% sequence identity
Md3	2,445 seq	Merged of scaffolds with 70% from scaffold extraction	SL1	1,030 seq	Constructed using Local Alignment, HC Similarity Measure and 50% sequence identity
Md4	2,703 seq	Merged of scaffolds with 80% from scaffold extraction	SL2	1,557 seq	Constructed using Local Alignment, HC Similarity Measure and 60% sequence identity
Md5	3,206 seq	Merged of scaffolds with 90% from scaffold extraction	SL3	2,028 seq	Constructed using Local Alignment, HC Similarity Measure and 70% sequence identity
SG1	1,626 seq	Constructed using Global Alignment, HB Similarity Measure and 50% sequence identity	SL4	2,369 seq	Constructed using Local Alignment, HC Similarity Measure and 80% sequence identity
SG3	1,991 seq	Constructed using Global Alignment, HB Similarity Measure and 70% sequence identity	SL5	3,003 seq	Constructed using Local Alignment, HC Similarity Measure and 90% sequence identity

4.7.1 Selection of best models

Model Calibration

To identify the optimal model, an exhaustive search for available AVP datasets was conducted to compile a comprehensive set of sequences (refer to Table 4.2 for evaluation details). It's important to note that for studies encompassing multiple types of AMPs, the search was limited to the antiviral subsections of these databases. This selection process was executed through multiple filtering stages.

Table 4.2: Description of the used "Target" Datasets

Dataset	Size	Positives	Negatives	Ref
TR_StarPep	4,642	2,321	2,321	[103]
TS_Starpep	1,246	623	623	[103]
Ex_Starpep	12,001	1,230	10,771	[103]
AVPIden	53,116	2,662	51,116	[118]
AMPfun	5,826	2,001	3,825	[120]
ENNAVIA_A	974	557	420	[119]
ENNAVIA_B	1,154	557	597	[119]
Imb	12,234	2,038 139 (Anti-CoV)	10,196	[142]
Thakur	1,056	604	452	[127]
Sharma	6,544	3,273	3,271	[117]
AI4AVP	20,222	2,934	17,288	[113]
ENNAVIA_C	465	109 (Anit-CoV)	356	[119]
ENNAVIA_D	469	110 (Anti-CoV)	359	[119]
HIPdb	981	981	-	[98]
Expanded	55,822	3,178	52,644	-
Reduced	27,692	1,419	26,273	-

The model selection procedure can be divided into two distinct sections. The first, described here, pertains to the calibration of model construction, wherein various hyperparameters were fine-tuned. These parameters encompassed the selection of the best "Query" Dataset, as outlined in Table 4.1 alignment algorithm, and the value of the cutoff parameter (r). The calibration process occurred across two evaluation rounds. In the initial round, the evaluation was conducted using datasets provided by [103]. Specifically, the TS_StarPep, TR_StarPep, and EX_StarPep datasets were utilized as targets for the Multi-Query Similarity Search (MQSS). This evaluation stage resulted in reducing the number of models from 210 to 80.

The subsequent stage encompassed six distinct datasets: AVPIden, AMPfun, ENNAVIA A, ENNAVIA B, Imb, and Thakur. Evaluating these models using the aforementioned datasets as targets further refined the selection, reducing the number to 50 models.

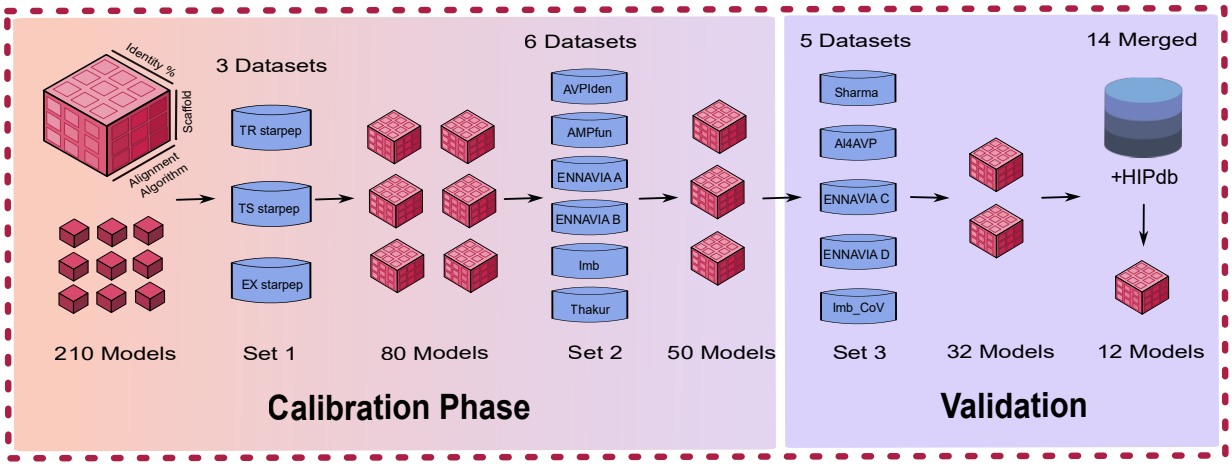


Figure 4.1: Calibrationa and Validation of MQSS

Model Validation

In the third stage of reduction, the performance of the remaining 50 models was rigorously assessed using an additional set of five datasets: Sharma, AI4AVP, ENNAVIA C, ENNAVIA D, and Imb_CoV. This meticulous evaluation process eventually led to the final selection of 32 models. Importantly, this evaluation phase included datasets specifically tailored to certain types of viruses such as SARS-CoV, enabling the assessment of the model's performance against more targeted AVPs.

To further refine the selection process, an "Expanded" dataset was constructed by aggregating all sequences from the 14 individual datasets (Figure.4.2). Furthermore, positive sequences demonstrating anti-HIV activity from the HIPdb[98] were integrated. This amalgamated dataset encompassed a total of 70,126 negative sequences and 20,136 positive sequences. Following the removal of sequence redundancy, the dataset contained 54,088 negative sequences and 4,745 unique positive sequences. To enhance accuracy, overlapping sequences that were reported both as positive and negative were excluded from the dataset, leading to the elimination of 1,567 overlapping sequences. The evaluation conducted using the Expanded dataset led to the identification of 12 models that exhibited optimal performance characteristics. The final selection of these 12 models was determined through an Average Ranking of the Friedman test. The Statistical test and corresponding significance test were carried out using the KEEL software ([143], https://sci2s.ugr.es/keel/development.php#x1-20001) and the Non-parametric Statistical Analysis Module. The described

workflow is better depicted in Figure.4.1

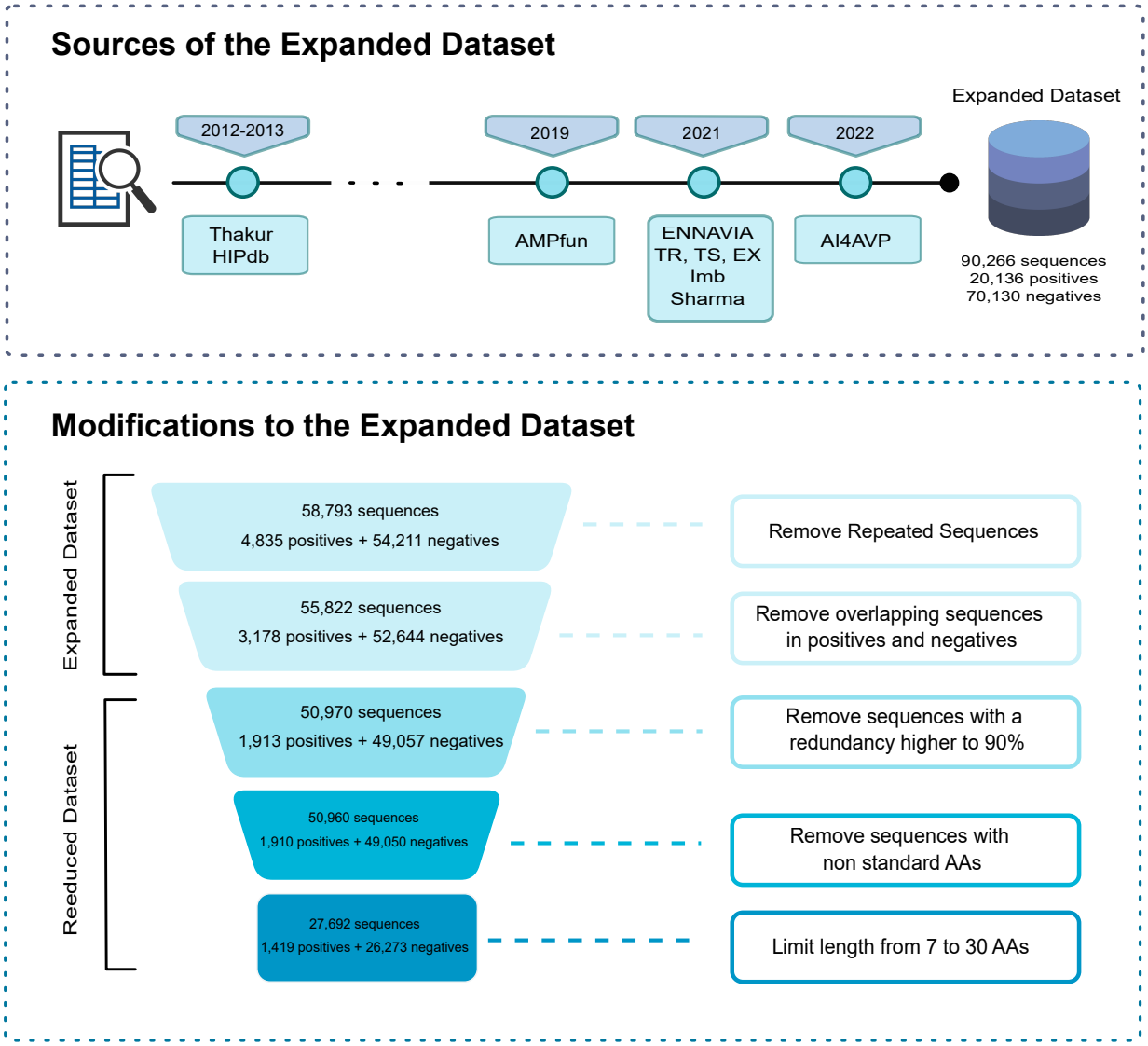


Figure 4.2: Construction and Modification of Expandend Dataset

4.7.2 Model Improvement

One significant challenge that necessitates addressing in these models is the precision of recalling positive sequences. Various approaches were undertaken to explore avenues for enhancing the detection of positive sequences.

A posteriori Modification

The first approach involved aggregating multiple models and employing a majority vote system to generate a new prediction for a given sequence. Given the total of 12 base models, there were 220 possible combinations of 3 models, and 720 possible combinations for groups of 5 and 7 base models. These amalgamated models were designated as meta models. The evaluation of these metamodels was conducted using the Expanded dataset. The concept of the majority vote system is visually depicted in Figure 4.3.

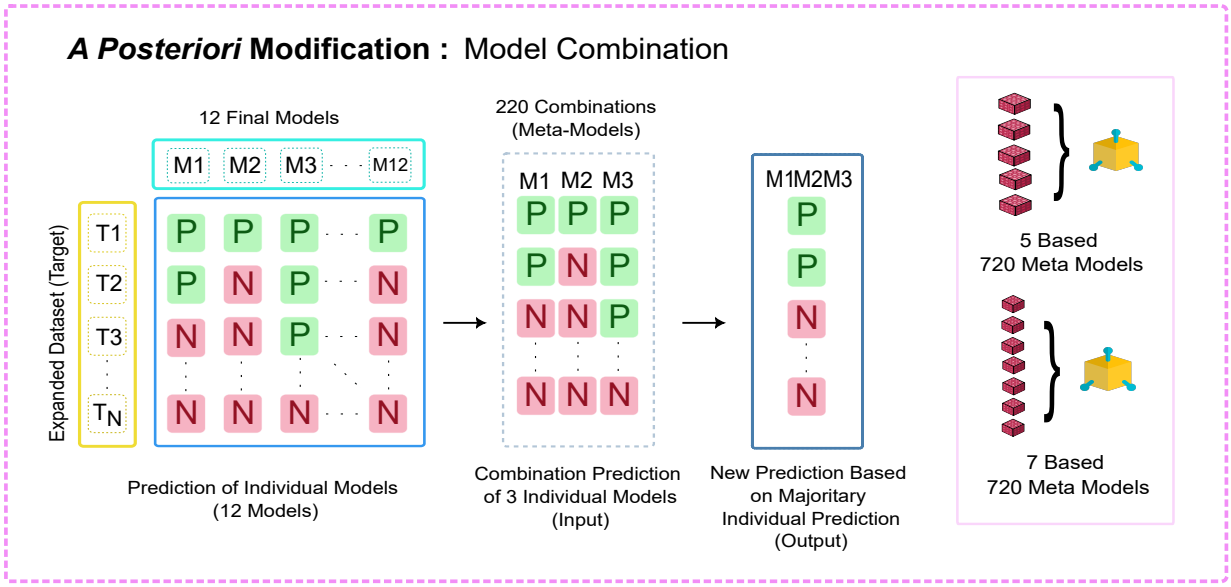


Figure 4.3: Process for Model Combination

A priori Modification

The second approach involves enhancing the "Query" Dataset. After identifying the top 3 base models, the scaffolds employed in the QMSS were chosen. These selected scaffolds were merged and any redundancy within the set was eliminated. New models were constructed using the refined query set while retaining the alignment algorithm and cutoff value from the parent models.

Query Enrichment

The primary challenge associated with the accuracy of positive sequences can also be attributed to the inadequate representation of specific sequence types absent in StarPepDB.

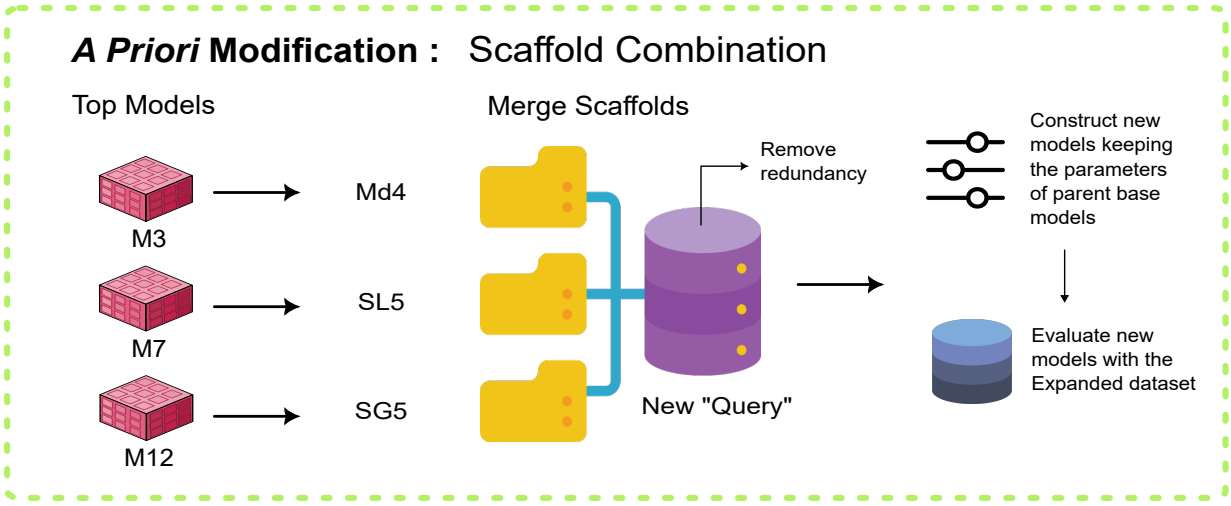


Figure 4.4: Process for Scaffold Combination

Consequently, the MQSS struggles to accurately predict such sequences. The second round of evaluations during the model selection's calibration stage was employed as a starting point to enhance the "Query" Dataset with additional sequences. Datasets where models exhibited notably poor performance were singled out. These datasets encompassed Thakur, ENNAVIA A, AMPfun, and Imb. Positive sequences from these four datasets were extracted in a single new dataset, with redundant sequences subsequently removed. This process resulted in 2403 unique sequences. To ensure the absence of these sequences from the currently top-performing model's scaffolds, a pairwise similarity comparison was conducted using Dover Analyzer.

With the newly curated and validated dataset , it was integrated into the StarPep toolbox to construct a HSPN, similar to the HSPNs developed previously for the entire AVP space. Following HSPN contruction, Scaffold Extraction was applied to the network, varying the centrality measure between Harmonica and Community Hub-Bridge. The alignment algorithm type was adjusted between local and global, considering only 80% and 90% as the sequence identity percentages. This process yielded 8 new scaffolds, designated as "external scaffolds." These scaffolds were utilized to develop new MQSS models, employing the same methodology as before, resulting in 112 new models, as depicted in Figure.4.5. These newly generated models underwent testing against the Expanded Dataset to identify the most effective external scaffolds.

Once the best-performing scaffolds were pinpointed, they were incorporated into the

scaffolds of the best-performing base models, while eliminating redundancy among the sequences. These enriched queries were employed once again in constructing models while retaining the parameters of the parent base models.

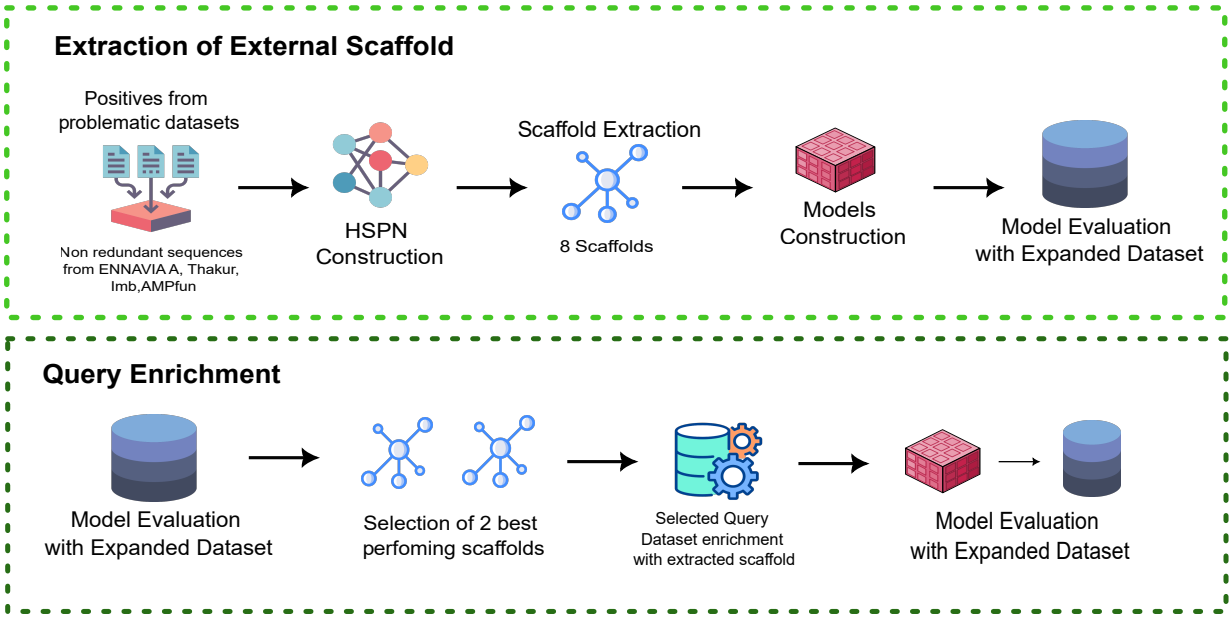


Figure 4.5: Process for Query Enrichment

4.7.3 Model Performance Evaluation

Performance is assessed across various datasets at every stage of the model selection, validation, and improvement process. This performance evaluation is conducted using commonly employed metrics in Machine Learning Based Predictors. These metrics encompass Accuracy (ACC), Sensitivity (SN), Specificity (SP), the Mathews Correlation Coefficient, the F1 Score, and the False Positive Rate (FPR) [144].

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

$$SN(Recall) = \frac{TP}{TP + FN} \tag{4.2}$$

$$SP = \frac{TN}{TN + FP} \tag{4.3}$$

$$Precision = \frac{TP}{TP + FN} \tag{4.4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.5}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.6}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.7}$$

Where TN are the true negatives, TP are the true positives, FN are the false positives and the FP are the false positives.

Comparison with State of the Art

To assess the robustness of the MQSSMs, a selection depicted in Table.4.3 of state-of-the-art predictors was made, primarily considering their availability through web services and ease of implementation. The table below provides a summary of the predictors used in this section. For the comparison with the predictors the Reduced Dataset was employed as it was stated previously:

Table 4.3: State of the Art predictors used for comparison

Predictor	Year	Algorithm	Implementation	Ref
AI4AVP	2022	CNN	https://axp.iis.sinica.edu.tw/AI4AVP/	[113]
iACVP	2022	RF	http://kurata35.bio.kyutech.ac.jp/iACVP/	[145]
PTPAMP	2022	SVM	http://www.nipgr.ac.in/PTPAMP/	[117]
seqpros	2022	MLP,LSTM	https://github.com/eotovic/seqpropstherapeutic	[130]
ProDcal	2021	RF,RNN	https://biocom-ampdiscover.cicese.mx/	[103]
AMPfun	2020	RF	http://fdblab.csie.ncu.edu.tw/AMPfun/index.html	[120]
FIRM-AVP	2020	RF,SVM, DL	https://github.com/pmartR/FIRM-AVP	[134]
Meta-iAVP	2019	hybrid	http://codes.bio/meta-iavp/	[121]
AntiVPP	2019	RF	https://github.com/bio-coding/AntiVPP	[146]
ClassAMP	2012	RF SVM	http://www.bicnirrh.res.in/classamp/	[126]
AVPpred	2012	SVM	http://crdd.osdd.net/servers/avppred/	[127]

4.8 Lead Discovery from Protein Cleavage

The objective is to propose antiviral sequences through Virtual Cleavage, using the models developed in the last section. Although the process involves more steps than just applying the models, the workflow for this section is explained in the diagram.

The initial point applying of the workflow (Figure.4.6) is three databases: the Starpep database, a human proteome database, and a cephalopod peptides database. The Starpep database contains the 45,120 sequences embedded in the StarPepDB. The human proteome database contains 43,000 novel cryptic AMPs scanned using a scoring function. Finally, the cephalopod database was crafted applying 13 enzymatic digestion protocols to the proteins found in cephalopods' salivary glands[147]. The peptides in these databases were filtered to retain only sequences with fewer than 35 amino acids and using standard amino acids. For the StarPep DB, sequences labeled as antiviral, toxic, or hemolytic were also removed.

After the initial filter, the M13+ MQSS model was applied. The remaining sequences were compared with all the positive sequences in the Expanded dataset and the experimentally negative sequences in the same dataset. Another filter was applied to remove sequences with a similarity higher than 90%.

The resulting group of potential antiviral sequences was subjected to some state-of-the-art predictors to increase the chances of identifying antiviral activity. The web servers used in this case included AMPfun, iACVP, meta-iAVP, AI4AVP, PTPAMP, and ProtDcal. Sequences that received a positive prediction from most the models were selected for the next filter.

Finally, the selected sequences were subjected to toxicity, hemolysis, and allergen predictors to eliminate sequences predicted as toxic, hemolytic, or allergen. The GRAVY index was also calculated to further refine the set of sequences presented. Certain physicochemical properties were calculated using the "peptides" package from R to provide more information. All the tools employed in this section are listed in Table.4.4

Table 4.4: Web-Available tools used for Virtual Cleavage

Web Server	Activity	URL	Ref
AMPfun	Antiviral	http://fdblab.csie.ncu.edu.tw/AMPfun/index.html	[120]
iACVP		http://kurata35.bio.kyutech.ac.jp/iACVP/	[145]
PTPAMP		http://www.nipgr.ac.in/PTPAMP/	[116]
Meta-iAVP		http://codes.bio/meta-iavp/	[121]
AI4AVP		http://axp.iis.sinica.edu.tw/AI4AVP/	[113]
ProtDcal		https://biocom-ampdiscover.cicese.mx/	[103]
ToxinPred	Toxicity	https://webs.iiitd.edu.in/raghava/toxinpred/algo.php	[148]
HemoPred	Hemolysis	http://codes.bio/hemopred/	[149]
ALGpred2	Allergens	https://webs.iiitd.edu.in/raghava/algpred2/algo.html	[150]
GRAVY	GRAVY	https://www.gravy-calculator.de/	

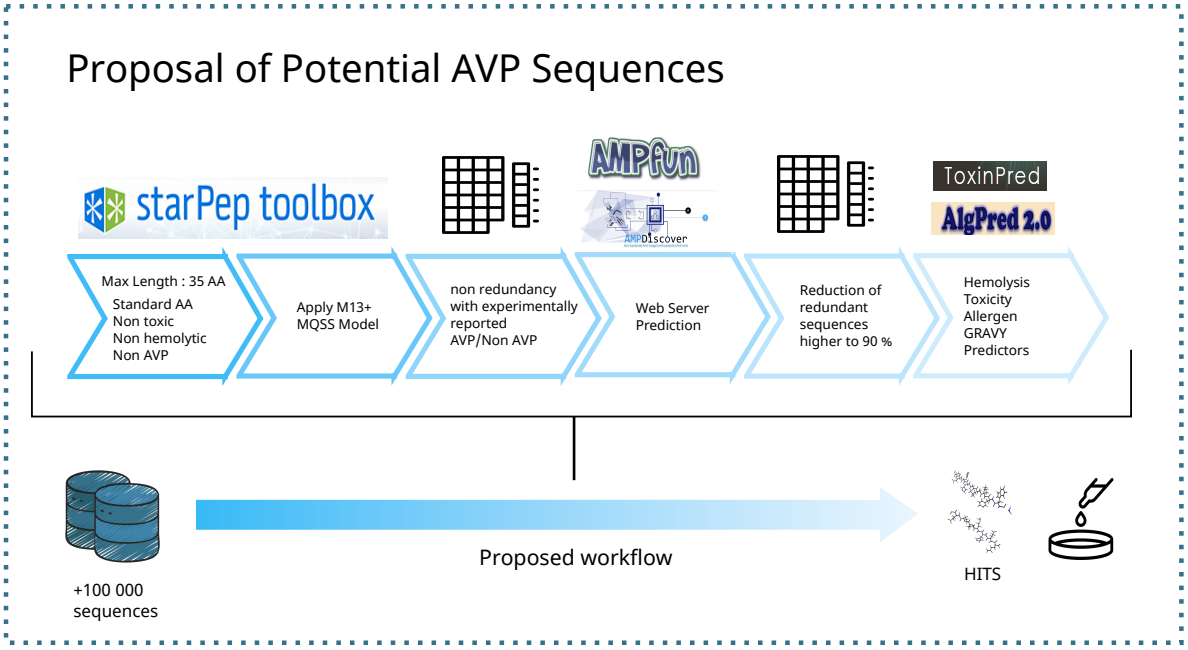


Figure 4.6: Workflow for Virtual Cleavage

To visually represent the selected groups of potential antiviral hits, and was created an HSPN using the sequences as nodes. Once the network was constructed, the Louvain algorithm for clustering was employed and it was calculated the Community Hub-Bridge Centrality for each node. This information guided us in hand-picking specific sequences based on their representativeness and significance within the network. To streamline the presented sequences, the Scaffold Extraction Algorithm was utilized. This process utilized

the HB Centrality calculation in combination with a local type of sequence alignment, considering a 50% sequence pairwise identity threshold.

For the final set of sequences, we estimated additional antimicrobial activities using the AMPfun and the AMPDiscover (ProtDcal Hierchal) web servers. These assessments covered antiparasitic, antifungal, and antibacterial activities.

Chapter 5

Results and Discussion

5.1 Metadata Complex Networks

Metadata Networks (MNs) serve as an essential starting point for the basic characterization of AVPs, utilizing the information available in StarPepDB. MNs provide researchers with a comprehensive view of the metadata associated with AVPs, enabling them to explore the distribution and interconnections of AVP sequences based on various attributes, including their sources, functions, or targets. These complex networks highlight the hierarchical relationships within the data, particularly between the "peptide" nodes and the corresponding "metadata" nodes. Additionally, other hierarchical structures may emerge among the "metadata" nodes due to the presence of redundant classification categories (Figure.5.1A).

Database MN : The databases that contribute the majority of AVPs in StarPepDB include SATPdb [151], AVPdb [97], DBAASP [152], DRAMP_General [153] and, LAMP_ Experimental [92] (Figure 1A). These databases serve as the top 5 most central nodes in the network, as measured by Betweenness centrality, and exhibit high connectivity, sharing a significant number of sequences. Among them, SATPdb stands out with connections to 3106 peptides, which account for 88.9% of the original subset, surpassing the next most central database (AVPdb) by a wide margin. The least connected databases are DRAMP_Clinical [153] and MilkAMP [154], with a node degree of 4 and 6, respectively.

While most sequences are associated with several databases, peripheral nodes connect to only one database, such as AVPdb, which contains 70 unique sequences visible in the bottom right corner of the network. This distinction is important as AVPdb specifically focuses on antiviral peptides, unlike SATPdb. Another example is the CyBase_Cyclotides [89], which connect to only 81 peptides but is noteworthy for including cyclic backbone peptides(Figure.5.1B).

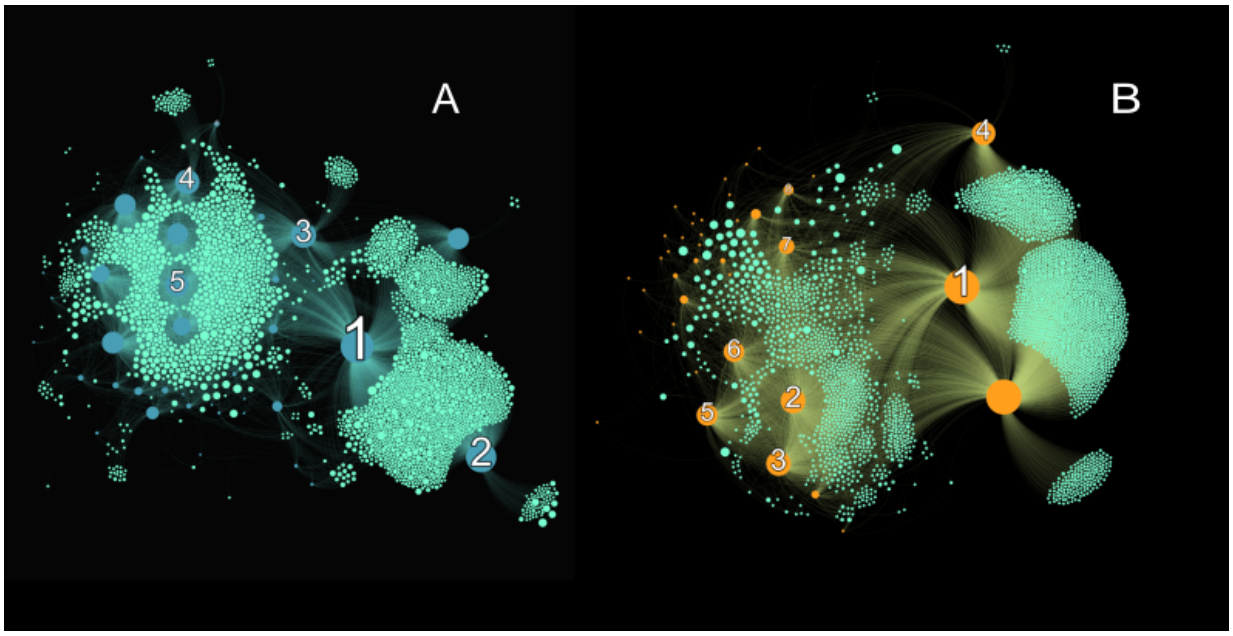


Figure 5.1: Metadata Networks (MN). (A) "Database" MN (B) "Function" MN Layout: Force Atlas2

Function MN: Exploring the relationship between antimicrobial peptides (AMPs) subclasses and AVPs reveals an intriguing approach. Beyond the AMP class itself, these associations can shed light on the peptides' action mechanisms and potential biological activities. The antiviral subclass was expected to be the most central node in the MN. Moving beyond that, the subsequent most central nodes from 1 to 6 include further subclassifications of antimicrobial peptides: Antimicrobial, Antibacterial, Antifungal, Anti Gram +, Anti Gram -, and Anti-HIV, in that order. The prominence of the anti-HIV function indicates the importance of targeted antiviral research specifically focused on HIV. Outside the antimicrobial category, the first two functions to appear are "toxic to mammals" and "hemolytic" suggesting a potential relationship between antiviral activity and the toxicity of these sequences. These observations align with findings reported in [78] regarding

hemolytic peptides. In contrast, the function with the fewest connections is "tumor-homing activity," which is only associated with one peptide(Figure.5.2A).

Origin MN: The Origin MN enables the association of selected peptide sequences with their respective sources, which range from synthetic constructs to isolates from various organisms. An interesting feature of this MN is its ability to link peptides with their origins and also relate the subcategories of origin to broader categories. This leads to additional edges labeled as "produced by" and "is a" with the latter relating to the subcategories of origin. In the "produced by" category, two main spheres are noticeable. The first sphere connects to the most central node representing "synthetic constructs", encompassing all reported synthetic AVPs. Although synthetic sources are the most common for AVPs, they only account for 13.5% of all sources, indicating the wide range of natural sources for AVPs. Some synthetic sequences have also been isolated from living organisms, primarily from *Homo sapiens*, *Bos taurus*, and *Rattus norvegicus*. A closer examination of this section reveals peptides that have been synthetically obtained and naturally isolated, such as StarPep_01104 and StarPep_00155, which have 22 reported sources each. The first peptide has 15 cross-references in StarPepDB and is reported as *mammalian tachykinin* peptide family [155]. The second one has 41 cross-references in StarPepDB, a peptide part of a transferase found in *Homo sapiens* and *Saccharomyces cerevisiae* [156]. In the outer sphere, sequences that have only been obtained from living organisms are found. These sequences are less central and more dispersed in comparison. The most common natural sources are the genus Homo and the family Homininae (Figure.5.2B).

On the other hand, the "is a" classification allows for identifying small clusters of taxonomically related categories. Additionally, some origins node do not appear to be directly related to any other origin node and are located on the periphery of the MN, such as *Macaca mulatta*, *Rana temporaria*, and *Odorrana andersonii*. This observation suggests the potential for further research on AVPs in species related to the mentioned ones.

Target MN: Like the previous MN, this complex network also consists of two types of edges labeled as "is a" and "assessed against". The "assessed against" edge is particularly useful for analyzing different targets. Two main sections can be observed: the inner circle,

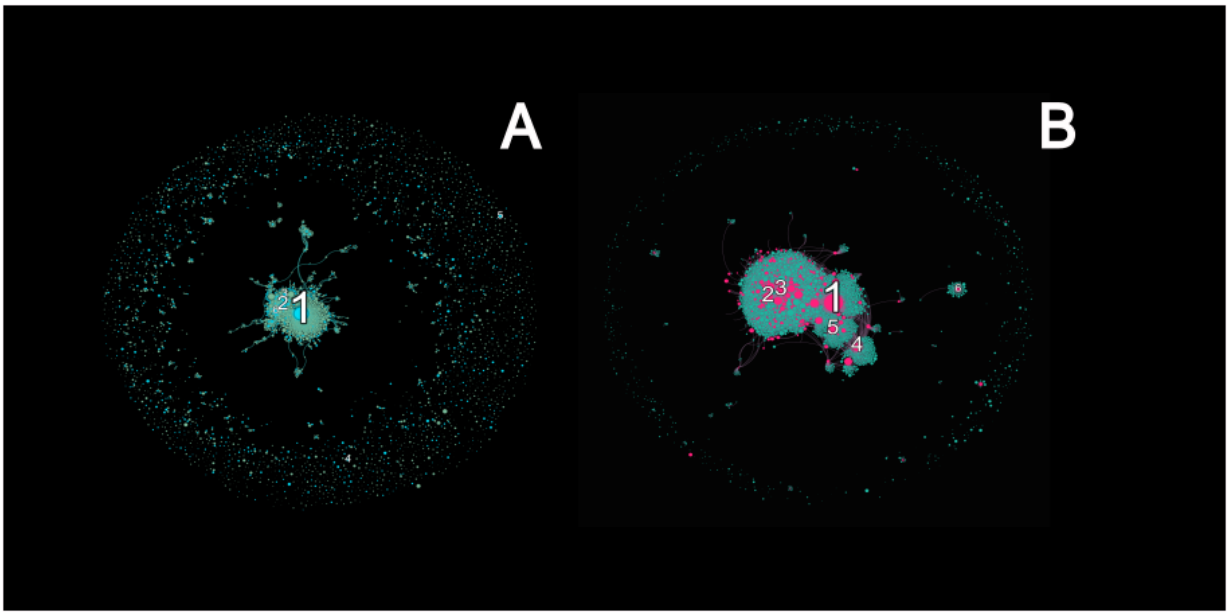


Figure 5.2: Metadata Networks (MNs). (A) "Origin" MN "produced by" edges (B) "Target" MN "assessed against" nodes Layout: Force Atlas2

which includes most nodes, and the outer ring, where nodes are more dispersed with limited connectivity. In the center, the most common targets for AVPs are HIV, Escherichia Coli, Staphylococcus Aureus, Hepatitis C, and Herpes Simplex. With 666 peptides connected to it, HIV represents 19% of the reported targets, emphasizing the importance of developing therapeutics against this specific virus. It is worth noting that the most common targets are viruses and bacteria, highlighting the close relationship between AVPs and AMPs. Another interesting observation from this section is the forming of a small cluster, situated between the inner and outer circles, with the most central node being the target Andes Virus. The outer ring includes peptides that are not specifically associated with any target but rather represent a taxonomic classification of a larger group of organisms. In the "is a" MN, the taxonomical relation and relative proximity of each target to the others are displayed.

5.2 Half Space Proximal Networks

In HSPNs for AVPs, the Euclidean metric serves as the most intuitive similarity measure, commonly employed to calculate distances in many applications. In an n-dimensional space (n molecular descriptors), distances between nodes are computed, and these dis-

tances are then aggregated into a 'similarity matrix'. While the visual interpretation of a similarity matrix leads to a similarity network, the conversion between the two requires the application of a threshold matrix. The selection of a similarity threshold determines the preservation of edges connecting different peptides (nodes) [47]. This study systematically varied the similarity threshold (t) was systematically varied from 0.3 to 0.9. Additionally, an HSPN without a threshold was included to facilitate a later comparison between networks with and without the threshold. The choice of the similarity threshold directly influences the density of connections within the complex network. While the number of nodes remains constant across all networks, the number of edges varies. Although this may appear to be a minor change, it significantly impacts the formation of communities and the presence of singletons (nodes that lack connections to any other node) within the network.

Once all the HSPNs with different similarity threshold (t) values were constructed and the topology was characterized using several parameters obtained in Gephi software . The different HSPNs constructed are displayed in Figure.5.3 Several parameters were considered in this analysis, including the number of edges, number of communities, number of singletons, density, modularity, ACC, average path length, and average degree, gathered in Table. 5.1

Figure.5.5 depicts the density, ACC, and modularity of each HSPN for all t used. These parameters were utilized to establish an HSPNs' characterization and discovery, which t value is optimal for representing AVPs. The significance of this study lies in the fact that selecting the most optimal t value is not an automated process; it requires careful consideration by researchers. Choosing the correct t value enhances complex network visualization and facilitates a better understanding of the relationships between communities. An HSPN with just one large community may not be desirable, just as a network with a high count of singletons can obscure important similarity information. Therefore, finding the right balance is crucial for obtaining a meaningful representation of AVPs and their subsequent exploration [47].

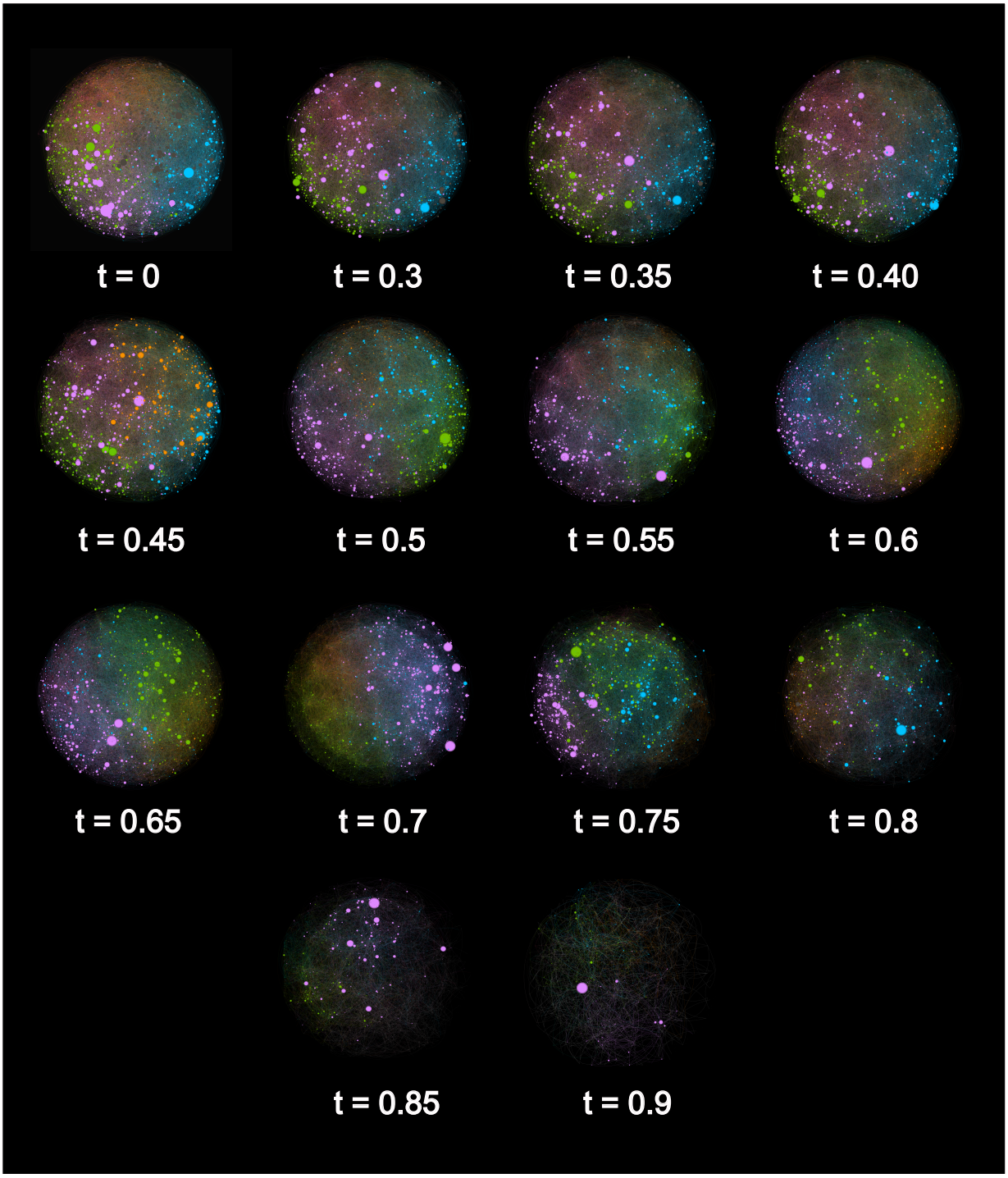


Figure 5.3: Comparison of HSPN visual density depending on t value

Table 5.1: Topology Characterization of HSPNs fro t = 0.3-0.9

t	Edges	Clusters	Singletons	Density	Modularity	ACC	Av. Path	Av.Degree
0	27725	10	0	0.005	0.463	0.024	7	3.591
0.3	27724	8	0	0.005	0.473	0.024	7	3.591
0.35	27722	8	0	0.005	0.472	0.024	7	3.592
0.4	27716	8	0	0.005	0.464	0.024	7	3.593
0.45	27698	7	0	0.005	0.476	0.024	7	3.596
0.5	27667	7	0	0.005	0.472	0.024	7	3.6
0.55	27564	8	0	0.005	0.478	0.024	8	3.611
0.6	27340	10	3	0.004	0.467	0.024	8	3.631
0.65	26676	15	9	0.004	0.48	0.025	8	3.685
0.7	25052	21	13	0.004	0.486	0.026	10	3.836
0.75	21474	74	75	0.004	0.518	0.027	18	4.305
0.8	14844	307	509	0.002	0.58	0.026	16	4.849
0.85	5739	1075	1726	0.001	0.79	0.024	23	6.449
0.9	1641	2356	3388	0	0.952	0.023	12	3.829

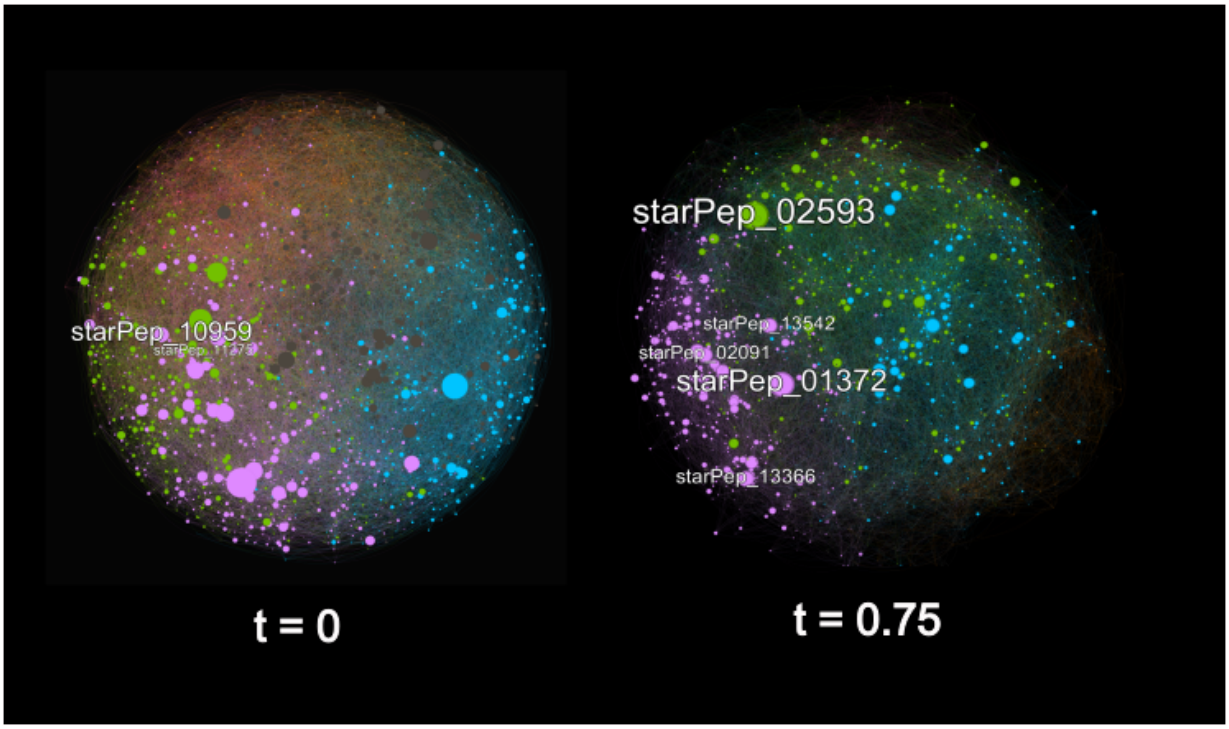


Figure 5.4: (I) t = 0 and (II) t = 0.75. A different color is assigned in each HSPN to show communities. Layout: Fruchterman Reingold

However, the construction of the HSPN presents a unique characteristic that sets it apart from traditional networks: it doesn't necessarily require a similarity threshold yet remains sparse. This attribute makes the network parameter-free, offering a significant

advantage as there is no need to determine an optimal similarity threshold. In this study, both the threshold-free network and the network with the optimal threshold are used in parallel. This approach enables a comprehensive comparison of the similarities and differences discovered, allowing for a more thorough analysis of the network dynamics. By utilizing both network versions, the study gains a broader perspective, enhancing the overall understanding of the HSPN network's characteristics

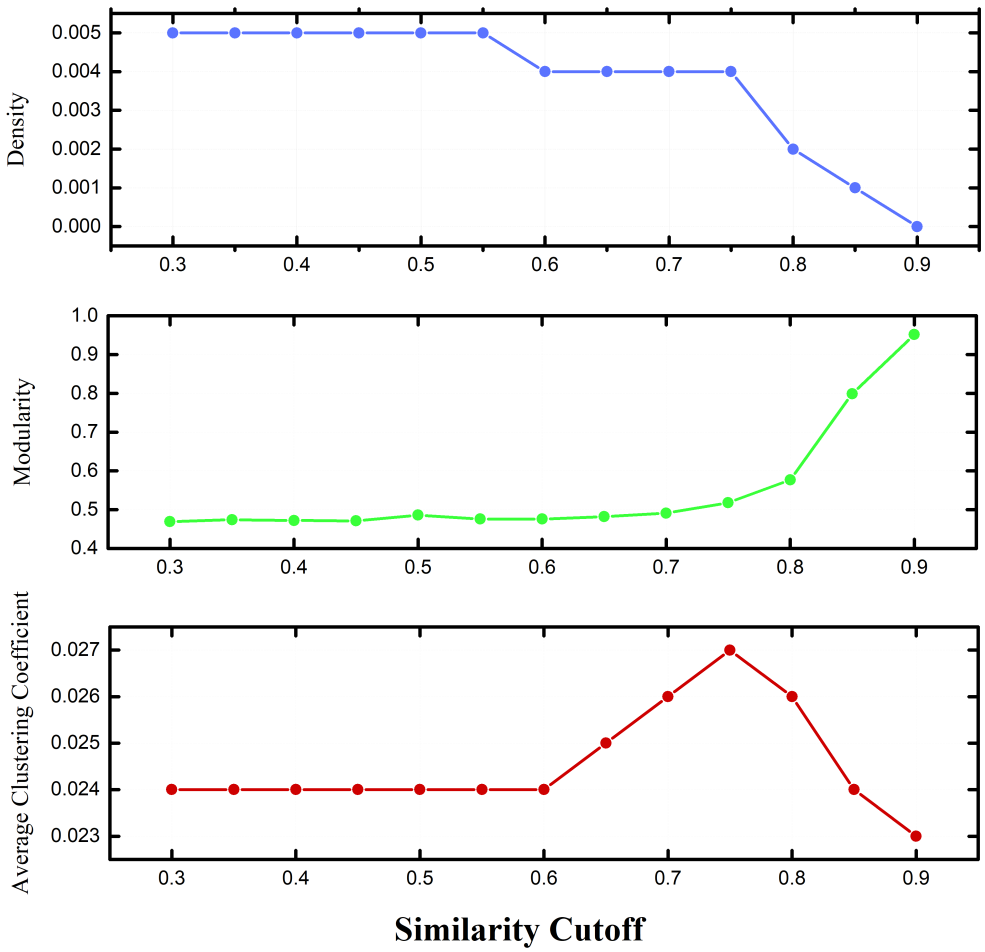


Figure 5.5: HSPN's characterization using different parameters.

The density of a network is expected to decrease as the number of edges satisfying the similarity threshold condition decreases. On the other hand, modularity is a parameter that reflects the presence of communities, and maximizing modularity is a goal of the Louvain clustering algorithm. As the t value increases, there is an increase in modularity, indicating the formation of fewer but more cohesive communities. This increase is partic-

ularly noticeable after t = 0.8. However, the correlation between ACC and t is not linear, as an increase in t does not necessarily lead to an increase in ACC. When plotting the different ACC values, a maximum is observed around t = 0.75, indicating the highest level of connectivity.

Based on these observations, the cut-off value of t = 0.75 was determined as the optimal choice. Figure.5.4 illustrates the HSPN at this cut-off (HSPN_OP) alongside the HSPN at t = 0 (HSPN_NC). Additionally, Figure.5.6 displays the graphical representation of the Louvain Clustering Algorithm, showcasing the 8 communities obtained from HSPN_NC. Each community is depicted in a distinctive color, and the size of the nodes reflects their centrality.

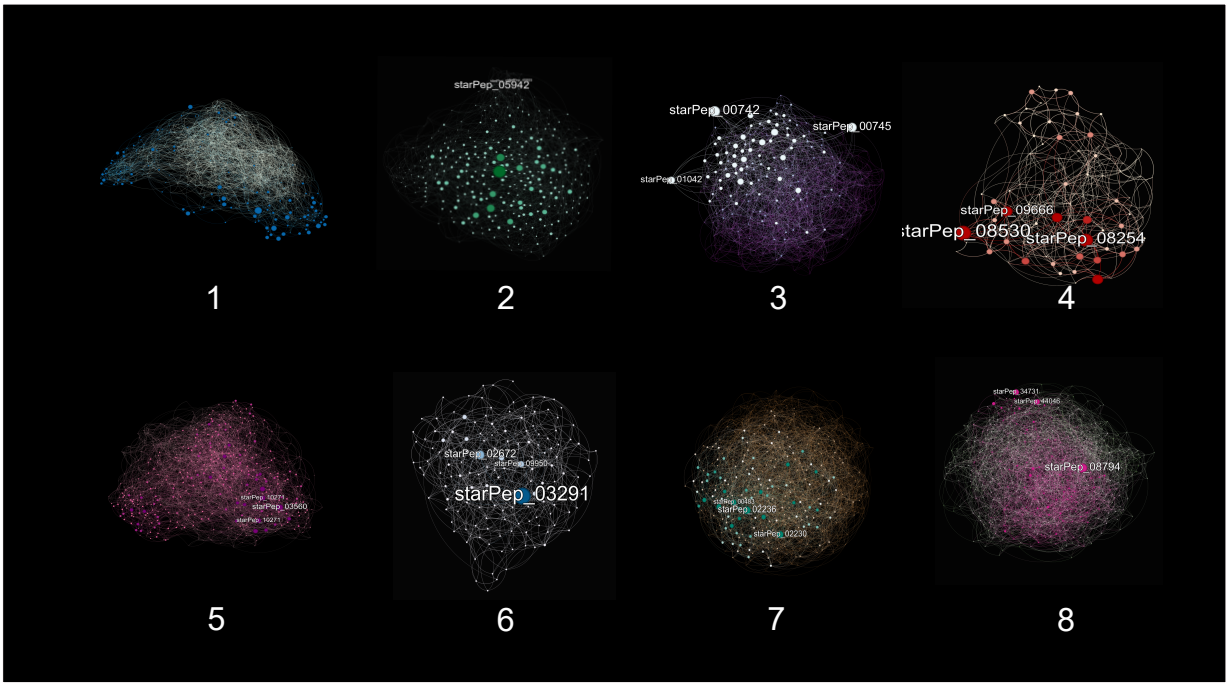


Figure 5.6: HSPNs (t = 0) for each 8 communities obtained by using the Louvain algorithm Layout: Force Atlas2

These communities range from 100 nodes (cluster 4) to 652 (cluster 1). It is worth noting that these communities can be further subdivided and characterized similarly to the whole-space networks, due to their still considered size. In a subsequent analysis, the most representative sequences within these 8 clusters, labeled in the figure, will be examined. A community analysis was not performed to HSPN_OP since the number of clusters extracted by the Louvain clustering algorithm is considerably larger (74 clusters) and the

number of singletons (75 singletons) is also high hindering a more rounded interpretation of the communities.

For an additional representation of the networks, the degree distribution was plotted noticing a higher degree distribution for the HSPN_NC, where HSPN_OP has a higher count of singletons and nodes with lower connectivity as it was aforementioned. The degree distributions along with a normal approximation are shown in Figure.5.7

The most central nodes (determined by HB centrality) were selected and further analyzed to delve deeper into the chemical space of these complex networks. These selected sequences were characterized using various molecular descriptors, allowing for a detailed examination of their chemical properties (Table.5.2). Additionally, a classification of the different AAs present in each sequence was performed using the "Peptides" package for R.

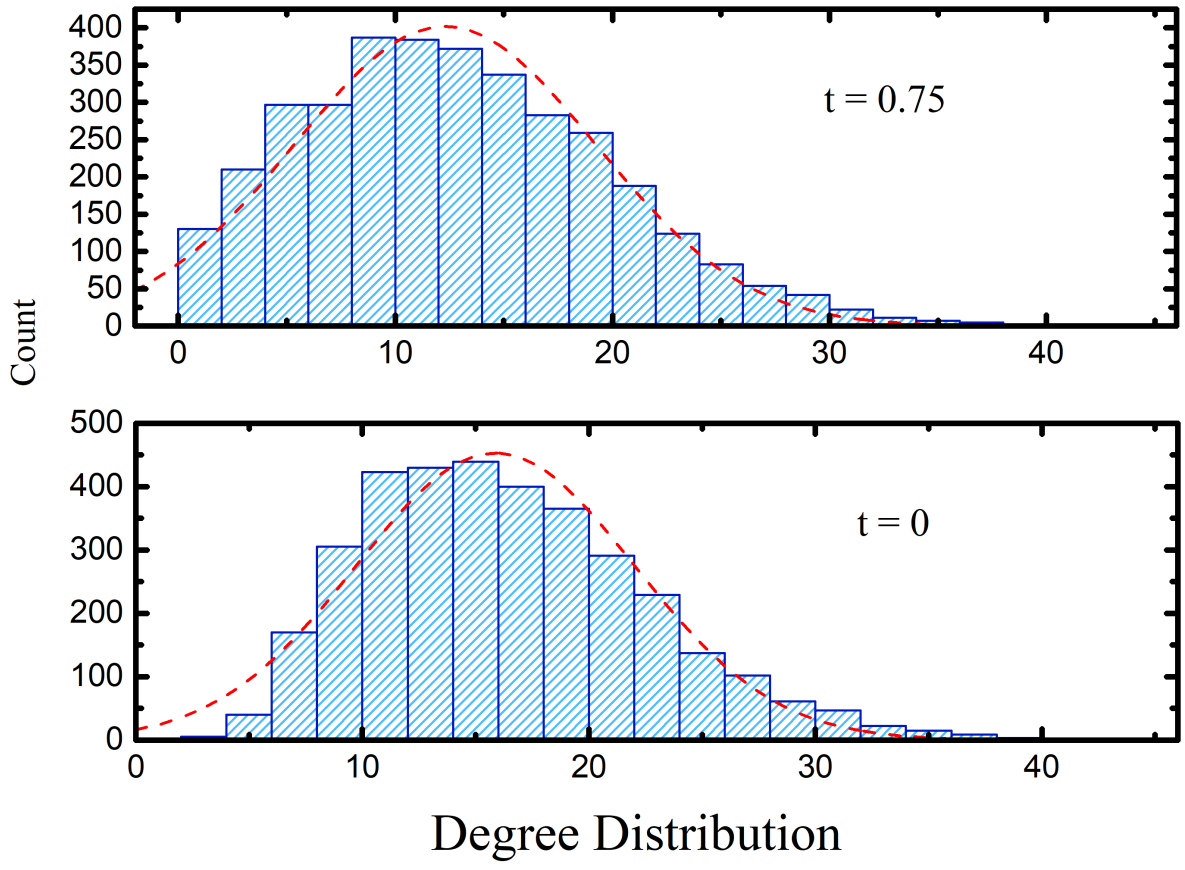


Figure 5.7: HSPNs' Degree distribution. (A) t = 0.75, HSPN with optimal similarity cutoff. (B) t = 0, HSPN with no cutoff (free parameter). The dashed red line indicates the normal fit for the respective distribution.

Upon careful examination of the central nodes within each HSPN, a noteworthy obser-

vation comes to light: the peptides represented in the network are not identical. However, this does not mean that the nodes highly central in HSPN_NC are absent in HSPN_OP. In fact, upon further analysis, it is revealed that the most central nodes in HSPN_OP, as shown in Table.5.2, are present in clusters 19 and 9, which correspond to clusters 1 and 3 in HSPN_NC. This emphasizes the significance of these particular clusters in terms of connectivity. Additionally, the most central peptides in HSPN_NC are found in three distinct clusters (1, 5, 8), these same peptides are present in HSPN_OP in clusters 9, 18, and 31. This demonstrates that, regardless of the HSPN considered, the same sequences are consistently grouped. However, it also highlights how HSPN_OP further fractures the space, making it more challenging to comprehend the distances between different groups of sequences. Despite this complexity, the underlying connectivity patterns remain preserved across the networks, and cluster 1 is considerably important in terms of centrality as it is also the largest one.

Table 5.2: Most Central sequences of each HSP and corresponding chemical features

IDa	Cluster	Aliphatic Index	Boman Index	Hydrophobicity	Isoelectric Point	Charge	Length
HSPN t = 0.75 (HSPN_OP)							
StarPep_02593	19	78.00	0.42	0.34	7.99	1.69	30
StarPep_01372	9	62.86	2.07	-00.90	12.25	8.09	28
StarPep_13366	9	86.67	3.56	-00.95	11.28	2.09	18
StarPep_02091	9	66.67	1.37	-00.01	8.17	3.50	63
StarPep_13542	9	91.00	1.53	-00.14	10.21	2.09	30
Mean (±SD)		77.04 (±10.93)	1.79 (±1.03)	-00.33 (±0.51)	9.98 (±1.68)	3.49 (±2.38)	33.8 (±15.26)
HSPN t = 0 (HSPN_NC)							
StarPep_02526	5	97.50	0.34	0.43	11.90	6.00	20
StarPep_08887	8	43.33	1.47	-00.39	8.16	1.06	9
StarPep_10907	1	46.52	3.66	-1.56	11.09	5.94	23
StarPep_01472	1	65.00	2.12	-00.07	11.16	5.75	30
StarPep_10501	5	76.11	1.43	-00.50	7.02	0.00	18
Mean (± SD)		65.69 (±19.94)	1.81 (±1.09)	-00.42 (±0.65)	9.87 (±1.91)	3.75 (±2.65)	20 (±6.84)

Jumping into a more chemical analysis of the AVPs space, one notable difference observed among the molecular descriptors used is the Aliphatic index, which is lower for

HSPN_OP than HSPN_NC. Additionally, the Hydrophobicity values are also lower for HSPN_OP. Furthermore, the average length of the most central peptides in HSPN_NC is approximately 13 residues longer than those in HSPN_OP. On the other hand, the differences between the two networks in terms of other properties such as Boman Index, Isoelectric Point, and Charge are less than 10%. These differences are best depicted in Table.5.2

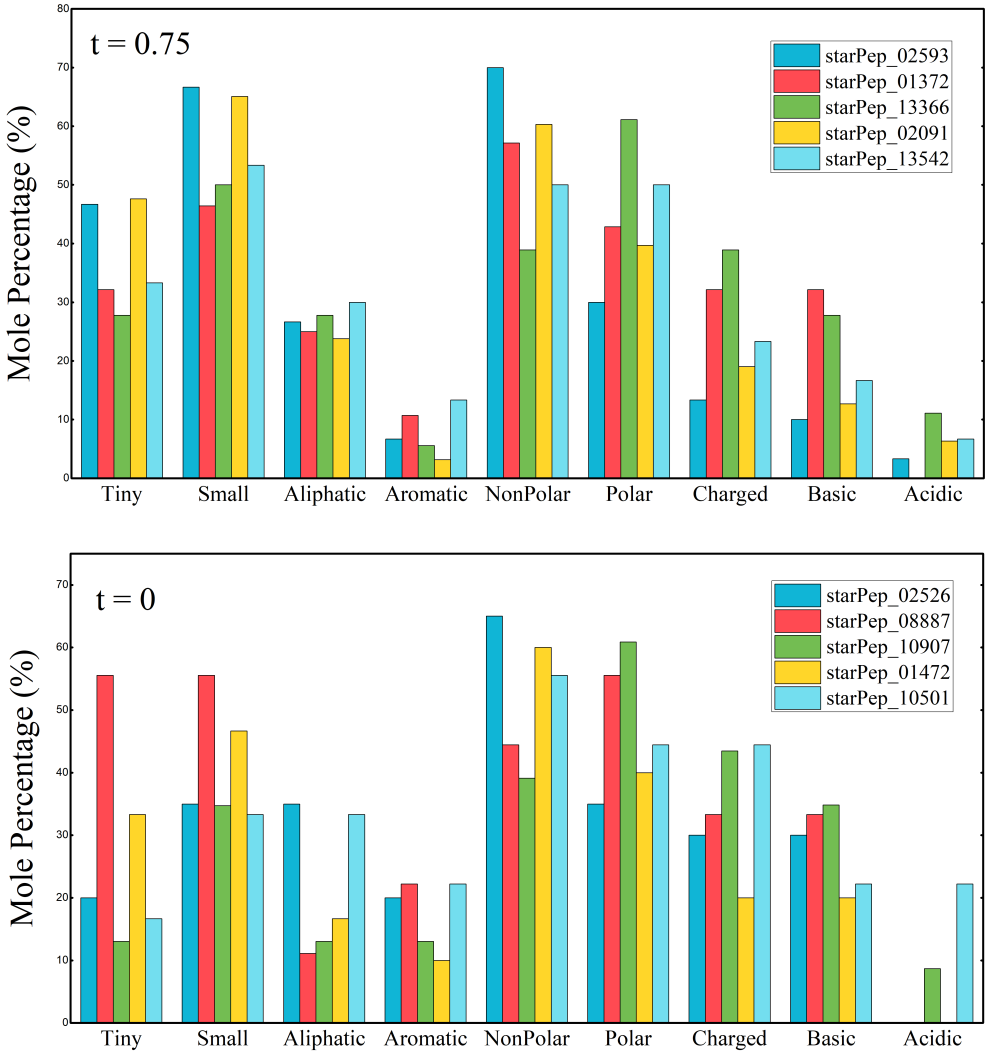


Figure 5.8: Occurrence of different types of AAs corresponding to the five most central nodes (Harmonic Centrality) of each HSPNs selected

In terms of AA representation, it can be observed that HSPN_OP exhibits a higher level of homogeneity in its representative peptides compared to HSPN_NC. This difference

is probably linked to the fact that many of the most central peptides in HSPN_OP are from the same cluster compared to HSPN_NC. Specifically, HSPN_NC demonstrates a larger mole percentage of aromatic, charged, and basic AAs. Conversely, HSPN_OP shows a more stable presence of AAs categorized as tiny, aliphatic residues, and acidic. However, both networks exhibit the highest representativity of non-polar AAs over any other category (Figure.5.8).

Furthermore, an extensive literature search was conducted to gain insights into the mechanisms of action associated with these AVPs. While many of these sequences have numerous cross-references in StarPepDB, only one reference was selected for each case to avoid excessive lengthening of this section. Focusing on a representative reference for each sequence makes the analysis more concise and manageable.

Most Central Nodes for HSPN_OP

- StarPep_02593. This peptide has been reported as Cycloviolacin-O17, with a sequence "GIPCGESCVWIPGISAAIGCSCKNKVCYRN", and extracted from Viola odorata (Sweet violet). Anti-HIV, antibacterial, and hemolytic biological activities for these cyclotides have been reported. It has also studied the relation between the cyclotide framework and the proteolytic stability of these peptides [157].

- StarPep_01372.It is reported as Kenojeinin I, with a sequence "GKQYFPKVGGRLSGKAPLAAKTHRRLKP" and has been isolated from the skin of a fermented skate. It has many cationic residues and shows antimicrobial activity. The researchers that purified this peptide suggest that the basic AA facilitates binding of and transport across the bacterial outer membrane, and the hydrophobic residues cause the inner membrane's disruption and permeation [158].

- StarPep_13366. This peptide is part of a Genome polyprotein, with a sequence "VATRDGKLPTTQLRRHID". It is related to the Hepatitis C virus (HCV) genotype 1a. This peptide is linked to the envelope proteins of E1 and E2 of HCV. Characterization of the E1 and E2 heterodimer suggests that the functional complex's pre-budding form of the functional complex, which will probably play an active role in the entry into host cells [159].

- StarPep_02091. It is a Ascaris suum antibacterial factor type antimicrobial peptide, named by the researchers abf-2. Abf-2 has the following sequence, "DIDFSTCARMDVPILKKAAQGLCITSCSMQNCGTGSCKKRSGRPTCVCYRC ANGGGDIPLGAL". This peptide was tested to have biological activity against Gram-positive and Gram-negative bacteria, and yeast [160].

- StarPep_13542. This peptide is a genome polyprotein, with a sequence "VSRRYLASLHKKALPTSVTFELLFDGTNPS", and has been linked to the envelope glycoproteins involved in cell infection of classical swine fever virus [161] .

Most central nodes for HSPN_NC:

- StarPep_02526. This peptide "FLFRVASKVFPALIGKFKKK" is referred to as D51, and it was designed to have antimicrobial properties by using a linguistic model of natural AMPs based on amphipathic properties; the peptide was tested against Gram-positive and Gram-negative bacteria [162].

- StarPep_08887. This peptide "CSLHSHKGC was reported as a cyclic peptide that inhibits the Andes Virus infection. The technique used for identifying this peptide was phage display using a cysteine-constrained cyclic nonapeptide-bearing library [163]

- StarPep_10907. Described as Envelope glycoprotein gp150 synthetized for the inhibition of the Feline immunodeficiency virus, with a sequence "KQRNRWEWRPDFKSKKVKISLPC". Although the action mechanism isn't clear, the author suggests that the inhibitory peptides may act by interacting with cell-surface molecules involved in viral infection [164].

- StarPep_01472. This peptide "IRNSLTCRFNFGICLPKRCPGRMRQIGTCF" was isolated from Cervus elaphus blood (deer), and it showed antimicrobial activity, especially against Gram-negative bacteria [165].

- StarPep_10501. This amphipathic helix-containing peptide was designed to interfere with HIV envelope glycoprotein and interfere with the steps involving membrane fusion. The sequence is "KAFEEVLAKKFYDKALWD" [166].

Similar to the observations made in the MNs analysis, there is a close relationship between the antiviral function and the broader classification of antimicrobial peptides. Out of the 10 sequences mentioned earlier, 3 primarily target Gram-positive and Gram-negative bacteria. Another notable similarity among these sequences is that both HSPNs contain at least one central peptide that targets HIV, the most common target in the "target" MN.

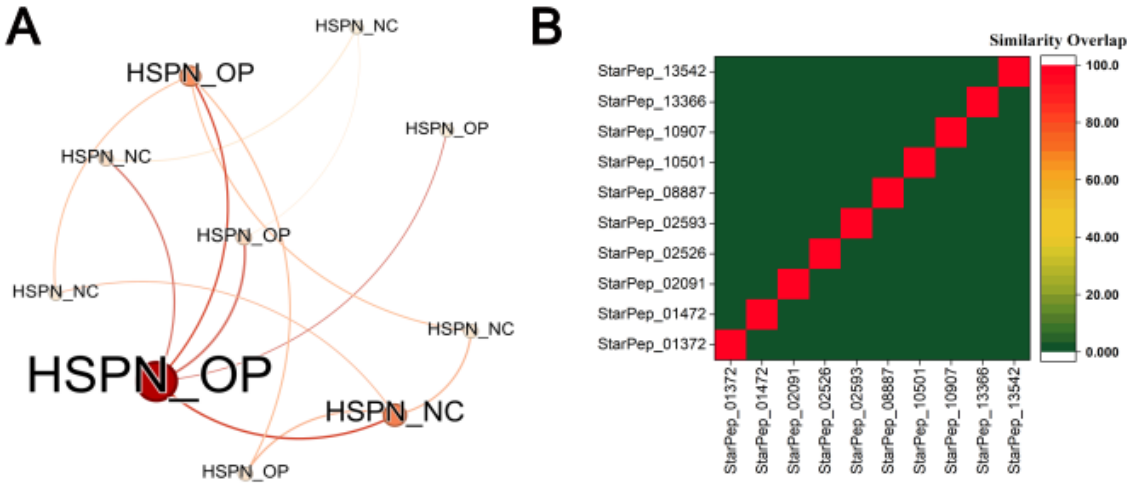


Figure 5.9: (A) Similarity Network between the most central nodes from HSPN_OP and HSPN_NC. Layout: Fruchterman Reingold (B) Similarity Overlap between the 10 top sequences of each HSPN

However, these sequences have additional underlying relationships and similarities to shed light on. A similarity network can be constructed by considering only these ten sequences. The network found in Figure.5.9A reveals sparse connections, with most nodes only connecting to two other sequences. The peptide that exhibits the highest number of connections is the StarPep_02091 sequence, previously identified as one of the most central nodes in the HSPN_OP. This can be attributed to the fact that it is the longest peptide among the selected sequences. To explore the study of sequence similarities in more depth, each of these peptides underwent analysis using Dover Analyzer (41) to generate a similarity overlap heatmap. Given the expectation of distinct characteristics among these sequences, a local type alignment algorithm was utilized to measure their similarity. As depicted in Figure.5.9A, these sequences demonstrate minimal compositional similarities.

Thus, it becomes imperative to delve further into the intracommunity to obtain a more comprehensive understanding of the nature of AVPs' chemical space. This analysis will be conducted at a later stage.

5.3 Scaffold Extraction

The primary objective of this section is to create a well-represented subset of the overall AVP chemical space. A total of 20 scaffolds were obtained from HSPN_OP and HSPN_NC by varying the alignment algorithm, centrality measure, and similarity threshold. The scaffold extraction tool available in the subnetwork mining section of the StarPep toolbox was utilized for this purpose. The scaffold extraction process allows for simplifying the network's topology based on a pre-defined similarity parameter. As a result, the general representation, and characteristics of the AVP chemical space are preserved, while significantly reducing the complexity of the network.

For the comparison of the scaffolds, Dover Analyzer software was utilized. Among the various computed results provided by that software, the primary focus was on the analysis of similarity overlaps between the scaffolds. These similarity overlaps, expressed as percentages, indicate the extent to which sequences are repeated when compared pairwise. The results of this comparison are visualized in the form of a heatmap graphic, where 8 different scaffolds are compared at the time, changing sourcing HSPN, alignment algorithm, and centrality measure. One notable observation from these heat maps is that they are not symmetrical. This lack of symmetry arises from an imbalance in the number of sequences in each scaffold. While two scaffolds may have the same number of redundant sequences, the redundancy percentage may vary for a smaller subset, representing a higher proportion of its total.

Table 5.3: Characterization of Scaffolds from HSPN_NC varying the alignment algorithm and Centrality Measure.

HB centrality				HC measure			
Identity Percent	Edges	Nodes	Coverage (%)	Identity Percent	Edges	Nodes	Coverage (%)
Local Alignment							
90	22,343	2,996	86	90	22,229	3,003	86
80	16,396	2,363	68	80	16,027	2,369	68
70	12,820	2,044	59	70	12,764	2,028	58
60	8,108	1,536	44	60	8,395	1,557	45
50	3,633	950	27	50	4,530	1,030	29
Global Alignment							
90	23,768	3,123	89	90	23,836	3,124	89
80	18,585	2,566	73	80	18,569	2,560	73
70	15,612	2,278	65	70	15,674	2,273	65
60	13,004	2,007	57	60	13,132	2,005	57
50	8,721	1,587	45	50	8,798	1,582	45

Table 5.4: Characterization of Scaffolds from HSPN_OP varying the alignment algorithm and Centrality Measure.

HB centrality				HC measure			
Identity Percent	Edges	Nodes	Coverage (%)	Identity Percent	Edges	Nodes	Coverage (%)
Local Alignment							
90	16,997	3,005	86	90	17,015	3,005	86
80	12,801	2,368	68	80	12,819	2,369	68
70	10,221	2,022	58	70	10,504	2,046	59
60	6,817	1,534	44	60	7,212	1,559	45
50	4,110	1,034	30	50	4,311	1,044	30
Global Alignment							
90	18,442	3,119	89	90	18,397	3,126	89
80	14,832	2,562	73	80	14,667	2.566	73
70	12,620	2,277	65	70	12,410	2,006	65
60	10,669	1,991	57	60	10,564	2,006	57
50	7,529	1,589	45	50	7,287	1,592	46

When comparing the various variables set for this experiment (Figure.5.10), a notable observation is that the section of the heatmaps comparing scaffolds derived from different parent HSPNs consistently shows a predominantly red color. This indicates that there is no

significant difference in representativeness and diversity between these two HSPNs. This finding holds important implications, suggesting that in future research involving HSPNs and AVPs, there may not be a need to extensively explore the network topology in search of an optimal similarity threshold (t-value). This result holds significance for the subsequent section of the study.

The difference in alignment algorithm yields the most noticeable effects, particularly in scaffolds characterized using different centrality measures. It is important to emphasize that the HB centrality adopts a more localized or community-based approach, whereas the HC measure is based on the shortest path between nodes. As a result, the choice of alignment algorithm has a lesser influence on scaffolds obtained using HC, and its influence becomes apparent only when the sequence's % identity is reduced to 60%. All the scaffold analysis is summarized in Table.5.3 and Table.5.4.

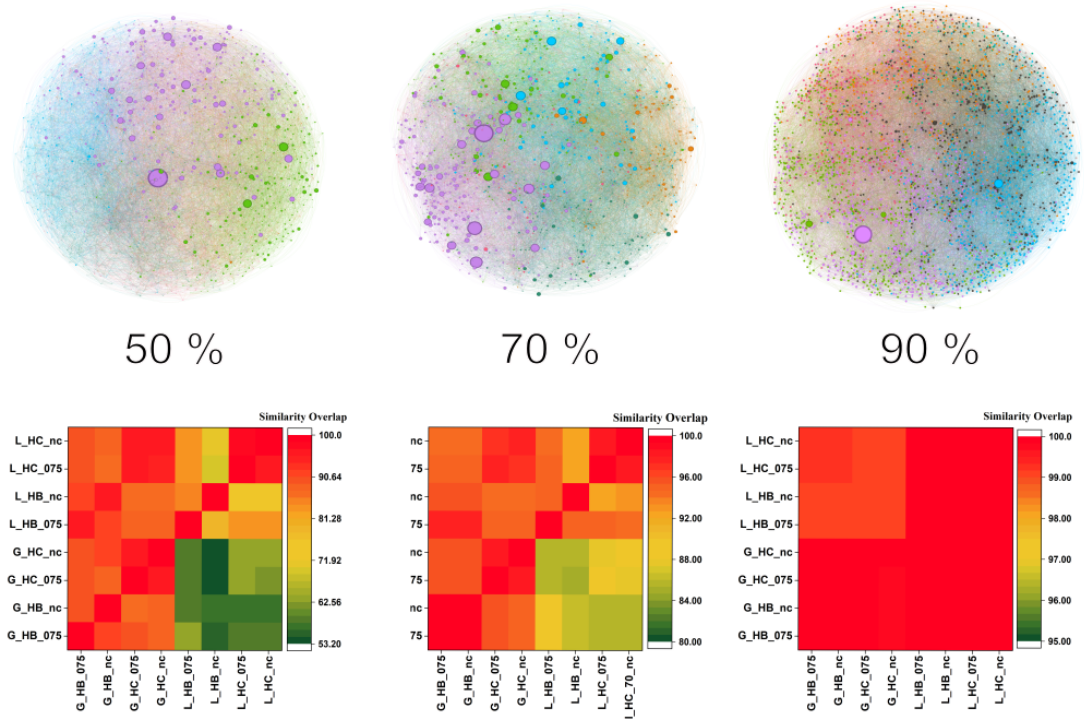


Figure 5.10: Scaffold Density Comparison and Similarity Overlap

Based on these findings, a set of scaffolds was carefully selected to construct representative and diverse subsets for various similarity cutoffs. Four merged subsets were constructed

from 80-50% of pairwise sequence's identity containing 2703, 2445, 2152, and 1872 positive sequences respectively. These subsets, consisting of representative sequences, are now being presented to the chemoinformatics/bioinformatics community as a condensed subset of the vast chemical space of AVPs. These results hold significant value as they can serve as training datasets for machine learning-based predictors and/or facilitate multi-query similarity searching. The primary advantage of these scaffolds lies in their high diversity and non-redundant representativity, which effectively prevent overfitting and family over-representation in predictive models. By utilizing these scaffolds, researchers can develop accurate prediction models, and effectively explore novel AVP candidates. These resources contribute to advancing AVP research and the development of effective antiviral therapies

5.4 Motif Discovery

The diversification of peptide sequences is immeasurable; however, the changes in the sequence must be within the limits of functionality. As a result, certain AAs tend to remain unaltered due to their direct involvement in the activity. When these conserved residues are identified across various sequences, they are called motifs, indicating an orthologous relationship [167]. The term "motif" is used in various contexts, often denoting a specific repetitive element. In a biological context, these elements represent recurrent patterns in biological entities.

To enhance our understanding of natural networks beyond their topology, it is valuable to identify motifs, which are significant patterns of interconnections. Network motifs refer to elements with a higher probability of appearing than to a randomly generated network [168]. In this case, the motifs will consist of AAs sets containing between 3 and 6 residues. Peptide motifs offer a significant advantage as compact building blocks with functional autonomy, potentially leading to more effective AVPs [169]. However, identifying linear motifs poses challenges due to their short length, the tendency to occur in different sequence sections, and limited conservation across unrelated species. Moreover, discovering these motifs is elusive using experimental techniques and often requires time-consuming experiments [170].

The search focused on the clusters obtained through modularity optimization of HSPN_NC

to identify antiviral motifs. Only the communities within this network were utilized, as the scaffold analysis from the previous section revealed no significant difference between HSPNs with and without an optimal cut-off. The graphical representation of each cluster can be found in Figure.5.6.

As mentioned earlier, an intracommunity analysis of the chemical space is required. The three most central nodes for each cluster are listed in Table.5.5, along with their sequences and relevant external information regarding their identity. It is worth noting that certain clusters exhibit specific similarities among their nodes. For instance, in clusters 1 and 2, the listed peptides are derived from plants and possess a cyclotide structure [171, 172, 173, 174, 175]. Nodes in clusters 3 and 6 share a notable similarity as they originate from amphibious sources [176, 177, 178, 179, 180]. Lastly, cluster 4 comprises nodes interacting with HIV [181, 182]. The remaining clusters do not exhibit such obvious biological relationships among their central nodes.

For the chemical characterization of these peptides, a distinct chemical profile was developed for each cluster. To facilitate comparison, the changes between clusters based on molecular descriptors are presented. Several important observations can be made from this comparison, see Figure. 5.11.

Table 5.5: List of Most Central AVPs corresponding to each Community in HSPN.

Cluster	Name	Sequence[a]	Reference
1	StarPep_02860	RTCMIKKEGWGK CLIDTTCAHSCKNRGYIGGDC KGMTRTCYCLVNC	Part of a plant defensin extracted from *Vigna radiata [171]*
	StarPep_00566	AACSDRAHGHI CESFKSFCKDSGR NGVKLRANCKKTCGLC	Antimicrobial peptide from *Aurelia aurita* with defensin feature [173]
	StarPep_02843	RECKTESNTFPGICI TKPPCRKACISEKFTDGHC SKLLRRCLCTKPC	Part of a floral defensin f rom *Nicotiana tabacum [172]*
2	StarPep_05942	ICGETCVGGTCNTPG CSCSWPVCTRNGLP	Plant cyclotide [174]
	StarPep_01071	GLPICGETCVGGTCN TPGCSCSWPVCTRN	Varv peptide E from *Viola arvensis [175]*
	StarPep_40805	TCVGGTCNTPGCSCSW PVCTRNGLPICGE	Produced by *Viola arvensis (StarPep DB)*
3	StarPep_00742	GVFTLIKGATQ LIGKTLGKELGKT GLEIMACKITKQC	Antimicrobial peptide extracted from Chinese odorous frog [176]
	StarPep_00745	GVFTLIKGATQ LIGKTLGKEVGKTG LELMACKITKQC	
	StarPep_01042	GLFPKINKKKA KTGVFNIIKTVGKEAGM DLIRTGIDTIGCKIKGEC	Antimicrobial peptide obtained from pickerel frog [177]
4	StarPep_08530	ATKALTE VIPLTEEAEC	Inhibitors targeting HIV-1 reverse transcriptase [181]
	StarPep_08254	AEAIPMSIP PEVKFNKPFVF	HIV-1 entry inhibitor [182]
	StarPep_09666	GAKALTE VIPLTEEAE	Inhibitors targeting HIV-1 reverse transcriptase [181]

Table 5.6: List of Most Central AVPs corresponding to each Community in HSPN.

Cluster	Name	Sequence[a]	Reference
5	StarPep_00500	HSDAVFTDNYTRLRKQ MAVKKYLNSILN	Vasoactive intestinal peptide [183]
	StarPep_13041	SQGVVESMNKELKK IIGQVRDQAEHLKTAY	Synthetic peptide from HIV type 1 integrase [184]
	StarPep_03560	QARSDIEKLKEAIRDTN KAVQSVQSSIGNLIVAIK	Fusion glycoprotein F0 related to Human parainfluenza 3 virus [185]
6	StarPep_03291	ILGAILPLVSGLLSSKL	Antimicrobial peptide from the skin secretions of the midwife toad [178]
	StarPep_02672	ILGAILPLVSGLLSNKL	Analogue of the frog skin peptide [179]
	StarPep_09950	GLVGTLLGHIGKAILGG	Antibacterial peptide from Skin Micro-Organs of the Orinoco Lime Treefrog [180]
7	StarPep_02236	GLWSKIKEAAKTAGKMAMGFVNDMV	Antimicrobial peptide from *Phyllomedusa distinta* [186]
	StarPep_02230	GLRSKIKEAAKTAG KMALGFVNDMA	Antimicrobial peptide Dermaseptin S9 [187]
	StarPep_00483	GLWSKIKEAA KAAGKAALNAVTGL VNQGDQPS	Antimicrobial peptide from *Phyllomedusa distinta* [186]
8	StarPep_08794	CNSHSPVHC	Cyclic peptide for Andes Virus inhibition [163]
	StarPep_34731	NXXLYSARGARGH	Aniviral/Antimicrobia (StarPep DB)
	StarPep_44046	XNXLYSARGARGH	

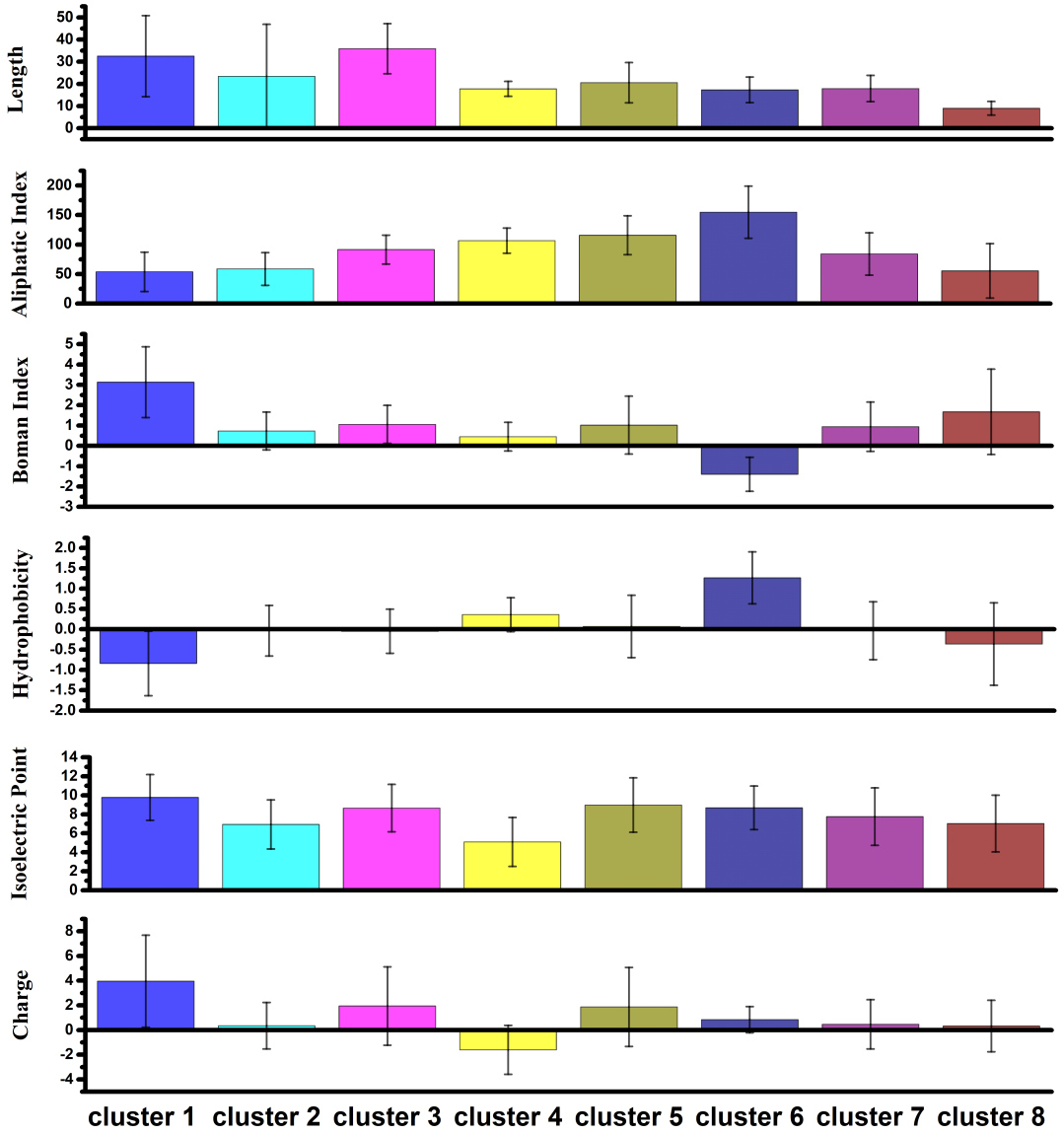


Figure 5.11: Chemical Characterization of each cluster

Firstly, Cluster 4 exhibits a negative charge on average, indicating a higher prevalence of acidic AAs, while Cluster 1 has the highest positive charge, indicating the opposite trend. Clusters 2, 7, and 8 are nearly neutral in charge. In terms of length, the longest sequences are found in Clusters 1 and 3, while the shortest sequences are present in Cluster 8. The aliphatic index is highest in Cluster 6 and lowest in Clusters 1, 2, and 8. The Boman index

is negative for Cluster 6, indicating a potential antimicrobial activity, whereas Cluster 1 has the highest Boman index value. Hydrophobicity is notably high in Cluster 6 and negative in Cluster 1. The isoelectric point shows relative homogeneity within each cluster, with Cluster 4 having the lowest value due to its negative charge. These properties are very important for the anti-viral ability of the peptide. Typically, AVPs are cationic and contain hydrophobic residues, the hydrophobicity and electrostatic interaction are important factors when fighting enveloped viruses [188].

This chemical information also provides insights into the similarity relations and communities within the chemical space. For example, clusters 1 and 2, which have relatively similar sources, may initially appear to belong to the same cluster. However, a closer examination of their distinct chemical properties reveals significant differences, justifying their formation as separate communities.

Once the community analysis was completed, the STREME algorithm from the MEME suite was employed for *an de novo* fixed-length motif discovery [140]. When provided with a dataset, STREME generates a control set by shuffling the letters of the primary sequences while preserving lower-order statistics. This process helps to identify only relevant motifs and avoids the discovery of non-relevant ones. STREME also provides significance statistics for each discovered motif, which is determined by comparing its occurrence in the primary sequences to that in the control set. The primary sets used for motif discovery consisted of the sequences found within the 8 clusters of HSPN_NC.

The validation process of the 42 motifs discovered by STREME involved two stages. In the first stage, the SEA algorithm [141] from the MEME suite was used in conjunction with four positive external datasets. This analysis aimed to evaluate the enrichment of the motifs within the positive datasets and determine their statistical significance. To ensure the reliability of the validation process, the external datasets used were subjected to an overlap analysis. It was observed that most of the datasets had a similarity overlap lower than 60%. However, an exception was found in the B-TS and TS datasets, where all the sequences in the B-TS dataset were also present in the TS dataset. By using non-redundant and diverse datasets for the validation process, the reliability and generalizability of the discovered motifs were enhanced. This further strengthens the significance and applicability of the motifs in understanding the functional properties of AVPs.

The second stage involved an "inverse-validation" using a unique negative dataset. This step allowed for the elimination of motifs that had a similar probability of appearing in negative sequences. The SEA algorithm [141], like STREME, constructs a control set by shuffling each of the primary sequences while preserving certain statistics. By comparing the occurrence of motifs in the primary sequences to that in the control set, SEA provides insights into the enrichment and significance of the motifs [141]. The validation process serves to confirm the relevance and reliability of the discovered motifs, ensuring that they are not mere chance occurrences but have true functional significance within the context of AVPs.

After the validation process, a final set of 33 motifs was obtained, these motifs along with their respective statistical significance are depicted in Table.5.8. These motifs were searched for in the most central sequences of each of the cluster listed before (Table marked in red). Just in the sequences from cluster one, no motif was found. To address this particularity a further examination of cluster 1 could improve the result. Nonetheless, the other clusters showed in many sequences the occurrence of more than one motif, especially clusters 2 and 4.

To gain further insights into these motifs, a comparison was conducted with existing state of art literature (see Table.5.9). Specifically, two other studies that reported logos of discovered antiviral sequences were examined for comparison [189, 142]. Additionally, the motifs were compared with known sequences of antiviral activity to identify any potential similarities or overlaps. This comparative analysis with existing studies and known antiviral sequences provides valuable context and supports the significance of the discovered motifs. By establishing connections and similarities with previously reported findings, the understanding of the antiviral properties and mechanisms of action associated with these motifs can be expanded, while also validating the reliability of the methodology here employed.

In a study conducted by Balachandran Manavalan et al. in 2022 [189], the researchers focused on comparing different machine-learning approaches for identifying peptides targeting SARS-CoV-2. The study reported statistically significant position-specific compositions of certain predictive sequences. Notably, their predictions highlighted a high occurrence of glycine in the first position of the sequence. Similarly, in our validated mo-

Table 5.7: Full list of validated motifs by SEA software, after removing motifs occurring in negative datasets.

Motif	Cluster	P-value	E-value	TP	Dataset
CYCR	1	0.00	0.00	279/1097 (25.4%)	1
		0.27	1.60	23/520 (4.4%)	2
		0.00	0.02	31/1935 (1.6%)	3
		0.50	3.00	11/217 (5.1%)	4
RRRRH		0.50	3.00	4/1097 (0.4%)	1
		0.43	2.55	15/520 (2.88 %)	2
		0.00	0.00	26/1935 (1.34%)	3
		0.17	1.03	7/217 (3.22%)	4
RRWWC		0.79	4.73	11/1097 (1.0%)	1
		0.23	1.36	5/520 (0.9%)	2
		0.16	0.98	16/1935 (0.8%)	3
		0.94	5.63	2/217 (0.9%)	4
YDISDD		0.99	5.94	4/1097 (0.4%)	1
		0.00	0.03	20/520 (3.8%)	2
		0.00	0.00	57/1935 (2.9%)	3
		0.05	0.29	12/217 (5.52%)	4
CGES	2	0.00	0.00	102 /1097 (9.3%)	1
		0.14	5.91	6 / 217 (2.8%)	4
GCSCK		0.00	0.00	96 /1097 (8.7%)	1
		0.09	3.58	10 /520 (1.9%)	2
		0.09	3.67	7 / 217 (3.2%)	4
VCYRN		0.00	0.00	126 /1097 (11.5%)	1
		0.01	0.01	16 /520 (3.1%)	2
GLPV		0.00	0.00	41 /1097 (3.7%)	1
GTCNTP		0.00	0.00	49 /1097 (4.5%)	1
		0.22	9.15	5 / 520 (0.9%)	2
		0.15	6.28	15 /217 (6.9%)	4
VWIPCI		0.00	0.00	62 /1097 (5.7%)	1
		0.01	0.42	13 /520 (2.5%)	2
		0.00	0.16	8 / 217 (3.7%)	4
SAAJ		0.06	2.27	275 /1097 (25.1%)	1
		0.00	0.00	43 /520 (8.3%)	2
		0.04	1.66	35 /217 (16.1%)	4
QAVG		0.14	5.87	170 /1097 (15.5%)	1
		0.06	2.52	4 / 520 (0.8%)	2

Table 5.8: Full list of validated motifs by SEA software , after removing motifs occurring in negative datasets.

Motif	Cluster	P-value	E-value	TP	Dataset
CKITG	3	0.00	0.00	110 /1097 (10.0 %)	1
GJMDT		0.00	0.00	63 /1097 (5.7%)	1
		0.11	4.41	5 / 520 (0.9%)	2
AGKSVA		0.00	0.00	81 /1097 (7.4%)	1
JFSKI		0.00	0.00	50 /1097 (4.6%)	1
		0.12	5.07	37 /520 (7.1%)	2
		0.03	1.17	23 /217 (10.6%)	4
LLDK		0.00	0.00	29 /1097 (2.6%)	1
		0.09	3.67	10 /217 (4.6%)	4
EAIPLT	4	0.06	2.52	4 / 520 (0.8%)	2
		0.00	0.08	9 / 217 (4.14%)	4
FNK		0.04	1.55	15 /520 (2.9%)	2
IPPEVK		0.20	8.29	14 /1097 (1.3%)	1
		0.11	4.47	5 / 217 (2.3%)	4
KKKKVV	5	0.00	0.00	26 /520 (5.0%)	2
		0.06	2.56	6 / 217 (2.8%)	4
ATYVL		0.00	0.00	41 /520 (7.9%)	2
TKKC		0.00	0.00	168 /1097 (15.3%)	1
WLRDI		0.20	8.10	38 /1097 (3.5%)	1
		0.00	0.01	23 /520 (4.4%9	2
		0.15	6.28	15 /217 (6.9%)	4
LSDFK		0.00	0.00	43 /520 (8.3%)	2
WDWIC		0.18	7.18	18 /520 (3.5%)	2
GLSGL	6	0.00	0.00	32 /1097 (2.9%)	1
		0.11	4.47	5 / 217 (2.3%)	4
GKK		0.19	7.72	153 /1097 (13.9%)	1
		0.12	5.06	17 /217 (7.8%)	4
FLPIV		0.00	0.00	84 /1097 (7.6%)	1
		0.17	6.91	7 / 520 (1.3%)	2
KAAGKA	7	0.00	0.00	122 /1097 (11.1%)	1
SLLGRM		0.03	1.22	27 /1097 (2.5%)	1
		0.01	0.44	11 /520 (2.1%)	2
YFL		0.07	2.93	29 /217 (13.4%)	4
HCKFWW		0.15	5.95	26 /1097 (2.4%)	1
		0.16	6.63	11 /520 (2.1%)	2

Table 5.9: Motifs Found in Literature Reports and other Bioinformatics Studies

MOTIF	Cluster	Reference Sequence	Reference
CYCR	1	CYCRTGRCATRERRSGTCIIQGRL	[190]
RRRRRH		RRRRRRRRHPAEPGSTVTTQNTASQTMS	[142]
CGES	2	IPCGESCVWIPCITA	[142]
VWIPCI		PCGESCVWIPCITA	[142]
QAVG		VYSRCGFAQTLYYDYGVTDMNTLANWVCLVQYESSFNDQAVGAINYNGTQDF GLFQINNKYWCQGAVSSSDSCGIACTSLLGNLSASWSCAQLVYQQQGFS AWYGWLNNCNGTAPSVADCF	[191]
FNK	3	GVTQNVLYENQKQIANQFNKAISQIQESLTTTSTALGKLQ	[15]
		NGIGVTQNVLYENQKQIANQFNKAISQIQESLTTTSTA	
		AASFNKAMTNIVDAFTGVNDAITQTSQALQTVATALNKIQDVVNQQ GNSLNHLTSQ	
		QIANQFNKAISQIQE	[142]
		KQFNKCSLATELSRLGVPKSELPDWVCLVQHESNFKTNWINKKNSNG SWDFGLFQINDKWWCEGHIRSHNTCNVKCEELVTEDIEKALECAK VIKRERGYKAWYGWLNNCQNKKPSVDECF	[191]
KKKKVV	5	KKKKVVAATYV	[142]
WLRDI		SWLRDIWDWICEVLS	[142]
WDWIC		SWLRDIWDWICEVLS	[142]
		SGSWLRDVWDWICTVLTDFKTWLQSKL	[21]
GKK	6	ILPLLKKFGKKFGKKVWKAL	[192]
		AFAFDVTRKINPETSAVERPEVSEYPEIPKGTKLQEFVMMDIEIEEEGADNR AETIQRIKCVPSQCNQICRVLGKKCGYCKNASTCVCLG	[191]

tifs, 18% of them also had glycine as the first residue and 21% of them contained glycine in other positions. Another motif, GKK, was found in the position-specific representation of positive samples estimated using the ENNAVIA-D model [119]. Interestingly, the results from this study [119], also revealed the frequent presence of lysine, leucine, asparagine, glutamic acid, and valine, although in different positions. These residues are also commonly observed in the motifs reported in our study, in particular lysine, leucine, and valine occupying, 36%, 30%, and 27% of the validated motifs respectively.

The second study we examined also focused on identifying anti-coronavirus peptides [142]. In their supplementary information, functional motifs for AVPs were reported. We found higher similarities between these motifs and those reported here. However, an important observation not included in the Table is the high occurrence of arginine, leucine, and valine residues in these sequences, consistent with the previous comparison and our reported motifs. Interestingly, while lysine was found to be highly prevalent in AVPs, its occurrence in non-AVPs was even higher, which is an important factor to consider. Other studies have also confirmed the frequent presence of leucine, glutamic acid, valine, and tryptophan in AVPs [120, 193].

To our knowledge, only 10 of the 33 motifs introduced in this study have been documented in existing literature, indicating the discovery of 23 entirely new motifs. These novel motifs offer promising prospects for designing and testing new sequences in peptide-based antiviral therapeutics. Notably, previous literature predominantly emphasizes longer sequences rather than concise motifs. Consequently, the findings of this study demonstrate a more versatile and "building block" nature compared to similar research works in the field. This underscores the effectiveness of utilizing similarity networks and data mining tools in uncovering novel motifs and advancing our understanding of peptide-based antiviral strategies.

5.5 Multi Query Similarity Search Models

This section addresses the challenge of handling an initial pool of 210 models through a scaled-down process. During the first round of the calibration stage, we evaluated these initial 210 models using three datasets: TR_Starpep, TS_Starpep, and EX_Starpep.

Notably, these datasets, primarily sourced from StarpepDB, exhibited fairly consistent model behavior (Figure.5.12). This result is extremely logical since these datasets and the used "Query" share a lot of sequences.

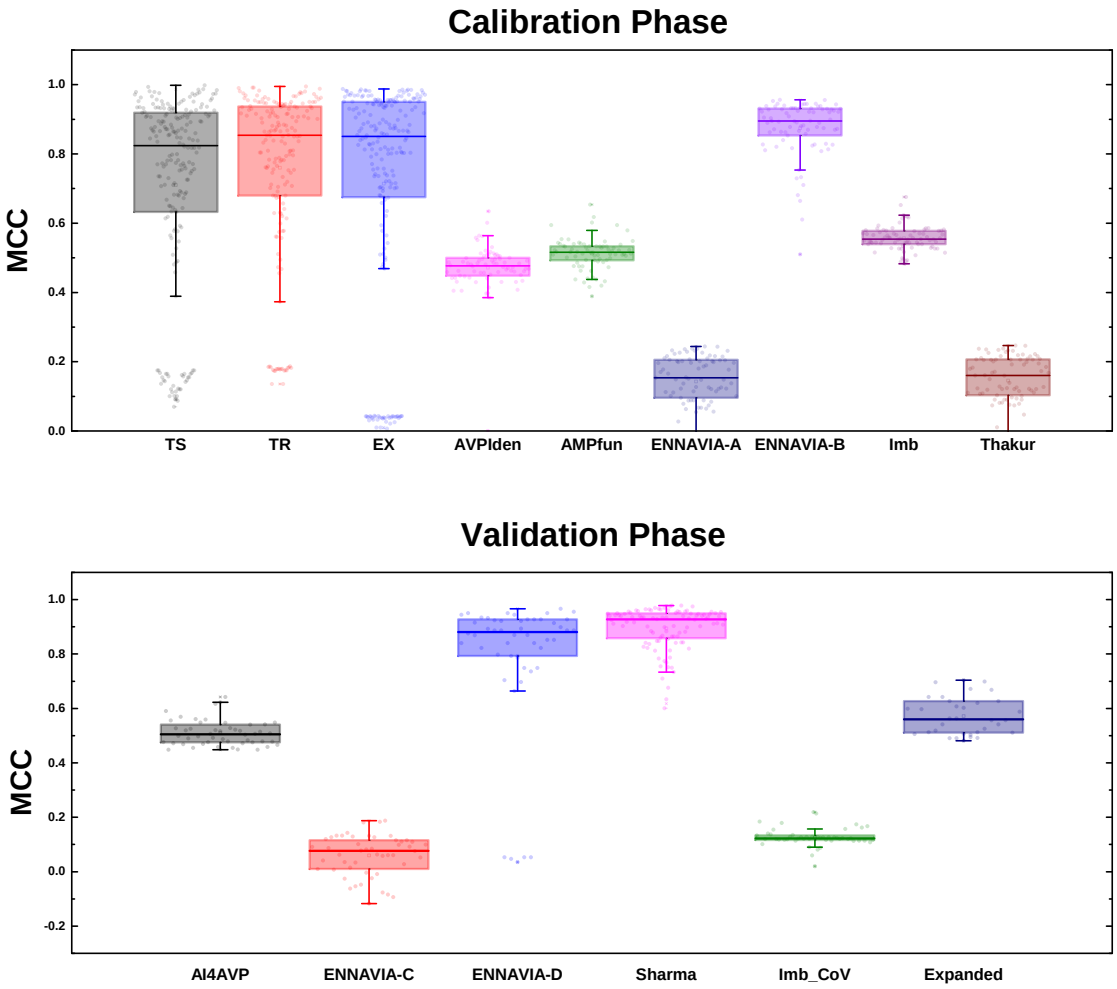


Figure 5.12: MCC Distribution In Calibration and Validation Stage

In the subsequent round of the calibration stage, we introduced six datasets unrelated to StarPepDB. As expected, during this round, the models' performance deteriorated compared to the first round. An essential insight that emerged during this phase is that the models excelled when working with datasets containing randomly generated negative sequences and non-experimentally tested sequences. This observation stemmed from the fact that many of these negative sequences bore a resemblance to the positive ones. As a result, the alignment-based methods struggled to distinguish the distinctive characteristics of each

group.

This issue became particularly evident in datasets such as ENNAVIA-A and Thakur, which included experimental sequences as negative datasets, while with ENNAVIA-B, which has the same amount of positive sequences, the models were far better at recognizing non-antiviral sequences. Additionally, it's crucial to note that the number of experimentally validated negative sequences is relatively limited compared to positive sequences. Therefore, the primary challenge during modeling was to enhance the recall of positive sequences in general.

After the calibration stage, certain trends in the parameters for the QMSS began to emerge. One initially expected observation was that richer scaffolds performed better, as more references fine-tuned the characterization of the AVPs chemical space. Scaffolds like Md4, Md5, SG4, SG5, SL4, and SL5 had the most variants of models.

Furthermore, global types of alignment tended to pair better with lower percentage sequence identity, whereas local types of alignment paired better with higher percentage sequence identity. Additionally, in less rich scaffolds, global types of alignment outperformed local types of alignment at any identity percentage.

5.5.1 Model Selection and Improvement

During the Validation test, one of the approaches taken was to test these models against datasets that contained sequences with a specific antiviral target like SARS-CoV. Datasets like ENNAVIA-C, ENNAVIA-D, and Imb_CoV provided these sequences. As it is shown in the Figure.5.12, both in ENNAVIA-C and IMB_CoV the base models performed really poorly. This behavior is understandable since the references for these models date to 2019, before the discovery of these sequences. This observation, remarks how important is the representation in the "Query" dataset used for the models. On the contrary, the models perform better in the ENNAVIA-D Dataset due to the fact that these datasets contained random negative sequences, a behavior already seen for these models in previous datasets.

Throughout the various stages of model selection, 32 models initially selected in the first round of the Validation stage were tested against the Expanded dataset. From this testing, 12 models emerged as the top performers, determined through a multi-variable

Friedman ranking approach. This group comprised 6 models based on global alignment and 6 on local alignment strategies. All the parameters for these models are summarized in Table.5.10.

At this point, the focus shifted towards enhancing the prediction of positive sequences. Various strategies were explored in pursuit of improved performance. The first approach involved post-processing the models, where combinations of 3, 5, and 7 models were constructed from the 12 final base models. These combinations were formed through a majority vote, with additional models added based on this selection criterion. Nevertheless, this approach did not result in a significant improvement in model performance. Notably, it was observed that the sequences predicted by the models remained largely unchanged, rendering the combinations ineffective since the active space predicted overlapped greatly. Among these combinations, the one comprising models M3, M7, and M12 demonstrated the best performance, prompting their selection for further enhancement.

Table 5.10: Parameters used for selected MQSSMs

Model	**Alignment**	**Identity %**	**Scaffold**
		Base Models	
M1	global	50	Md3
M2	global	50	Md4
M3	global	70	Md4
M4	local	70	SG4
M5	local	80	SL5
M6	global	40	SG4
M7	local	90	SL5
M8	global	40	SG5
M9	global	50	SG5
M10	local	70	SG5
M11	local	80	SG5
M12	local	90	SG5
		Modified Models	
M13	global	90	Fusion
M3+	global	70	Md4+
M7+	local	90	SL5+
M12+	local	90	SG5+
M13+	global	90	Fusion+
E1	global	90	EG5
E2	global	90	EL5

The second approach, an *a priori* modification, entailed extracting and combining the scaffolds of models M3, M7, and M12 (md4,SL5, SG5) into a single scaffold while eliminating duplicated sequences. This scaffold comprised 3206 unique sequences. Testing this modified scaffold revealed a slight improvement in predictions by models using global alignment with a similarity threshold of 90. This improvement is due to an increase in the representativity of the space with more sequences.

Despite these changes providing valuable insights into the specificities of the MQSSMs, their overall performance remained unsatisfactory. As indicated in the figure, the base models tested in the Calibration Phase performed poorly on datasets such as Thakur, ENNAVIA-A, AMPfun, and AVPiden as the Figure.5.12 shows. This suggests that many sequences in these datasets were not adequately represented in the scaffolds used for the QMSS. Furthermore, several of these datasets contained many experimentally validated negative sequences, adding to the complexity of predictions.

In response to these challenges, a new HSPN was constructed by aggregating the positive sequences from these problematic target datasets. The total number of sequences used for the HSPN was 2403 sequences. The resulting HSPN produced 8 scaffolds, the best 2 references were selected to enhance the current scaffolds. This enhancement introduced new models, namely M3+, M7+, M12+, and M13+, denoted by the "+" signifying their enriched nature. Now these scaffolds contained 3155, 3437, 3472, and 3606 sequences respectively, To retain scaffolds that did not overlap with the sequences found in StarPepDB, models E1 and E2 were crafted using the external scaffold. The scaffolds used for E1 and E2 contained 1517 and 1261 sequences respectively. This increased the total number of models for analysis to 10, adding 6 enriched models to the already existing M3, M7, M12, and M13. All the particularities of the models are shown in Table.5.10.

Subsequently, these 10 models were rigorously tested across the 15 databases (the 14 datasets from the workflow and the Expanded Dataset), and a Friedman ranking was employed to reduce the number of top-performing models by half, taking into account the metrics of ACC, SP, SN, MCC, and F1. The results of this ranking revealed that the best-performing methods were M3+, M13+, M7, M12, and E1, marked in gray in Table.5.11. Is important to notice that although the models that had more sequences as references are in the top 5, models like E1, M7, and M12, which contain fewer sequences than their

enriched counterparts, perform similarly, meaning that the key for the reference is not in the number but in their diversity and range of representativity. These top 5 models were then benchmarked against existing predictors from the literature to assess their current performance in comparison with other available tools.

Table 5.11: Models Performance Evaluation for the Expanded Dataset

Model Name	**ACC**	**SP**	**SN**	**MCC**	**FPR**	**F1 Score**
E1	0.966	0.995	0.481	0.624	0.005	0.614
E2	0.961	0.995	0.398	0.562	0.005	0.540
M12	0.958	0.962	0.891	0.704	0.038	0.708
M12+	0.736	0.724	0.937	0.330	0.276	0.288
M13	0.964	0.980	0.694	0.667	0.020	0.686
M13+	0.969	0.979	0.802	0.731	0.021	0.746
M3	0.935	0.944	0.782	0.568	0.056	0.577
M3+	0.935	0.939	0.876	0.609	0.061	0.606
M7	0.958	0.964	0.873	0.699	0.036	0.705
M7+	0.736	0.724	0.933	0.329	0.276	0.287

5.5.2 Comparison with state of the art

To ensure a fair comparison, all sequences overlapping between the models' scaffolds (M3+,M7,M12,M13+,E1) and the Reduced dataset were removed. This step lowered the number of positive sequences to just 116, while the count of negative sequences remained unchanged. It's worth noting that many of the negative sequences in the Reduced dataset are part of the training datasets from the external predictors, but these sequences were retained. This slight adjustment places the models at a considerable disadvantage in performance evaluation. A total of 14 External Predictors were tested, evaluating metrics including Accuracy (ACC), Specificity (SP), Sensitivity (SN), Mathews Correlation Coefficient (MCC), the False Positive Rate (FPR), and the F1 Score. Among these metrics, MCC is the most relevant as it's not affected by imbalanced data, as is the case here, and is the one on which the following analysis is most based.

Table 5.12: Performance Comparison With *State-of-the-Art* Predictors

Model Name	ACC	SP	SN	MCC	FPR	F1 Score
M3+	0.929	0.930	0.603	0.137	0.070	0.069
M7	0.970	0.972	0.448	0.163	0.028	0.115
M12	0.968	0.971	0.466	0.165	0.029	0.114
M13+	0.983	0.986	0.422	**0.214**	0.014	0.180
E1	0.993	**0.996**	0.190	0.184	0.004	**0.187**
AI4AVP	0.387	0.385	0.905	0.039	0.615	0.013
AI4AVP(DA)	0.379	0.376	0.871	0.034	0.624	0.012
FIRM-AVP	0.647	0.647	0.595	0.034	0.353	0.015
Meta-iAVP	0.594	0.593	0.647	0.032	0.407	0.014
seqpros	0.119	0.116	**0.940**	0.011	0.884	0.009
AMPfun	0.463	0.462	0.784	0.033	0.538	0.013
iACVP	0.893	0.895	0.517	0.088	0.105	0.041
PTPAMP	0.825	0.827	0.336	0.028	0.173	0.017
ClassAMP	0.795	0.798	0.310	0.018	0.202	0.013
AntiVPP	0.732	0.734	0.457	0.028	0.266	0.015
ProtDcalRF	**0.995**	1.000	0.000	-0.001	**0.000**	0.000
ProtDcalHier	**0.995**	**0.999**	0.000	-0.002	0.001	0.000
ProtDcalRNN	0.950	0.954	0.034	-0.004	0.046	0.006
AVPpred	0.902	0.904	0.371	0.062	0.096	0.032

With the modification to the Reduced dataset, the performance of the QMSSMs has noticeably dropped. This decline is particularly evident in the SN and MCC values, while ACC and SP remain relatively stable due to the significant class imbalance between positive and negative cases. The reduced sensitivity highlights a significant and consistent deficit in the recall of positive sequences. This failure to recover true positives also impacts the MCC, dropping from 0.731 to 0.214 for the M13+ model (Table.5.12). Correspondingly, the F1 score also declines, as it relies on both recall and precision.

Despite the unsatisfactory results, the QMSS models outperform the external predictors overall. From the performance results obtained from the external predictors, two different tendencies can be inferred from Figure.5.13. Some prediction models excel at recognizing most positive sequences, achieving high SN but at the expense of a high rate of false positives. Others are proficient at identifying all negative sequences but misclassify many positive ones, a characteristic of the presented QMSS models. Most deep learning-based models fall into the first category, while traditional machine learning models fall into the

second.

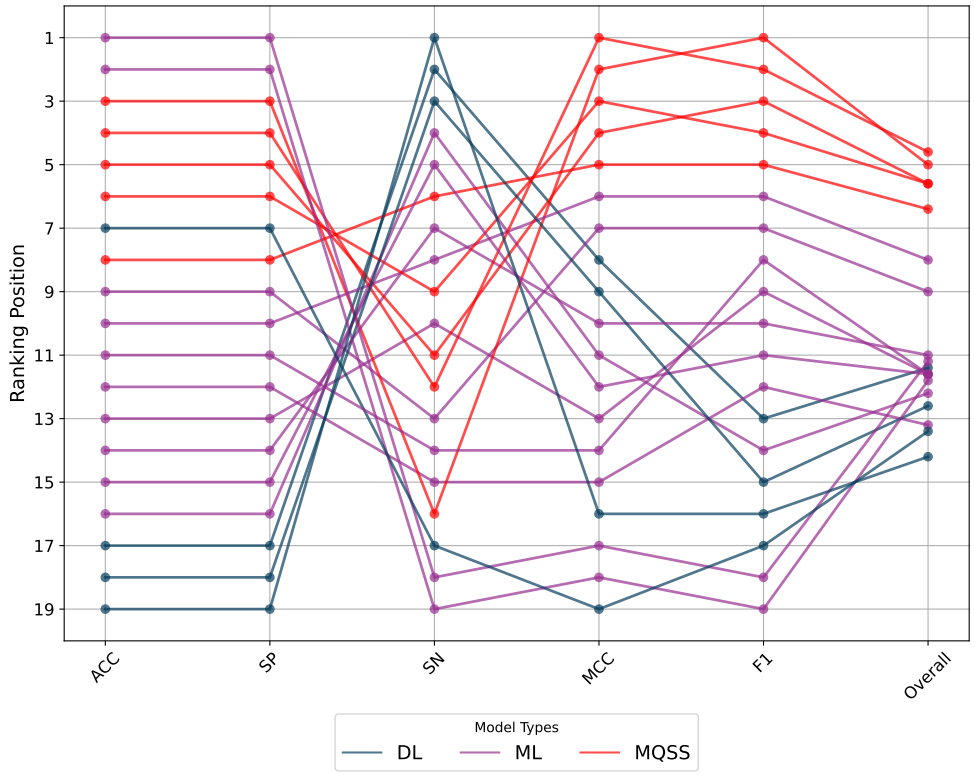


Figure 5.13: Ranking Change Based on Metrics

In general, no single predictor performs well in all categories, corroborating the findings of Garcia-Jacas et al. that the use of deep learning methods for AVP prediction is not justified. Expanding on that remark, none of the tested machine learning models are performing well enough, indicating that significant work is still required to enhance these models. The primary issue here lies in the quality of the training data, which is insufficiently numerous and representative. Most positive sequences are similar, with up to 90% similarity. Additionally, the validated negative sequences closely resemble the positive ones. The complexity of the architectures employed is not the problem; instead, the data availability has been a long-standing challenge to collect and analyze.

Another common challenge encountered when evaluating the *state-of-the-art* predictors was the accessibility issue. Table.3.2 list numerous predictors, many of which were quite challenging to assess. Firstly, several of the web servers proved to be poorly constructed and were either not currently operational or malfunctioned regularly, including servers less than 2 years old. Additionally, many of the repositories provided by researchers lacked

clear instructions, making their implementation more complicated. This issue aligns with concerns raised by [111] regarding the availability of source codes. In this context, the MQSSMs stand out as they are readily accessible through the StarPep toolbox standalone software, which features an intuitive interface, simplifying their usage.

Although the MQSSMs require substantial improvement to address their current deficits, they still outperform the ML models. A Friedman Ranking, listed in Table.5.13, considering MCC, ACC, SP, SN, and F1, was conducted to support these findings. Nonetheless, the computational resources required by the QMSSMs are significantly lower, and they don't have length limitations, offering crucial advantages to consider.

In selecting a prediction model, the researcher's specific requirements must be considered. Researchers may require many potential positive sequences or a smaller set with a lower likelihood of false positives. For the purpose of synthesizing potential sequences, the second approach seems more useful due to the high resource demands of experimental procedures.

Table 5.13: Ranking of all predictors evaluated

Predictor	Friedman Ranking	Predictor	Friedman Ranking
M13+	4.6	Meta	11.6
E1	5	PTPAMP	11.6
M7	5.6	AntiVPP	11.6
M12	5.6	ProtDcalHier	11.8
M3+	6.4	AMPfun	12.2
iACVP	8	PC6	12.6
AVPpred	9	ClassAMP	13.2
FIRM	11	ProtDcalRNN	13.4
ProtDcalRF	11.2	seqpros	14.2
AI4AVP	11.4		

5.6 Proposal of New AVPs

A significant aspect of this study involves proposing new potential antiviral peptide sequences. These predictions are based on the tools introduced in various sections and the previously developed MQSS models. Three different databases served as the starting point for virtual cleavage to initiate this process.

Starting Point: StarPepDB

Following the workflow outlined in the scheme, the initial number of sequences found in StarPepDB is 45,120. Of these, 34,093 sequences have a length of less than 35 amino acids (AAs). From this subset, 7,633 are labeled as toxic, hemolytic, or both, and only 26,327 sequences do not contain non-standard AAs. After applying the first predictive model (M13+), 1,256 sequences remained. Within this subset, 402 sequences did not overlap with experimentally validated antiviral and antiviral-related sequences. These sequences then underwent evaluation using various web predictors for antiviral activity, toxicity, hemolysis, allergenicity, and a GRAVY calculator. Following this rigorous filtering process, a final set of 12 sequences was obtained from StarPepDB.

Starting Point: Human Proteome

Similarly, the starting point for this database consisted of 42,999 sequences. Among them, 27,999 sequences had a maximum length of 35 AAs and did not contain non-standard AAs. After applying the M13+ model, the sequence count was reduced to 6,835. Subsequently, by eliminating sequences that overlapped with experimentally validated sequences and those with redundancy exceeding 90%, the count dropped to 1,268 sequences. These sequences were subjected to various web servers to evaluate antiviral activity, hemolysis, toxicity, allergenicity, and GRAVY score calculation. After these additional filtration steps, 32 sequences were obtained from the Human Proteome.

Starting Point: Cephalopods

The starting point for this dataset was 68,694 sequences that had already been tested for toxicity and hemolytic activity. By restricting the peptide length to 35 AAs, the dataset was reduced to 66,337 sequences. Applying the M13+ model to this subset reduced the potential AVPs to 13,801 sequences. After removing sequences with a similarity greater than 90%, the count lowered to 13,195. These sequences then underwent evaluation through various web server predictors, GRAVY calculation, and allergen prediction, resulting in 63 sequences identified as potential AVPs.

The utilization of numerous web servers is essential to predict various properties crucial for therapeutic development during the virtual cleavage, in addition to prediction for

anntiviral activity. Take, for instance, antimicrobial peptides (AMPs). To be considered suitable for systemic applications, these peptides must exhibit low toxicity towards erythrocytes[194]. Moreover, toxicity and immunogenicity represent significant concerns when considering them for therapeutic purposes [195]. Furthermore, the prediction of the GRAVY Index ensures that these sequences possess hydrophilic characteristics, which in turn suggest a more globular behavior and increased solubility.

A union of the chosen peptides from the aforementioned starting points was created to narrow down the selection, totaling 107 sequences. However, this number was still considered too large. These sequences were combined into a single input, and an HSPN was constructed using them. Subsequently, a Scaffold Extraction was conducted utilizing the Community Hub-Bridge Centrality and a local alignment with a 50% sequence identity threshold. This procedure reduced the pool of sequences to 92, which was still considered too extensive.

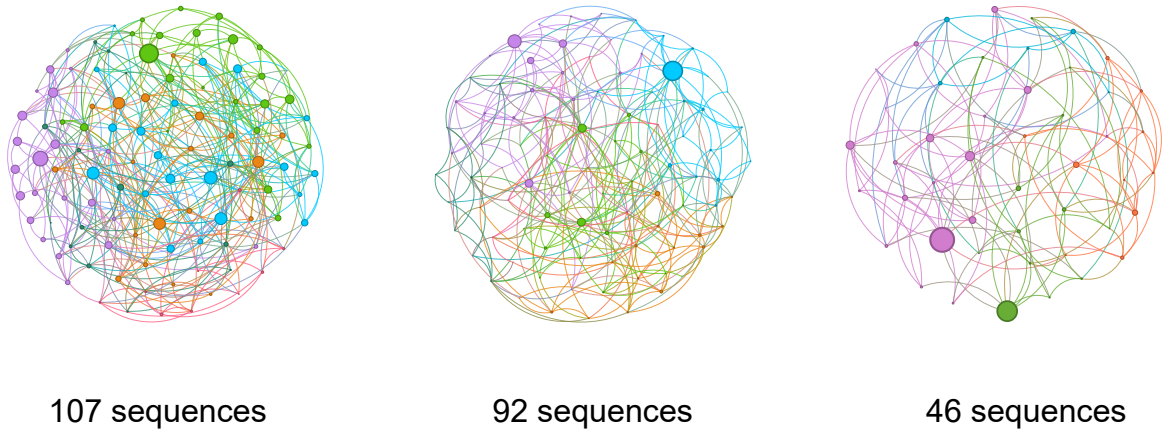


Figure 5.14: Comparison of Different HSPNs constructed from the hits sequences

A visual analysis was performed to further refine the selection, and sequences were manually selected based on how they were clustered using the Louvain Clustering Algorithm and their centrality within the network. This final curation step resulted in the presentation of 46 sequences in the Table.5.14, accompanied by relevant physicochemical features.Among these features some annotations are necessary, all the proposed peptides have a positive charge. Cationic peptides, due to the high presence of lysine and arginine,

have been extensively regarded as antimicrobial peptides due to their ability to attack microorganisms directly [196, 197]. These observations support further the potential antiviral activity of the 46 sequences. To give a graphical representation of the last reduction step HSPNs were constructed based on the 107,92 and 46 sequences selections (Figure.5.14).

Finally, this study also includes predictions for other antimicrobial activities, which are presented in Table 5.16. These predictions were generated using ProtDcal to assess antibacterial, antifungal, and antiparasitic activities as well as for the AMPfun predictor. This addition is grounded in the premise that Antiviral Peptides share close relations with other antimicrobial active peptides, as highlighted in the Metadata Network depicted in Figure.5.1.

Table 5.14: Proposed Peptide Sequences as AVP Hits

Identifier	Sequence	Length	MW	Ip	Charge	GRAVY
P27877	LQRDFLKQRPTKLKSLIRLVKHWYQNCKKKLGKLP	35	4306.28	11.32	10.02	-0.92
SRR5204441	KRHLRLITRWGE	12	1564.86	12.19	3.09	-1.17
P27484	QKLLKKIGISWNKKYRKQHGPLRKFLQLHSQIFLL	35	4289.23	11.82	9.18	-0.59
P27345	KINRPRTELKKKLFKRRRVLNRERRLRHRVVGAVI	35	4393.36	12.85	13.09	-1.11
P27727	LLRKYRRELQLRKKCHNELVRLKGNIRVIARVRPV	35	4325.29	12.28	10.03	-0.65
SRR3105558	RVTLCHCSGQGCHAPSRFAVH	21	2266.60	8.45	2.08	-0.03
SRR2047107	GTRPAIVRRGKSITPVF	17	1855.22	12.79	4.00	-0.09
SRR2047107_	NSGGSHKIVNWGPKAVY	17	1814.03	10.21	2.09	-0.58
SRR5204442	KTICGGVGNQKRVSWCSG	18	1880.17	9.93	2.87	-0.34
SRR3105558	GQVSSTIASGRPKPCGRSPSY	21	2135.38	10.43	2.93	-0.73
SRR6349992	IWTLPRRKIVNRYVLPFE	18	2300.78	11.34	3.00	-0.15
SRR2047107	CSLVYHRQGHTAAPCVHW	18	2065.36	8.16	1.15	-0.14
SRR2047107	SITSTPVRNWYLFHRPPKG	19	2256.60	11.44	3.09	-0.77
P27545	LVQRRKKRLRHRVPPRKPEPLVKLHRTALHWACLK	35	4318.30	12.62	11.21	-0.97
P27679	ARLWLHVLPTLPGTLCLRIFRWGPRRRRQGSRTLL	35	4197.07	12.96	8.03	-0.21
SRR3105558	RTNYLWIFRPGQSLRRLLS	19	2376.79	12.48	4.00	-0.47
P27200	ALLWRGRIPGRQWIGKHRRPRFVSLRAKQNMIRRL	35	4324.23	13.38	11.09	-0.79
P26071	FRLQRALRFLLGRLPWRVASGARRFRRWLNRYSY	34	4339.14	12.85	10.00	-0.58
P27157	RLLRGLWQRWRRYKYRFVPWIALNLSHNPRTLRYV	35	4585.44	12.49	9.09	-0.71
P27739	RWGQAISKSKLSRRWRRWNRFNRRRCRAAVKSVTF	35	4421.17	13.08	12.94	-1.26
P26563	ALCTLRGRRCHCLPFPKRGMQRWMLPLRRGARLL	34	4064.04	12.48	8.90	-0.29
SRR5204441	PPGKGRSRIFSRLGPKWE	18	2068.41	12.21	4.00	-1.23
P26773	ILDPWVYILFRRAVLRRLQPRLSTRPRRSLTLWP	34	4246.13	12.70	7.00	-0.20
P27743	PAWWFMSVKVKLKRKKRQKTPNRKLPRRGAPTLRP	35	4257.21	13.19	13.00	-1.30

Table 5.15: Proposed Peptide Sequences as AVP Hits 2

Identifier	Sequence	Length	MW	Ip	Charge	GRAVY
P27877	LQRDFLKQRPTKLKSLIRLVKHWYQNCKKKLGKLP	35	4306.28	11.32	10.02	-0.92
SRR5204441	KRHLRLITRWGE	12	1564.86	12.19	3.09	-1.17
P27484	QKLLKKIGISWNKKYRKQHGPLRKFLQLHSQIFLL	35	4289.23	11.82	9.18	-0.59
P27345	KINRPRTELKKKLFKRRRVLNRERRLRHRVVGAVI	35	4393.36	12.85	13.09	-1.11
P27727	LLRKYRRELQLRKKCHNELVRLKGNIRVIARVRPV	35	4325.29	12.28	10.03	-0.65
SRR3105558	RVTLCHCSGQGCHAPSRFAVH	21	2266.60	8.45	2.08	-0.03
SRR2047107	GTRPAIVRRGKSITPVF	17	1855.22	12.79	4.00	-0.09
SRR2047107_	NSGGSHKIVNWGPKAVY	17	1814.03	10.21	2.09	-0.58
SRR5204442	KTICGGVGNQKRVSWCSG	18	1880.17	9.93	2.87	-0.34
SRR3105558	GQVSSTIASGRPKPCGRSPSY	21	2135.38	10.43	2.93	-0.73
SRR6349992	IWTLPRRKIVNRYVLPFE	18	2300.78	11.34	3.00	-0.15
SRR2047107	CSLVYHRQGHTAAPCVHW	18	2065.36	8.16	1.15	-0.14
SRR2047107	SITSTPVRNWYLFHRPPKG	19	2256.60	11.44	3.09	-0.77
P27545	LVQRRKKRLRHRVPPRKPEPLVKLHRTALHWACLK	35	4318.30	12.62	11.21	-0.97
P27679	ARLWLHVLPTLPGTLCLRIFRWGPRRRRQGSRTLL	35	4197.07	12.96	8.03	-0.21
SRR3105558	RTNYLWIFRPGQSLRRLLS	19	2376.79	12.48	4.00	-0.47
P27200	ALLWRGRIPGRQWIGKHRRPRFVSLRAKQNMIRRL	35	4324.23	13.38	11.09	-0.79
P26071	FRLQRALRFLLGRLPWRVASGARRFRRWLNRYSY	34	4339.14	12.85	10.00	-0.58
P27157	RLLRGLWQRWRRYKYRFVPWIALNLSHNPRTLRYV	35	4585.44	12.49	9.09	-0.71
P27739	RWGQAISKSKLSRRWRRWNRFNRRRCRAAVKSVTF	35	4421.17	13.08	12.94	-1.26
P26563	ALCTLRGRRCHCLPFPKRGMQRWMLPLRRGARLL	34	4064.04	12.48	8.90	-0.29
SRR5204441	PPGKGRSRIFSRLGPKWE	18	2068.41	12.21	4.00	-1.23
P26773	ILDPWVYILFRRAVLRRLQPRLSTRPRRSLTLWP	34	4246.13	12.70	7.00	-0.20
P27743	PAWWFMSVKVKLKRKKRQKTPNRKLPRRGAPTLRP	35	4257.21	13.19	13.00	-1.30

Table 5.16: Other Potential Antimicrobial Activities of the proposed sequences

	Antibacterial		**Antifungal**		**AntiParasitic**	
Peptide ID	ProtDcal	AMPfun	ProtDcal	AMPfun	ProtDcal	AMPfun
P27877	✓	×	✓	✓	✓	×
SRR5204441	✓	×	✓	×	✓	×
P27484	✓	✓	✓	×	✓	×
P27345	✓	×	×	×	✓	✓
P27727	✓	×	×	×	✓	✓
SRR3105558	✓	✓	✓	✓	✓	×
SRR2047107	✓	✓	✓	×	✓	×
SRR2047107_	✓	✓	✓	×	×	×
SRR5204442	✓	✓	✓	✓	✓	×
SRR3105558	✓	✓	✓	×	✓	×
SRR6349992	✓	✓	×	×	✓	×
SRR2047107	✓	✓	✓	×	✓	×
SRR2047107	✓	✓	✓	×	✓	×
P27545	✓	✓	✓	×	✓	×
P27679	✓	✓	✓	×	✓	×
SRR3105558	✓	✓	✓	×	✓	×
P27200	✓	✓	×	×	✓	✓
P26071	✓	✓	✓	×	✓	×
P27157	✓	✓	✓	×	✓	×
P27739	✓	✓	✓	×	✓	×
P26563	✓	✓	✓	×	✓	✓
SRR5204441	✓	✓	✓	×	✓	×
P26773	✓	×	✓	×	✓	×
P27743	✓	✓	×	×	✓	×
SRR2047107	✓	✓	✓	✓	✓	×
SRR6349992	✓	✓	×	×	✓	×
DN19901	×	×	✓	×	✓	×
SRR3105321	×	✓	×	×	×	✓
SRR5204441	✓	✓	✓	×	✓	×
SRR3105558	✓	✓	✓	×	✓	×
SRR3105558	✓	✓	✓	×	✓	×
DN11116	×	✓	✓	×	✓	✓
DN8078	✓	✓	✓	×	✓	×
starPep44946	×	×	×	×	×	×
SRR6349992	✓	✓	×	×	✓	×
SRR6349992	✓	×	×	×	✓	×
SRR3105321	×	✓	×	×	×	×
SRR5204441	✓	✓	✓	×	✓	×
SRR2047107	✓	✓	✓	×	✓	×
SRR6349992	✓	✓	×	×	✓	×
SRR6349992	✓	✓	×	×	✓	×
SRR5204441	✓	✓	✓	×	✓	×
SRR3105558	✓	✓	✓	×	✓	×
P27042	✓	✓	×	×	✓	✓
SRR2047107	✓	✓	✓	×	✓	×
starPep37607	✓	✓	✓	×	✓	✓

Chapter 6

Conclusions

The chemical space of AVPs, which comprises 3494 sequences (StarPepDB (36)), was effectively represented using the HSPN implemented in the StarPep toolbox (37). The optimal similarity threshold (t) for the HSPNs was determined to be 0.75, although no major representational differences were found when compared with an HSPN t = 0. The most representative peptides within the HSPNs were identified using HB centrality. These peptides were characterized based on their chemical and biological properties. The Louvain clustering algorithm was applied to the HSPN with optimal cut-off and without cut-off, resulting in the identification of distinct communities within the AVP chemical space. These communities present in HSPN (t = 0) were individually studied, and cluster profiles were created based on various molecular descriptors and literature-based validations.

In addition to HSPN construction, the AVPs' chemical space was further explored using Metadata Networks, which incorporated metadata information such as database, function, origin, and target. This provided a comprehensive understanding of the AVP landscape. These networks revealed that the majority of AVPs are of synthetic origin and are frequently associated with other antimicrobial peptides. Additionally, the most commonly assessed target for antiviral therapeutics was HIV.

To obtain representative and diverse subsets of AVPs, scaffold extraction was performed with similarity thresholds ranging from 90% to 50%. Changes in alignment algorithm and centrality measures were studied during scaffold extraction. The scaffold comparison revealed no significant difference between the HSPNs with and without cut-off, further supporting the robustness of the network representation. Additionally, it was observed that

the most representative centrality-alignment algorithm combination was the global alignment algorithm and the Community Hub-Bridge Centrality, as well as the local alignment with the Harmonic Centrality.

Furthermore, 42 potential motifs were discovered using the STREME algorithm, and subsequent validation using the SEA algorithm was conducted on 33 motifs. The validation process involved measuring the relative enrichment of the motifs in four external datasets. This analysis confirmed the motifs' significance and potential functional relevance, with 23 completely novel motifs. The identified motifs provide valuable insights into potential functional patterns within AVPs, contributing to the understanding and discovery of AVPs.

Moreover, a significant milestone in this research is the successful design and implementation of Multi-Query Similarity Search Models (MQSSMs). These models were meticulously developed based on the insights gained during chemical space exploration and scaffold extraction stages. This innovative approach, grounded in structural similarities concept, leading to analogous biological activities, has played a pivotal role in refining model selection.

Furthermore, creating the largest dataset of Antiviral sequences to date is another remarkable achievement, as the availability of diverse data is one of the main concerns of the design of AVPs predictors. The extensive evaluation and various filtering stages culminated in the selection of five final MQSSMs, each rigorously assessed using various metrics. The highest-performing model achieved the following: ACC=0.986, SP=0.930, SN=0.422, MCC=0.214, FPR=0.014, and F1=0.180. Compared to *state-of-the-art* machine learning-based predictors, the MQSS models outperformed the 14 predictors tested. This analysis underscored the current limitations and deficiencies of these models. The MQSSMs demonstrated superior capabilities in predicting AVP sequences while addressing the challenges posed by variable-length sequences and imbalanced data.

Building upon the refinement of the MQSSMs, a virtual cleavage process was undertaken to propose potential Antiviral Sequences. This process was applied to three different datasets, collectively comprising more than 100,000 sequences. Additionally, these sequences underwent scrutiny for toxicity, hemolytic activity, allergen activity, and other antimicrobial activities. Herein, we report 46 potential antiviral peptides with sequences shorter than 35 amino acids, which were predicted as non-toxic, non-hemolytic,

non-allergenic, and antiviral by the presented MQSSMs and several online web predictors.

The MQSSMs performance was measured and results were derived from the Multi-Query Similarity Search (MQSS) method. This innovative approach, grounded in the principle that structural similarities often lead to analogous biological activities, has been instrumental in refining model selection. The meticulous evaluation conducted using the Expanded dataset resulted in the identification of 12 models that exhibited unparalleled performance characteristics. When juxtaposed against existing state-of-the-art predictors, these models emerged superior, setting a new gold standard in AVP discovery.

Bibliography

[1] L. W. Enquist, "Virology in the 21st Century," *Journal of Virology*, vol. 83, no. 11, pp. 5296–5308, Jun. 2009. [Online]. Available: https://journals.asm.org/doi/10.1128/JVI.00151-09

[2] R. Sanjuán and P. Domingo-Calap, "Mechanisms of viral mutation," *Cellular and Molecular Life Sciences*, vol. 73, no. 23, pp. 4433–4448, Dec. 2016. [Online]. Available: http://link.springer.com/10.1007/s00018-016-2299-6

[3] R. M. Meganck and R. S. Baric, "Developing therapeutic approaches for twenty-first-century emerging infectious viral diseases," *Nature Medicine*, vol. 27, no. 3, pp. 401–410, Mar. 2021. [Online]. Available: https://www.nature.com/articles/s41591-021-01282-0

[4] P. L. Yang, "Antiviral therapeutics," *ACS Infectious Diseases*, vol. 7, no. 6, pp. 1297–1297, Jun. 2021. [Online]. Available: https://doi.org/10.1021/acsinfecdis.1c00271

[5] A. Mahmoud, "New vaccines: challenges of discovery," *Microbial Biotechnology*, vol. 9, no. 5, pp. 549–552, Sep. 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/1751-7915.12397

[6] Z. Lou, Y. Sun, and Z. Rao, "Current progress in antiviral strategies," *Trends in Pharmacological Sciences*, vol. 35, no. 2, pp. 86–102, Feb. 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165614713002265

[7] X. Liu, P. Zhan, L. Menéndez-Arias, and V. Poongavanam, Eds., *Antiviral Drug Discovery and Development*, ser. Advances in Experimental Medicine and

Biology. Singapore: Springer Singapore, 2021, vol. 1322. [Online]. Available: https://link.springer.com/10.1007/978-981-16-0267-2

[8] L. Wang, N. Wang, W. Zhang, X. Cheng, Z. Yan, G. Shao, X. Wang, R. Wang, and C. Fu, "Therapeutic peptides: current applications and future directions," *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, p. 48, Feb. 2022. [Online]. Available: https://www.nature.com/articles/s41392-022-00904-4

[9] J. L. Lau and M. K. Dunn, "Therapeutic peptides: Historical perspectives, current development trends, and future directions," *Bioorganic & Medicinal Chemistry*, vol. 26, no. 10, pp. 2700–2707, Jun. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968089617310222

[10] C. Lamers, "Overcoming the shortcomings of peptide-based therapeutics," *Future Drug Discovery*, vol. 4, no. 2, p. FDD75, Jun. 2022. [Online]. Available: https://www.future-science.com/doi/10.4155/fdd-2022-0005

[11] A. Henninot, J. C. Collins, and J. M. Nuss, "The Current State of Peptide Drug Discovery: Back to the Future?" *Journal of Medicinal Chemistry*, vol. 61, no. 4, pp. 1382–1414, Feb. 2018. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jmedchem.7b00318

[12] G. Castel, M. Chtéoui, B. Heyd, and N. Tordo, "Phage display of combinatorial peptide libraries: Application to antiviral research," *Molecules*, vol. 16, no. 5, pp. 3499–3518, Apr. 2011. [Online]. Available: https://doi.org/10.3390/molecules16053499

[13] M. S. Mousavi Maleki, S. Sardari, A. Ghandehari Alavijeh, and H. Madanchi, "Recent Patents and FDA-Approved Drugs Based on Antiviral Peptides and Other Peptide-Related Antivirals," *International Journal of Peptide Research and Therapeutics*, vol. 29, no. 1, p. 5, Nov. 2022. [Online]. Available: https://link.springer.com/10.1007/s10989-022-10477-z

[14] E. P. Carter, C. G. Ang, and I. M. Chaiken, "Peptide Triazole Inhibitors of HIV-1: Hijackers of Env Metastability," *Current Protein & Peptide Science*, vol. 24, no. 1, pp. 59–77, Jan. 2023. [Online]. Available: https://www.eurekaselect.com/205833/article

[15] H. Heydari, R. Golmohammadi, R. Mirnejad, H. Tebyanian, M. Fasihi-Ramandi, and M. Moosazadeh Moghaddam, "Antiviral peptides against Coronaviridae family: A review," *Peptides*, vol. 139, p. 170526, May 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0196978121000346

[16] M. Agamennone, M. Fantacuzzi, G. Vivenzio, M. C. Scala, P. Campiglia, F. Superti, and M. Sala, "Antiviral Peptides as Anti-Influenza Agents," *International Journal of Molecular Sciences*, vol. 23, no. 19, p. 11433, Sep. 2022. [Online]. Available: https://www.mdpi.com/1422-0067/23/19/11433

[17] H. Jenssen, "Anti herpes simplex virus activity of lactoferrin/lactoferricin – an example of antiviral activity of antimicrobial protein/peptide," *Cellular and Molecular Life Sciences CMLS*, vol. 62, no. 24, pp. 3002–3013, Dec. 2005. [Online]. Available: https://link.springer.com/10.1007/s00018-005-5228-7

[18] Y. Becker, "Dengue fever virus and Japanese encephalitis virus synthetic peptides, with motifs to fit HLA class I haplotypes prevalent in human populations in endemic regions, can be used for application to skin Langerhans cells to prime antiviral CD8+ cytotoxic T cells (CTLs)—A novel approach to the protection of humans," *Virus Genes*, vol. 9, no. 1, pp. 33–45, Sep. 1994. [Online]. Available: http://link.springer.com/10.1007/BF01703433

[19] J. Y. Park, S. Y. Yang, Y. C. Kim, J.-C. Kim, Q. L. Dang, J. J. Kim, and I. S. Kim, "Antiviral peptide from Pseudomonas chlororaphis O6 against tobacco mosaic virus (TMV)," *Journal of the Korean Society for Applied Biological Chemistry*, vol. 55, no. 1, pp. 89–94, Feb. 2012. [Online]. Available: http://link.springer.com/10.1007/s13765-012-0015-2

[20] H. Jenssen, T. J. Gutteberg, O. Rekdal, and T. Lejon, "Prediction of Activity, Synthesis and Biological Testing of anti-HSV Active Peptides," *Chemical Biology*

<*html_ent glyph="@amp;" ascii="&"/> *Drug Design*, vol. 68, no. 1, pp. 58–66, Jul. 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1747-0285.2006.00412.x

[21] J. A. Jackman, V. V. Costa, S. Park, A. L. C. V. Real, J. H. Park, P. L. Cardozo, A. R. Ferhan, I. G. Olmo, T. P. Moreira, J. L. Bambirra, V. F. Queiroz, C. M. Queiroz-Junior, G. Foureaux, D. G. Souza, F. M. Ribeiro, B. K. Yoon, E. Wynendaele, B. De Spiegeleer, M. M. Teixeira, and N.-J. Cho, "Therapeutic treatment of Zika virus infection using a brain-penetrating antiviral peptide," *Nature Materials*, vol. 17, no. 11, pp. 971–977, Nov. 2018. [Online]. Available: https://www.nature.com/articles/s41563-018-0194-2

[22] D. Cook and L. Holder, "Graph-based data mining," *IEEE Intelligent Systems*, vol. 15, no. 2, pp. 32–41, Mar. 2000. [Online]. Available: http://ieeexplore.ieee.org/document/850825/

[23] J. L. Medina-Franco, N. Sánchez-Cruz, E. López-López, and B. I. Díaz-Eufracio, "Progress on open chemoinformatic tools for expanding and exploring the chemical space," *Journal of Computer-Aided Molecular Design*, vol. 36, no. 5, pp. 341–354, May 2022. [Online]. Available: https://link.springer.com/10.1007/s10822-021-00399-1

[24] J. L. Medina-Franco, A. L. Chávez-Hernández, E. López-López, and F. I. Saldívar-González, "Chemical Multiverse: An Expanded View of Chemical Space," *Molecular Informatics*, vol. 41, no. 11, p. 2200116, Nov. 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/minf.202200116

[25] J. L. Medina-Franco, "Interrogating novel areas of chemical space for drug discovery using chemoinformatics," *Drug Development Research*, vol. 73, no. 7, pp. 430–438, Oct. 2012. [Online]. Available: https://doi.org/10.1002/ddr.21034

[26] J. Medina-Franco, K. Martinez-Mayorga, M. Giulianotti, R. Houghten, and C. Pinilla, "Visualization of the chemical space in drug discovery," *Current Computer Aided-Drug Design*, vol. 4, no. 4, pp. 322–333, Dec. 2008. [Online]. Available: https://doi.org/10.2174/157340908786786010

[27] M. von Korff and K. Hilpert, "Assessing the predictive power of unsupervised visualization techniques to improve the identification of GPCR-focused compound libraries," *Journal of Chemical Information and Modeling*, vol. 46, no. 4, pp. 1580–1587, Apr. 2006. [Online]. Available: https://doi.org/10.1021/ci060037o

[28] D. I. Osolodkin, E. V. Radchenko, A. A. Orlov, A. E. Voronkov, V. A. Palyulin, and N. S. Zefirov, "Progress in visual representations of chemical space," *Expert Opinion on Drug Discovery*, vol. 10, no. 9, pp. 959–973, Jun. 2015. [Online]. Available: https://doi.org/10.1517/17460441.2015.1060216

[29] J. L. Medina-Franco, J. J. Naveja, and E. López-López, "Reaching for the bright StARs in chemical space," *Drug Discovery Today*, vol. 24, no. 11, pp. 2162–2169, Nov. 2019. [Online]. Available: https://doi.org/10.1016/j.drudis.2019.09.013

[30] G. M. Maggiora and J. Bajorath, "Chemical space networks: a powerful new paradigm for the description of chemical space," *Journal of Computer-Aided Molecular Design*, vol. 28, no. 8, pp. 795–802, Aug. 2014. [Online]. Available: http://link.springer.com/10.1007/s10822-014-9760-0

[31] A. de la Vega de León and J. Bajorath, "Chemical space visualization: transforming multidimensional chemical spaces into similarity-based molecular networks," *Future Medicinal Chemistry*, vol. 8, no. 14, pp. 1769–1778, Sep. 2016. [Online]. Available: https://doi.org/10.4155/fmc-2016-0023

[32] H. D. Boekhout, "Combining graph mining and deep learning in molecular activity prediction," 2015.

[33] M. Vogt, D. Stumpfe, G. M. Maggiora, and J. Bajorath, "Lessons learned from the design of chemical space networks and opportunities for new applications," *Journal of Computer-Aided Molecular Design*, vol. 30, no. 3, pp. 191–208, Mar. 2016. [Online]. Available: https://doi.org/10.1007/s10822-016-9906-3

[34] B. I. Díaz-Eufracio, O. Palomino-Hernández, R. A. Houghten, and J. L. Medina-Franco, "Exploring the chemical space of peptides for drug discovery: a

focus on linear and cyclic penta-peptides," *Molecular Diversity*, vol. 22, no. 2, pp. 259–267, Feb. 2018. [Online]. Available: https://doi.org/10.1007/s11030-018-9812-9

[35] A. Holzinger, C. Stocker, M. Bruschi, A. Auinger, H. Silva, H. Gamboa, and A. Fred, "On applying approximate entropy to ECG signals for knowledge discovery on the example of big sensor data," in *Active Media Technology*. Springer Berlin Heidelberg, 2012, pp. 646–657. [Online]. Available: https://doi.org/10.1007/978-3-642-35236-2_64

[36] A. Holzinger, M. Dehmer, and I. Jurisica, "Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions," *BMC Bioinformatics*, vol. 15, no. S6, p. I1, May 2014. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S6-I1

[37] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996. [Online]. Available: https://doi.org/10.1145/240455.240464

[38] A. Varnek and I. I. Baskin, "Chemoinformatics as a theoretical chemistry discipline," *Molecular Informatics*, vol. 30, no. 1, pp. 20–32, Jan. 2011. [Online]. Available: https://doi.org/10.1002/minf.201000100

[39] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data," *PLOS ONE*, vol. 10, no. 12, p. e0144059, Dec. 2015. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0144059

[40] M. L. Zepeda-Mendoza and O. Resendis-Antonio, "Bipartite Graph," in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer New York, 2013, pp. 147–148. [Online]. Available: http://link.springer.com/10.1007/978-1-4419-9863-7_1370

[41] M. Newman, *Networks: An Introduction*. Oxford University Press, 03 2010. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

[42] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Feb. 2004. [Online]. Available: https://doi.org/10.1103/physreve.69.026113

[43] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.0601602103

[44] L. A. Zahoránszky, G. Y. Katona, P. Hári, A. Málnási-Csizmadia, K. A. Zweig, and G. Zahoránszky-Köhalmi, "Breaking the hierarchy - a new cluster selection mechanism for hierarchical clustering methods," *Algorithms for Molecular Biology*, vol. 4, no. 1, p. 12, Dec. 2009. [Online]. Available: https://almob.biomedcentral.com/articles/10.1186/1748-7188-4-12

[45] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998. [Online]. Available: https://www.nature.com/articles/30918

[46] Q. K. Telesford, K. E. Joyce, S. Hayasaka, J. H. Burdette, and P. J. Laurienti, "The ubiquity of small-world networks," *Brain Connectivity*, vol. 1, no. 5, pp. 367–375, Dec. 2011. [Online]. Available: https://doi.org/10.1089/brain.2011.0038

[47] G. Zahoránszky-Kőhalmi, C. G. Bologa, and T. I. Oprea, "Impact of similarity threshold on the topology of molecular similarity networks and clustering outcomes," *Journal of Cheminformatics*, vol. 8, no. 1, p. 16, Dec. 2016. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-016-0127-5

[48] R. K. Kaur, M. Kaur, and A. Kaur, "Using cluster analysis for protein secondary structure prediction," *International Journal of Computer Applications*, vol. 4, no. 12, pp. 20–22, Aug. 2010. [Online]. Available: https://doi.org/10.5120/877-1248

[49] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, Feb. 2010. [Online]. Available: https://doi.org/10.1016/j.physrep.2009.11.002

[50] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct. 2008. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008

[51] U. Brandes and C. Pich, "CENTRALITY ESTIMATION IN LARGE NETWORKS," *International Journal of Bifurcation and Chaos*, vol. 17, no. 07, pp. 2303–2318, Jul. 2007. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218127407018403

[52] V. A. Traag, L. Waltman, and N. J. Van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, p. 5233, Mar. 2019. [Online]. Available: https://www.nature.com/articles/s41598-019-41695-z

[53] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, Jan. 1978. [Online]. Available: https://doi.org/10.1016/0378-8733(78)90021-7

[54] T. W. Valente, K. Coronges, C. M. Lakon, and E. Costenbader, "How correlated are network centrality measures?" *Connections*, vol. 28 1, pp. 16–26, 2008.

[55] S. Oldham, B. Fulcher, L. Parkes, A. Arnatkevic

iūtė, C. Suo, and A. Fornito, "Consistency and differences between centrality measures across distinct classes of networks," *PLOS ONE*, vol. 14, no. 7, p. e0220061, Jul. 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0220061

[56] M. Jalili, A. Salehzadeh-Yazdi, Y. Asgari, S. S. Arab, M. Yaghmaie, A. Ghavamzadeh, and K. Alimoghaddam, "CentiServer: A comprehensive resource, web-based application and r package for centrality analysis," *PLOS ONE*, vol. 10, no. 11, p. e0143111, Nov. 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0143111

[57] Z. Ghalmane, M. E. Hassouni, and H. Cherifi, "Immunization of networks with non-overlapping community structure," 2018, publisher: arXiv Version Number: 1. [Online]. Available: https://arxiv.org/abs/1806.05637

[58] S. Rajeh, M. Savonnet, E. Leclercq, and H. Cherifi, "Investigating Centrality Measures in Social Networks with Community Structure," 2022, publisher: arXiv Version Number: 1. [Online]. Available: https://arxiv.org/abs/2201.12914

[59] L. Aguilera-Mendoza, Y. Marrero-Ponce, C. R. García-Jacas, E. Chavez, J. A. Beltran, H. A. Guillen-Ramirez, and C. A. Brizuela, "Automatic construction of molecular similarity networks for visual graph mining in chemical space of bioactive peptides: an unsupervised learning approach," *Scientific Reports*, vol. 10, no. 1, p. 18074, Oct. 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-75029-1

[60] M. Marchiori and V. Latora, "Harmony in the small-world," *Physica A: Statistical Mechanics and its Applications*, vol. 285, no. 3-4, pp. 539–546, Oct. 2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378437100003113

[61] Y. Ruan, J. Tang, Y. Hu, H. Wang, and L. Bai, "Efficient Algorithm for the Identification of Node Significance in Complex Network," *IEEE Access*, vol. 8, pp. 28 947–28 955, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8985345/

[62] E. Chavez, S. Dobrev, E. Kranakis, J. Opatrny, L. Stacho, H. Tejeda, and J. Urrutia, "Half-Space Proximal: A New Local Test for Extracting a Bounded Dilation Spanner of a Unit Disk Graph," in *Principles of Distributed Systems*, J. H. Anderson, G. Prencipe, and R. Wattenhofer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 3974, pp. 235–245, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/11795490_19

[63] ——, "Half-space proximal: A new local test for extracting a bounded dilation spanner of a unit disk graph," in *Principles of Distributed Systems*, J. H. Anderson, G. Prencipe, and R. Wattenhofer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 235–245.

[64] C. Both, N. Dehmamy, R. Yu, and A.-L. Barabási, "Accelerating network layouts using graph neural networks," *Nature Communications*, vol. 14, no. 1, Mar. 2023. [Online]. Available: https://doi.org/10.1038/s41467-023-37189-2

[65] A. Noack, "Modularity clustering is force-directed layout," *Physical Review E*, vol. 79, no. 2, Feb. 2009. [Online]. Available: https://doi.org/10.1103/physreve.79.026102

[66] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, Nov. 1991. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/spe.4380211102

[67] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software," *PLoS ONE*, vol. 9, no. 6, p. e98679, Jun. 2014. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0098679

[68] J. Chao, F. Tang, and L. Xu, "Developments in algorithms for sequence alignment: A review," *Biomolecules*, vol. 12, no. 4, p. 546, Apr. 2022. [Online]. Available: https://doi.org/10.3390/biom12040546

[69] Z. Xia, Y. Cui, A. Zhang, T. Tang, L. Peng, C. Huang, C. Yang, and X. Liao, "A Review of Parallel Implementations for the Smith–Waterman Algorithm," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 14, no. 1, pp. 1–14, Mar. 2022. [Online]. Available: https://link.springer.com/10.1007/s12539-021-00473-0

[70] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, Mar. 1970. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0022283670900574

[71] F. Grisoni, D. Ballabio, R. Todeschini, and V. Consonni, *Molecular Descriptors for Structure–Activity Applications: A Hands-On Approach*. Springer New York, 2018, p. 3–53. [Online]. Available: http://dx.doi.org/10.1007/978-1-4939-7899-1_1

[72] "Thermostability and aliphatic index of globular proteins," *The Journal of Biochemistry*, Oct. 1980. [Online]. Available: https://doi.org/10.1093/oxfordjournals.jbchem.a133168

[73] H. G. Boman, "Antibacterial peptides: basic facts and emerging concepts," *Journal of Internal Medicine*, vol. 254, no. 3, pp. 197–215, Sep. 2003. [Online]. Available: https://doi.org/10.1046/j.1365-2796.2003.01228.x

[74] H. Cid, M. Bunster, M. Canales, and F. Gazitúa, "Hydrophobicity and structural classes in proteins," *"Protein Engineering, Design and Selection"*, vol. 5, no. 5, pp. 373–375, 1992. [Online]. Available: https://doi.org/10.1093/protein/5.5.373

[75] H. J. Cleaves, "Isoelectric point," in *Encyclopedia of Astrobiology*. Springer Berlin Heidelberg, 2011, pp. 858–858. [Online]. Available: https://doi.org/10.1007/978-3-642-11274-4_819

[76] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, May 1982. [Online]. Available: https://doi.org/10.1016/0022-2836(82)90515-0

[77] Y. C. Martin, J. L. Kofron, and L. M. Traphagen, "Do structurally similar molecules have similar biological activity?" *Journal of Medicinal Chemistry*, vol. 45, no. 19, pp. 4350–4358, Aug. 2002. [Online]. Available: https://doi.org/10.1021/jm020155c

[78] K. Castillo-Mendieta, G. Agüero-Chapin, N. Santiago Vispo, E. A. Márquez, Y. Perez-Castillo, S. J. Barigye, and Y. Marrero-Ponce, "Peptide Hemolytic Activity Analysis using Visual Data Mining of Similarity-based Complex Networks," MATHEMATICS & COMPUTER SCIENCE, preprint, Mar. 2023. [Online]. Available: https://www.preprints.org/manuscript/202303.0322/v1

[79] M. Romero, Y. Marrero-Ponce, H. Rodríguez, G. Agüero-Chapin, A. Antunes, L. Aguilera-Mendoza, and F. Martinez-Rios, "A novel network science and similarity-searching-based approach for discovering potential tumor-homing peptides from antimicrobials," *Antibiotics*, vol. 11, no. 3, p. 401, Mar. 2022. [Online]. Available: https://doi.org/10.3390/antibiotics11030401

[80] S. Ayala-Ruano, Y. Marrero-Ponce, L. Aguilera-Mendoza, N. Pérez, G. Agüero-Chapin, A. Antunes, and A. C. Aguilar, "Network science and group fusion similarity-based searching to explore the chemical space of antiparasitic peptides,"

ACS Omega, vol. 7, no. 50, pp. 46 012–46 036, Dec. 2022. [Online]. Available: https://doi.org/10.1021/acsomega.2c03398

[81] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, Dec. 1937. [Online]. Available: https://doi.org/10.1080/01621459.1937.10503522

[82] D. Chicco and G. Jurman, "The matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining*, vol. 16, no. 1, Feb. 2023. [Online]. Available: https://doi.org/10.1186/s13040-023-00322-4

[83] D. Chicco, N. Tötsch, and G. Jurman, "The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, no. 1, Feb. 2021. [Online]. Available: https://doi.org/10.1186/s13040-021-00244-z

[84] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan. 2020. [Online]. Available: https://doi.org/10.1186/s12864-019-6413-7

[85] A. D. Baxevanis, "The Importance of Biological Databases in Biological Discovery," *Current Protocols in Bioinformatics*, vol. 34, no. 1, Jun. 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0101s34

[86] G. Agarwal and R. Gabrani, "Antiviral Peptides: Identification and Validation," *International Journal of Peptide Research and Therapeutics*, vol. 27, no. 1, pp. 149–168, Mar. 2021. [Online]. Available: https://link.springer.com/10.1007/s10989-020-10072-0

[87] S. Ramazi, N. Mohammadi, A. Allahverdi, E. Khalili, and P. Abdolmaleki, "A review on antimicrobial peptides databases and the computational tools,"

Database, vol. 2022, p. baac011, Mar. 2022. [Online]. Available: https://academic.oup.com/database/article/doi/10.1093/database/baac011/6550847

[88] S. Seebah, A. Suresh, S. Zhuo, Y. H. Choong, H. Chua, D. Chuon, R. Beuerman, and C. Verma, "Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides," *Nucleic Acids Research*, vol. 35, no. Database, pp. D265–D268, Jan. 2007. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl866

[89] C. K. L. Wang, Q. Kaas, L. Chiche, and D. J. Craik, "CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering," *Nucleic Acids Research*, vol. 36, no. Database, pp. D206–D210, Dec. 2007. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm953

[90] J.-H. Jhong, L. Yao, Y. Pang, Z. Li, C.-R. Chung, R. Wang, S. Li, W. Li, M. Luo, R. Ma, Y. Huang, X. Zhu, J. Zhang, H. Feng, Q. Cheng, C. Wang, K. Xi, L.-C. Wu, T.-H. Chang, J.-T. Horng, L. Zhu, Y.-C. Chiang, Z. Wang, and T.-Y. Lee, "dbAMP 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data," *Nucleic Acids Research*, vol. 50, no. D1, pp. D460–D470, Jan. 2022. [Online]. Available: https://academic.oup.com/nar/article/50/D1/D460/6445964

[91] M. Pirtskhalava, A. A. Amstrong, M. Grigolava, M. Chubinidze, E. Alimbarashvili, B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D. E. Hurt, and M. Tartakovsky, "DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics," *Nucleic Acids Research*, vol. 49, no. D1, pp. D288–D297, Jan. 2021. [Online]. Available: https://academic.oup.com/nar/article/49/D1/D288/5957160

[92] X. Zhao, H. Wu, H. Lu, G. Li, and Q. Huang, "LAMP: A Database Linking Antimicrobial Peptides," *PLoS ONE*, vol. 8, no. 6, p. e66557, Jun. 2013. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0066557

[93] X. Kang, F. Dong, C. Shi, S. Liu, J. Sun, J. Chen, H. Li, H. Xu, X. Lao, and H. Zheng, "DRAMP 2.0, an updated data repository of antimicrobial peptides," *Scientific Data*, vol. 6, no. 1, p. 148, Aug. 2019. [Online]. Available: https://www.nature.com/articles/s41597-019-0154-y

[94] E. A. Gómez, P. Giraldo, and S. Orduz, "InverPep: A database of invertebrate antimicrobial peptides," *Journal of Global Antimicrobial Resistance*, vol. 8, pp. 13–17, Mar. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2213716516301291

[95] U. Gawde, S. Chakraborty, F. H. Waghu, R. S. Barai, A. Khanderkar, R. Indraguru, T. Shirsat, and S. Idicula-Thomas, "CAMPR4: a database of natural and synthetic antimicrobial peptides," *Nucleic Acids Research*, vol. 51, no. D1, pp. D377–D383, Jan. 2023. [Online]. Available: https://academic.oup.com/nar/article/51/D1/D377/6825344

[96] G. Wang, X. Li, and Z. Wang, "APD3: the antimicrobial peptide database as a tool for research and education," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1087–D1093, Jan. 2016. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1278

[97] A. Qureshi, N. Thakur, H. Tandon, and M. Kumar, "AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1147–D1153, Jan. 2014. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1191

[98] A. Qureshi, N. Thakur, and M. Kumar, "HIPdb: A Database of Experimentally Validated HIV Inhibiting Peptides," *PLoS ONE*, vol. 8, no. 1, p. e54908, Jan. 2013. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0054908

[99] Q. Zhang, X. Chen, B. Li, C. Lu, S. Yang, J. Long, H. Chen, J. Huang, and B. He, "A database of anti-coronavirus peptides," *Scientific Data*, vol. 9, no. 1, p. 294, Jun. 2022. [Online]. Available: https://www.nature.com/articles/s41597-022-01394-3

[100] A. Zouhir, O. Khamessi, S. Kamoun, C. Hkimi, H. Othman, A. Cherif, B. Mahjoubi, T. Jr, K. Sebei, and K. Ghedira, "AntiCoV-DB: A novel database resource of Anti COVID- 19, Anti CoronaVirus, Natural products and peptides," In Review, preprint, Feb. 2023. [Online]. Available: https://www.researchsquare.com/article/rs-2579195/v1

[101] Y. Liu, Y. Zhu, X. Sun, T. Ma, X. Lao, and H. Zheng, "DRAVP: A Comprehensive Database of Antiviral Peptides and Proteins," *Viruses*, vol. 15, no. 4, p. 820, Mar. 2023. [Online]. Available: https://www.mdpi.com/1999-4915/15/4/820

[102] L. Aguilera-Mendoza, Y. Marrero-Ponce, J. A. Beltran, R. Tellez Ibarra, H. A. Guillen-Ramirez, and C. A. Brizuela, "Graph-based data integration from bioactive peptide databases of pharmaceutical interest: toward an organized collection enabling visual network analysis," *Bioinformatics*, vol. 35, no. 22, pp. 4739–4747, Nov. 2019. [Online]. Available: https://academic.oup.com/bioinformatics/article/35/22/4739/5474901

[103] S. A. Pinacho-Castellanos, C. R. García-Jacas, M. K. Gilson, and C. A. Brizuela, "Alignment-Free Antimicrobial Peptide Predictors: Improving Performance by a Thorough Analysis of the Largest Available Data Set," *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 3141–3157, Jun. 2021. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jcim.1c00251

[104] Y. B. Ruiz-Blanco, G. Agüero-Chapin, S. Romero-Molina, A. Antunes, L.-R. Olari, B. Spellerberg, J. Münch, and E. Sanchez-Garcia, "Abp-finder: A tool to identify antibacterial peptides and the gram-staining type of targeted bacteria," *Antibiotics*, vol. 11, no. 12, p. 1708, Nov 2022. [Online]. Available: http://dx.doi.org/10.3390/antibiotics11121708

[105] G. Agüero-Chapin, A. Antunes, J. R. Mora, N. Pérez, E. Contreras-Torres, J. R. Valdes-Martini, F. Martinez-Rios, C. H. Zambrano, and Y. Marrero-Ponce, "Complex Networks Analyses of Antibiofilm Peptides: An Emerging Tool for Next Generation Antimicrobials Discovery," MEDICINE & PHARMACOLOGY, preprint, Mar. 2023. [Online]. Available: https://www.preprints.org/manuscript/202303.0193/v1

[106] H. ElAbd, Y. Bromberg, A. Hoarfrost, T. Lenz, A. Franke, and M. Wendorff, "Amino acid encoding for deep learning applications," *BMC Bioinformatics*, vol. 21, no. 1, Jun. 2020. [Online]. Available: https://doi.org/10.1186/s12859-020-03546-x

[107] S. Spänig and D. Heider, "Encodings and models for antimicrobial peptide classification for multi-resistant pathogens," *BioData Mining*, vol. 12, no. 1, p. 7, Dec. 2019. [Online]. Available: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-019-0196-x

[108] W. R. Pearson, "Selecting the right similarity-scoring matrix," *Current Protocols in Bioinformatics*, vol. 43, no. 1, Oct. 2013. [Online]. Available: https://doi.org/10.1002/0471250953.bi0305s43

[109] S. R. Eddy, "Where did the BLOSUM62 alignment score matrix come from?" *Nature Biotechnology*, vol. 22, no. 8, pp. 1035–1036, Aug. 2004. [Online]. Available: https://doi.org/10.1038/nbt0804-1035

[110] I. Diakou, E. Papakonstantinou, L. Papageorgiou, K. Pierouli, K. Dragoumani, D. Spandidos, F. Bacopoulou, G. Chrousos, E. Eliopoulos, and D. Vlachakis, "Novel computational pipelines in antiviral structure-based drug design (Review)," *Biomedical Reports*, vol. 17, no. 6, p. 97, Oct. 2022. [Online]. Available: http://www.spandidos-publications.com/10.3892/br.2022.1580

[111] J. Yan, J. Cai, B. Zhang, Y. Wang, D. F. Wong, and S. W. I. Siu, "Recent Progress in the Discovery and Design of Antimicrobial Peptides Using Traditional Machine Learning and Deep Learning," *Antibiotics*, vol. 11, no. 10, p. 1451, Oct. 2022. [Online]. Available: https://www.mdpi.com/2079-6382/11/10/1451

[112] S. Basith, B. Manavalan, T. Hwan Shin, and G. Lee, "Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening," *Medicinal Research Reviews*, vol. 40, no. 4, pp. 1276–1314, Jul. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/med.21658

[113] T.-T. Lin, Y.-Y. Sun, C.-T. Wang, W.-C. Cheng, I.-H. Lu, C.-Y. Lin, and S.-H. Chen, "AI4AVP: an antiviral peptides predictor in deep learning approach

with generative adversarial network data augmentation," *Bioinformatics Advances*, vol. 2, no. 1, p. vbac080, Jan. 2022. [Online]. Available: https://academic.oup.com/bioinformaticsadvances/article/doi/10.1093/bioadv/vbac080/6774982

[114] C. R. García-Jacas, S. A. Pinacho-Castellanos, L. A. García-González, and C. A. Brizuela, "Do deep learning models make a difference in the identification of antimicrobial peptides?" *Briefings in Bioinformatics*, vol. 23, no. 3, p. bbac094, May 2022. [Online]. Available: https://academic.oup.com/bib/article/doi/10.1093/bib/bbac094/6563422

[115] H. Kurata, S. Tsukiyama, and B. Manavalan, "iACVP: markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model," *Briefings in Bioinformatics*, vol. 23, no. 4, Jul. 2022. [Online]. Available: https://doi.org/10.1093/bib/bbac265

[116] M. Jaiswal, A. Singh, and S. Kumar, "PTPAMP: prediction tool for plant-derived antimicrobial peptides," *Amino Acids*, vol. 55, no. 1, pp. 1–17, Jul. 2022. [Online]. Available: https://doi.org/10.1007/s00726-022-03190-0

[117] R. Sharma, S. Shrivastava, S. K. Singh, A. Kumar, A. K. Singh, and S. Saxena, "Deep-avppred: Artificial intelligence driven discovery of peptide drugs for viral infections," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, pp. 5067–5074, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244660199

[118] Y. Pang, L. Yao, J.-H. Jhong, Z. Wang, and T.-Y. Lee, "AVPIden: a new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab263, Nov. 2021. [Online]. Available: https://academic.oup.com/bib/article/doi/10.1093/bib/bbab263/6323205

[119] P. B. Timmons and C. M. Hewage, "ENNAVIA is a novel method which employs neural networks for antiviral and anti-coronavirus activity prediction for therapeutic peptides," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab258, Nov.

2021. [Online]. Available: https://academic.oup.com/bib/article/doi/10.1093/bib/bbab258/6326528

[120] C.-R. Chung, T.-R. Kuo, L.-C. Wu, T.-Y. Lee, and J.-T. Horng, "Characterization and identification of antimicrobial peptides with different functional activities," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 1098–1114, May 2020. [Online]. Available: https://academic.oup.com/bib/article/21/3/1098/5498047

[121] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, and W. Shoombuatong, "Meta-iAVP: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation," *International Journal of Molecular Sciences*, vol. 20, no. 22, p. 5743, Nov. 2019. [Online]. Available: https://doi.org/10.3390/ijms20225743

[122] L. Wei, C. Zhou, R. Su, and Q. Zou, "PEPred-suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning," *Bioinformatics*, vol. 35, no. 21, pp. 4272–4280, Apr. 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz246

[123] W. Lin and D. Xu, "Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types," *Bioinformatics*, vol. 32, no. 24, pp. 3745–3752, Aug. 2016. [Online]. Available: https://doi.org/10.1093/bioinformatics/btw560

[124] A. Qureshi, H. Tandon, and M. Kumar, "AVP-ICsub50/subpred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (ICsub50/sub)," *Biopolymers*, vol. 104, no. 6, pp. 753–763, Nov. 2015. [Online]. Available: https://doi.org/10.1002/bip.22703

[125] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "iAMP-2l: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, May 2013. [Online]. Available: https://doi.org/10.1016/j.ab.2013.01.019

[126] S. Joseph, S. Karnik, P. Nilawe, V. K. Jayaraman, and S. Idicula-Thomas, "ClassAMP: A prediction tool for classification of antimicrobial peptides,"

IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 5, pp. 1535–1538, Sep. 2012. [Online]. Available: https://doi.org/10.1109/tcbb.2012.89

[127] N. Thakur, A. Qureshi, and M. Kumar, "AVPpred: collection and prediction of highly effective antiviral peptides," *Nucleic Acids Research*, vol. 40, no. W1, pp. W199–W204, Jul. 2012. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks450

[128] H. Zhang, K. M. Saravanan, Y. Wei, Y. Jiao, Y. Yang, Y. Pan, X. Wu, and J. Z. H. Zhang, "Deep learning-based bioactive therapeutic peptide generation and screening," *Journal of Chemical Information and Modeling*, vol. 63, no. 3, pp. 835–845, Feb. 2023. [Online]. Available: https://doi.org/10.1021/acs.jcim.2c01485

[129] S. Zhang and X. Li, "Pep-CNN: An improved convolutional neural network for predicting therapeutic peptides," *Chemometrics and Intelligent Laboratory Systems*, vol. 221, p. 104490, Feb. 2022. [Online]. Available: https://doi.org/10.1016/j.chemolab.2022.104490

[130] E. Otović, M. Njirjak, D. Kalafatovic, and G. Mauša, "Sequential properties representation scheme for recurrent neural network-based prediction of therapeutic peptides," *Journal of Chemical Information and Modeling*, vol. 62, no. 12, pp. 2961–2972, Jun. 2022. [Online]. Available: https://doi.org/10.1021/acs.jcim.2c00526

[131] Y. Pang, L. Yao, J. Xu, Z. Wang, and T.-Y. Lee, "Integrating transformer and imbalanced multi-label learning to identify antimicrobial peptides and their functional activities," *Bioinformatics*, vol. 38, no. 24, pp. 5368–5374, Nov. 2022. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac711

[132] Y. Pang, L. Yao, J.-H. Jhong, Z. Wang, and T.-Y. Lee, "AVPIden: a new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches," *Briefings in Bioinformatics*, vol. 22, no. 6, Jul. 2021. [Online]. Available: https://doi.org/10.1093/bib/bbab263

[133] X. Xiao, Y.-T. Shao, X. Cheng, and B. Stamatovic, "iAMP-CA2l: a new CNN-BiLSTM-SVM classifier based on cellular automata image for identifying

antimicrobial peptides and their functional types," *Briefings in Bioinformatics*, vol. 22, no. 6, Jun. 2021. [Online]. Available: https://doi.org/10.1093/bib/bbab209

[134] A. S. Chowdhury, S. M. Reehl, K. Kehn-Hall, B. Bishop, and B.-J. M. Webb-Robertson, "Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance," *Scientific Reports*, vol. 10, no. 1, Nov. 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-76161-8

[135] J. F. B. Lissabet, L. H. Belén, and J. G. Farias, "AntiVPP 1.0: A portable tool for prediction of antiviral peptides," *Computers in Biology and Medicine*, vol. 107, pp. 127–130, Apr. 2019. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2019.02.011

[136] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," 2009, publisher: Unpublished. [Online]. Available: http://rgdoi.net/10.13140/2.1.1341.1520

[137] D. Osorio, P. Rondón-Villarreal, and R. Torres, "Peptides: A Package for Data Mining of Antimicrobial Peptides," *The R Journal*, vol. 7, no. 1, p. 4, 2015. [Online]. Available: https://journal.r-project.org/archive/2015/RJ-2015-001/index.html

[138] L. Aguilera-Mendoza, Y. Marrero-Ponce, R. Tellez-Ibarra, M. T. Llorente-Quesada, J. Salgado, S. J. Barigye, and J. Liu, "Overlap and diversity in antimicrobial peptide databases: compiling a non-redundant set of sequences," *Bioinformatics*, vol. 31, no. 15, pp. 2553–2559, Aug. 2015. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv180

[139] T. L. Bailey, "STREME: accurate and versatile sequence motif discovery," *Bioinformatics*, vol. 37, no. 18, pp. 2834–2840, Sep. 2021. [Online]. Available: https://academic.oup.com/bioinformatics/article/37/18/2834/6184861

[140] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, "The MEME Suite," *Nucleic Acids Research*, vol. 43, no. W1, pp. W39–W49, Jul. 2015. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv416

[141] T. L. Bailey and C. E. Grant, "SEA: Simple Enrichment Analysis of motifs," Bioinformatics, preprint, Aug. 2021. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2021.08.23.457422

[142] Y. Pang, Z. Wang, J.-H. Jhong, and T.-Y. Lee, "Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1085–1095, Mar. 2021. [Online]. Available: https://academic.oup.com/bib/article/22/2/1085/6120286

[143] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, and F. Herrera, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Computing*, vol. 13, no. 3, pp. 307–318, May 2008. [Online]. Available: https://doi.org/10.1007/s00500-008-0323-y

[144] K. Y. Chang and J.-R. Yang, "Analysis and prediction of highly effective antiviral peptides based on random forests," *PLoS ONE*, vol. 8, no. 8, p. e70166, Aug. 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0070166

[145] H. Kurata, S. Tsukiyama, and B. Manavalan, "iACVP: markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model," *Briefings in Bioinformatics*, vol. 23, no. 4, p. bbac265, Jul. 2022. [Online]. Available: https://academic.oup.com/bib/article/doi/10.1093/bib/bbac265/6623727

[146] J. F. Beltrán Lissabet, L. H. Belén, and J. G. Farias, "AntiVPP 1.0: A portable tool for prediction of antiviral peptides," *Computers in Biology and Medicine*, vol. 107, pp. 127–130, Apr. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482519300587

[147] D. Almeida, D. Domínguez-Pérez, A. Matos, G. Agüero-Chapin, H. Osório, V. Vasconcelos, A. Campos, and A. Antunes, "Putative antimicrobial peptides of the posterior salivary glands from the cephalopod octopus vulgaris revealed by exploring a composite protein database," *Antibiotics*, vol. 9, no. 11, p. 757, Oct. 2020. [Online]. Available: https://doi.org/10.3390/antibiotics9110757

[148] S. Gupta, P. Kapoor, K. Chaudhary, A. Gautam, R. Kumar, and G. P. S. Raghava, "Peptide toxicity prediction," in *Methods in Molecular Biology*. Springer New York, Dec. 2014, pp. 143–157. [Online]. Available: https://doi.org/10.1007/978-1-4939-2285-7_7

[149] T. S. Win, A. A. Malik, V. Prachayasittikul, J. E. S. Wikberg, C. Nantasenamat, and W. Shoombuatong, "HemoPred: a web server for predicting the hemolytic activity of peptides," *Future Medicinal Chemistry*, vol. 9, no. 3, pp. 275–291, Mar. 2017. [Online]. Available: https://doi.org/10.4155/fmc-2016-0188

[150] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, and G. P. S. Raghava, "AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes," *Briefings in Bioinformatics*, Nov. 2020. [Online]. Available: https://doi.org/10.1093/bib/bbaa294

[151] S. Singh, K. Chaudhary, S. K. Dhanda, S. Bhalla, S. S. Usmani, A. Gautam, A. Tuknait, P. Agrawal, D. Mathur, and G. P. Raghava, "SATPdb: a database of structurally annotated therapeutic peptides," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1119–D1126, Jan. 2016. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1114

[152] G. Gogoladze, M. Grigolava, B. Vishnepolsky, M. Chubinidze, P. Duroux, M.-P. Lefranc, and M. Pirtskhalava, "dbaasp : database of antimicrobial activity and structure of peptides," *FEMS Microbiology Letters*, vol. 357, no. 1, pp. 63–68, Aug. 2014. [Online]. Available: https://academic.oup.com/femsle/article-lookup/doi/10.1111/1574-6968.12489

[153] G. Shi, X. Kang, F. Dong, Y. Liu, N. Zhu, Y. Hu, H. Xu, X. Lao, and H. Zheng, "DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides," *Nucleic Acids Research*, vol. 50, no. D1, pp. D488–D496, Jan. 2022. [Online]. Available: https://academic.oup.com/nar/article/50/D1/D488/6352447

[154] J. Théolier, I. Fliss, J. Jean, and R. Hammami, "MilkAMP: a comprehensive database of antimicrobial peptides of dairy origin," *Dairy Science & Technology*,

vol. 94, no. 2, pp. 181–193, Mar. 2014. [Online]. Available: http://link.springer.com/10.1007/s13594-013-0153-2

[155] I. R. Chandrashekar and S. M. Cowsik, "Three-Dimensional Structure of the Mammalian Tachykinin Peptide Neurokinin A Bound to Lipid Micelles," *Biophysical Journal*, vol. 85, no. 6, pp. 4002–4011, Dec. 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006349503748140

[156] J. Dong, S. Wang, and J. York, "Crystal structure of yeast N-terminal acetyltransferase NatE (IP6) in complex with a bisubstrate To Be Published,." [Online]. Available: 10.2210/pdb4xnh/pdb

[157] D. Ireland, M. Colgrave, and D. Craik, "A novel suite of cyclotides from *Viola odorata* : sequence variation and the implications for structure, function and stability," *Biochemical Journal*, vol. 400, no. 1, pp. 1–12, Nov. 2006. [Online]. Available: https://portlandpress.com/biochemj/article/400/1/1/93899/A-novel-suite-of-cyclotides-from-Viola-odorata

[158] S.-H. Cho, B.-D. Lee, H. An, and J.-B. Eun, "Kenojeinin I, antimicrobial peptide isolated from the skin of the fermented skate, Raja kenojei," *Peptides*, vol. 26, no. 4, pp. 581–587, Apr. 2005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0196978104005108

[159] A. Op De Beeck, C. Voisset, B. Bartosch, Y. Ciczora, L. Cocquerel, Z. Keck, S. Foung, F.-L. Cosset, and J. Dubuisson, "Characterization of Functional Hepatitis C Virus Envelope Glycoproteins," *Journal of Virology*, vol. 78, no. 6, pp. 2994–3002, Mar. 2004. [Online]. Available: https://journals.asm.org/doi/10.1128/JVI.78.6.2994-3002.2004

[160] Y. Kato, T. Aizawa, H. Hoshino, K. Kawano, K. Nitta, and H. Zhang, "abf-1 and abf-2, ASABF-type antimicrobial peptide genes in Caenorhabditis elegans," *Biochemical Journal*, vol. 361, no. 2, pp. 221–230, Jan. 2002. [Online]. Available: https://portlandpress.com/biochemj/article/361/2/221/39666/abf-1-and-abf-2-ASABF-type-antimicrobial-peptide

[161] X. Li, L. Wang, D. Zhao, G. Zhang, J. Luo, R. Deng, and Y. Yang, "Identification of host cell binding peptide from an overlapping peptide library for inhibition of classical swine fever virus infection," *Virus Genes*, vol. 43, no. 1, pp. 33–40, Aug. 2011. [Online]. Available: http://link.springer.com/10.1007/s11262-011-0595-7

[162] C. Loose, K. Jensen, I. Rigoutsos, and G. Stephanopoulos, "A linguistic model for the rational design of antimicrobial peptides," *Nature*, vol. 443, no. 7113, pp. 867–869, Oct. 2006. [Online]. Available: https://www.nature.com/articles/nature05233

[163] P. R. Hall, B. Hjelle, H. Njus, C. Ye, V. Bondu-Hawkins, D. C. Brown, K. A. Kilpatrick, and R. S. Larson, "Phage Display Selection of Cyclic Peptides That Inhibit Andes Virus Infection," *Journal of Virology*, vol. 83, no. 17, pp. 8965–8969, Sep. 2009. [Online]. Available: https://journals.asm.org/doi/10.1128/JVI.00606-09

[164] S. Lombardi, C. Massi, E. Indino, C. L. Rosa, P. Mazzetti, M. L. Falcone, P. Rovero, A. Fissi, O. Pieroni, P. Bandecchi, F. Esposito, F. Tozzini, M. Bendinelli, and C. Garzelli, "Inhibition of Feline Immunodeficiency Virus Infectionin Vitroby Envelope Glycoprotein Synthetic Peptides," *Virology*, vol. 220, no. 2, pp. 274–284, Jun. 1996. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0042682296903153

[165] C. Treffers, L. Chen, R. C. Anderson, and P.-L. Yu, "Isolation and characterisation of antimicrobial peptides from deer neutrophils," *International Journal of Antimicrobial Agents*, vol. 26, no. 2, pp. 165–169, Aug. 2005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924857905001457

[166] B. J. Owens, G. M. Anantharamaiah, J. B. Kahlon, R. V. Srinivas, R. W. Compans, and J. P. Segrest, "Apolipoprotein A-I and its amphipathic helix peptide analogues inhibit human immunodeficiency virus-induced syncytium formation." *Journal of Clinical Investigation*, vol. 86, no. 4, pp. 1142–1150, Oct. 1990. [Online]. Available: http://www.jci.org/articles/view/114819

[167] K. Shiba, "Natural and artificial peptide motifs: their origins and the application of motif-programming," *Chem. Soc. Rev.*, vol. 39, no. 1, pp. 117–126, 2010. [Online]. Available: http://xlink.rsc.org/?DOI=B719081F

[168] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002. [Online]. Available: https://www.science.org/doi/10.1126/science.298.5594.824

[169] P. Tompa, N. Davey, T. Gibson, and M. Babu, "A Million Peptide Motifs for the Molecular Biologist," *Molecular Cell*, vol. 55, no. 2, pp. 161–169, Jul. 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1097276514005620

[170] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. D. Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell, "Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks," *PLoS Biology*, vol. 3, no. 12, p. e405, Nov. 2005. [Online]. Available: https://dx.plos.org/10.1371/journal.pbio.0030405

[171] Y.-J. Liu, C.-S. Cheng, S.-M. Lai, M.-P. Hsu, C.-S. Chen, and P.-C. Lyu, "Solution structure of the plant defensin VrD1 from mung bean and its possible role in insecticidal activity against bruchids," *Proteins: Structure, Function, and Bioinformatics*, vol. 63, no. 4, pp. 777–786, Mar. 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/prot.20962

[172] F. T. Lay, H. J. Schirra, M. J. Scanlon, M. A. Anderson, and D. J. Craik, "The Three-dimensional Solution Structure of NaD1, a New Floral Defensin from Nicotiana alata and its Application to a Homology Model of the Crop Defense Protein alfAFP," *Journal of Molecular Biology*, vol. 325, no. 1, pp. 175–188, Jan. 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022283602011038

[173] T. V. Ovchinnikova, S. V. Balandin, G. M. Aleshina, A. A. Tagaev, Y. F. Leonova, E. D. Krasnodembsky, A. V. Men'shenin, and V. N. Kokryakov, "Aurelin, a novel antimicrobial peptide from jellyfish Aurelia aurita with structural features of defensins and channel-blocking toxins," *Biochemical and Biophysical Research Communications*, vol. 348, no. 2, pp. 514–523, Sep. 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006291X06016305

[174] D. J. Craik, N. L. Daly, T. Bond, and C. Waine, "Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif,"

Journal of Molecular Biology, vol. 294, no. 5, pp. 1327–1336, Dec. 1999. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022283699933831

[175] U. Göransson, T. Luijendijk, S. Johansson, L. Bohlin, and P. Claeson, "Seven Novel Macrocyclic Polypeptides from *Viola a rvensis*," *Journal of Natural Products*, vol. 62, no. 2, pp. 283–286, Feb. 1999. [Online]. Available: https://pubs.acs.org/doi/10.1021/np9803878

[176] X. Yang, W.-H. Lee, and Y. Zhang, "Extremely Abundant Antimicrobial Peptides Existed in the Skins of Nine Kinds of Chinese Odorous Frogs," *Journal of Proteome Research*, vol. 11, no. 1, pp. 306–319, Jan. 2012. [Online]. Available: https://pubs.acs.org/doi/10.1021/pr200782u

[177] Y. J. Basir, F. C. Knoop, J. Dulka, and J. Conlon, "Multiple antimicrobial peptides and peptides related to bradykinin and neuromedin N isolated from skin secretions of the pickerel frog, Rana palustris," *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, vol. 1543, no. 1, pp. 95–105, Nov. 2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167483800001916

[178] J. M. Conlon, A. Demandt, P. F. Nielsen, J. Leprince, H. Vaudry, and D. C. Woodhams, "The alyteserins: Two families of antimicrobial peptides from the skin secretions of the midwife toad Alytes obstetricans (Alytidae)," *Peptides*, vol. 30, no. 6, pp. 1069–1073, Jun. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S019697810900093X

[179] J. M. Conlon, M. Mechkarska, K. Arafat, S. Attoub, and A. Sonnevend, "Analogues of the frog skin peptide alyteserin-2a with enhanced antimicrobial activities against Gram-negative bacteria: ALYTESERIN-2: STRUCTURE-ACTIVITY," *Journal of Peptide Science*, vol. 18, no. 4, pp. 270–275, Apr. 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/psc.2397

[180] C. Muñoz-Camargo, V. Salazar, L. Barrero-Guevara, S. Camargo, A. Mosquera, H. Groot, and E. Boix, "Unveiling the Multifaceted Mechanisms of Antibacterial

Activity of Buforin II and Frenatin 2.3S Peptides from Skin Micro-Organs of the Orinoco Lime Treefrog (Sphaenorhynchus lacteus)," *International Journal of Molecular Sciences*, vol. 19, no. 8, p. 2170, Jul. 2018. [Online]. Available: http://www.mdpi.com/1422-0067/19/8/2170

[181] A. Agopian, E. Gros, G. Aldrian-Herrada, N. Bosquet, P. Clayette, and G. Divita, "A New Generation of Peptide-based Inhibitors Targeting HIV-1 Reverse Transcriptase Conformational Flexibility," *Journal of Biological Chemistry*, vol. 284, no. 1, pp. 254–264, Jan. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0021925820683132

[182] J. Münch, L. Ständker, K. Adermann, A. Schulz, M. Schindler, R. Chinnadurai, S. Pöhlmann, C. Chaipan, T. Biet, T. Peters, B. Meyer, D. Wilhelm, H. Lu, W. Jing, S. Jiang, W.-G. Forssmann, and F. Kirchhoff, "Discovery and Optimization of a Natural HIV-1 Entry Inhibitor Targeting the gp41 Fusion Peptide," *Cell*, vol. 129, no. 2, pp. 263–275, Apr. 2007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867407003285

[183] Y. Umetsu, T. Tenno, N. Goda, M. Shirakawa, T. Ikegami, and H. Hiroaki, "Structural difference of vasoactive intestinal peptide in two distinct membrane-mimicking environments," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1814, no. 5, pp. 724–730, May 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1570963911000604

[184] D. Krebs, R. G. Maroun, F. Sourgen, F. Troalen, D. Davoust, and S. Fermandjian, "Helical and coiled-coil-forming properties of peptides derived from and inhibiting human immunodeficiency virus type 1 integrase assessed by 1H-NMR . Use of NH temperature coefficients to probe coiled-coil structures," *European Journal of Biochemistry*, vol. 253, no. 1, pp. 236–244, Apr. 1998. [Online]. Available: http://doi.wiley.com/10.1046/j.1432-1327.1998.2530236.x

[185] M. K. Spriggs, R. A. Olmsted, S. Venkatesan, J. E. Coligan, and P. L. Collins, "Fusion glycoprotein of human parainfluenza virus type 3: Nucleotide sequence of the gene, direct identification of the cleavage-activation site, and comparison with

other paramyxoviruses," *Virology*, vol. 152, no. 1, pp. 241–251, Jul. 1986. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0042682286903880

[186] R. M. Verly, C. M. D. Moraes, J. M. Resende, C. Aisenbrey, M. P. Bemquerer, D. Piló-Veloso, A. P. Valente, F. C. Almeida, and B. Bechinger, "Structure and Membrane Interactions of the Antibiotic Peptide Dermadistinctin K by Multidimensional Solution and Oriented 15N and 31P Solid-State NMR Spectroscopy," *Biophysical Journal*, vol. 96, no. 6, pp. 2194–2203, Mar. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006349509003191

[187] O. Lequin, A. Ladram, L. Chabbert, F. Bruston, O. Convert, D. Vanhoye, G. Chassaing, P. Nicolas, and M. Amiche, "Dermaseptin S9, an -Helical Antimicrobial Peptide with a Hydrophobic Core and Cationic Termini," *Biochemistry*, vol. 45, no. 2, pp. 468–480, Jan. 2006. [Online]. Available: https://pubs.acs.org/doi/10.1021/bi051711i

[188] M. Jabeen, P. Biswas, M. T. Islam, and R. Paul, "Antiviral Peptides in Antimicrobial Surface Coatings—From Current Techniques to Potential Applications," *Viruses*, vol. 15, no. 3, p. 640, Feb. 2023. [Online]. Available: https://www.mdpi.com/1999-4915/15/3/640

[189] B. Manavalan, S. Basith, and G. Lee, "Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting SARS-CoV-2," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab412, Jan. 2022. [Online]. Available: https://academic.oup.com/bib/article/doi/10.1093/bib/bbab412/6378599

[190] S. Singh, H. N. Banavath, P. Godara, B. Naik, V. Srivastava, and D. Prusty, "Identification of antiviral peptide inhibitors for receptor binding domain of SARS-CoV-2 omicron and its sub-variants: an in-silico approach," *3 Biotech*, vol. 12, no. 9, p. 198, Sep. 2022. [Online]. Available: https://link.springer.com/10.1007/s13205-022-03258-4

[191] A. Moretta, R. Salvia, C. Scieuzo, A. Di Somma, H. Vogel, P. Pucci, A. Sgambato, M. Wolff, and P. Falabella, "A bioinformatic study of antimicrobial

peptides identified in the Black Soldier Fly (BSF) Hermetia illucens (Diptera: Stratiomyidae)," *Scientific Reports*, vol. 10, no. 1, p. 16875, Oct. 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-74017-9

[192] A. Tucs, D. P. Tran, A. Yumoto, Y. Ito, T. Uzawa, and K. Tsuda, "Generating Ampicillin-Level Antimicrobial Peptides with Activity-Aware Generative Adversarial Networks," *ACS Omega*, vol. 5, no. 36, pp. 22 847–22 851, Sep. 2020. [Online]. Available: https://pubs.acs.org/doi/10.1021/acsomega.0c02088

[193] W. Shi, Z. Qi, C. Pan, N. Xue, A. K. Debnath, J. Qie, S. Jiang, and K. Liu, "Novel anti-HIV peptides containing multiple copies of artificially designed heptad repeat motifs," *Biochemical and Biophysical Research Communications*, vol. 374, no. 4, pp. 767–772, Oct. 2008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006291X08014629

[194] A. Oddo and P. R. Hansen, "Hemolytic activity of antimicrobial peptides," in *Methods in Molecular Biology*. Springer New York, Dec. 2016, pp. 427–435. [Online]. Available: https://doi.org/10.1007/978-1-4939-6737-7_31

[195] A. S. Rathore, A. Arora, S. Choudhury, P. Tijare, and G. P. S. Raghava, "ToxinPred 3.0: An improved method for predicting the toxicity of peptides," Aug. 2023. [Online]. Available: https://doi.org/10.1101/2023.08.11.552911

[196] R. E. Hancock, "Cationic peptides: effectors in innate immunity and novel antimicrobials," *The Lancet Infectious Diseases*, vol. 1, no. 3, pp. 156–164, Oct. 2001. [Online]. Available: https://doi.org/10.1016/s1473-3099(01)00092-5

[197] R. E. Hancock and R. Lehrer, "Cationic peptides: a new source of antibiotics," *Trends in Biotechnology*, vol. 16, no. 2, pp. 82–88, Feb. 1998. [Online]. Available: https://doi.org/10.1016/s0167-7799(97)01156-6