UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Químicas e Ingeniería

Título: Virtual Screening of Antiangiogenic Cancer Treatment Peptides Based on Network Science and a Similarity Searching Based Approach

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

Autora:

Andrea Karolina Valarezo Albán

Tutora:

Ph.D Hortensia Rodríguez

Co-tutor:

Ph.D Yovani Marrero-Ponce

Urcuquí, Mayo - 2024

Autoría

Yo, **Andrea Karolina Valarezo Albán**, con cédula de identidad 1724867526, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así como, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el/la autor(a) del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Mayo - 2024.

Andrea Karolina Valarezo Albán

1724867526

Autorización de publicación

Yo, **Andrea Karolina Valarezo Albán**, con cédula de identidad 1724867526, cedo a la Universidad 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 Superior.

Urcuquí, Mayo - 2024.

Andrea Karolina Valarezo Albán

1724867526

This page is intentionally left blank.

This work is completely dedicated to my dad, who has always been my source of inspiration, support, and guidance.

Acknowledgements

Throughout my journey, I gained knowledge and experience, experienced significant personal growth, and set new goals for myself. I am deeply grateful to everyone who has been part of this transformative experience.

My greatest thanks, especially to my parents, Inés and Rene, whose unwavering support, sacrifices and infinite love have been my pillars of strength. To my brothers Josue and Mateo, who have been unconditional companions, offering me support and encouragement at all times. I want to express my deep appreciation to my friends Pablo, Sol, Joss, Nathy, and Mabe, whose companionship, guidance, and shared experiences have enriched my journey and filled my life with joy.

A special acknowledgment to my esteemed advisors, Ph.D. Hortensia Rodriguez and Ph.D. Yovani Marrero-Ponce, for their invaluable guidance, mentorship and introduction to the intricate fields of therapeutic peptides and bioinformatics. His unwavering support and confidence in my abilities have been instrumental in my development.

In closing, I would like to express my gratitude to the School of Chemical Sciences and Engineering and the committed professors who have taught and guided me throughout the Common Core curriculum, providing me with the essential foundation for developing my undergraduate thesis.

Andrea Karolina Valarezo Albán

This page is intentionally left blank.

Resumen

Los tratamientos convencionales contra el cáncer, como la quimioterapia o los fármacos quimioterapéuticos, presentan limitaciones debido a su toxicidad, resistencia y baja especificidad. La investigación actual se centra en terapias alternativas, destacando la importancia del soporte vascular en el crecimiento tumoral. El objetivo es inhibir la angiogénesis tumoral, explorando moléculas antiangiogénicas como péptidos y proteínas, que ofrecen ventajas sobre los fármacos convencionales. Debido a la escasez de fármacos antiangiogénicos, la investigación propuesta pretende examinar las secuencias de péptidos antiangiogénicos (AAP) notificadas mediante un enfoque novedoso que se basa en la ciencia de redes y la minería de datos. Para abordar este trabajo, se construyeron redes de metadatos (MENTs) cuyo análisis permite profundizar en el análisis, proporcionando una mejor comprensión de las características biológicas, los patrones y las relaciones generales entre las AAPs. Además, se examinó el espacio químico de los PAA mediante redes espaciales proximales (HSPNs) utilizando seis medidas de disimilitud bidireccionales, y el examen del efecto del punto de corte (t). A continuación, para garantizar la diversidad y evitar la sobrerrepresentación de las redes, se realizó una extracción de andamiajes que dio lugar a 96 subconjuntos. Este proceso ayudó a limitar el trabajo al análisis de tres métricas y a considerar para cada una sólo t = 0,00. Por último, se descubrieron y enriquecieron 37 motivos antiangiogénicos potenciales, lo que proporcionó información valiosa para el futuro desarrollo de fármacos más eficaces y selectivos, minimizando los efectos secundarios y mejorando la eficiencia de los tratamientos contra el cáncer.

Palabras clave: peptido antiangiogénico, cancer, ciencia de redes, minería visual, espacio Químico, redes de espacio proximal, redes de metadatos, descubrimiento de motivos, enriquecimiento de motivos, caja de herramientas StarPep.

Abstract

Conventional cancer treatments, such as chemotherapy or chemotherapeutic drugs, face limitations due to their toxicity, resistance and low specificity. Current research is focused on alternative therapies, highlighting the importance of vascular support in tumor growth. The aim is to inhibit tumor angiogenesis, exploring anti-angiogenic molecules such as peptides and proteins, which offer advantages over conventional drugs. Due to the scarce supply of anti-angiogenic drugs, the proposed research aims to examine reported anti-angiogenic peptide sequences (AAPs) using a novel approach that relies on network science and data mining. To address this work, metadata networks (MENTs) were constructed whose analysis allows for deeper analysis, providing an improved understanding of biological characteristics, patterns, and general relationships between AAPs. In addition, the chemical space of AAPs was examined through proximal space networks (HSPNs) using six bidirectional dissimilarity measures and examination of the effect of the cutoff point (t). Then, a scaffold extraction was performed to ensure diversity and avoid overrepresenting the networks, resulting in 96 subsets. This process helped to limit the work to the analysis of three metrics and to consider for each one only t=0.00. Finally, 37 potential anti-angiogenic motifs were discovered and enriched, providing valuable information for the future development of more effective and selective drugs, minimizing side effects and improving the efficiency of cancer treatments.

Keywords: antiangiogenic peptide, cancer, network science, visual mining, chemical space, proximal space networks, metadata networks, motif discovery, motif enrichment, StarPep toolbox.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer, a formidable adversary to human health, stands as the second leading cause of death globally [1]. Contemporary cancer treatments, including radiotherapy, chemotherapy, and surgical procedures, have made remarkable strides. However, they often fall short due to their indiscriminate effects on both cancer and healthy cells, leading to debilitating side effects [2]. The need for innovative, less invasive alternatives is palpable in the medical and scientific community [3].

This quest for novel approaches has spotlighted the potential of peptides in cancer research [4]. Peptides offer enticing advantages such as precise targeting. Unlike proteins, peptides possess a more compact structure and shorter length, facilitating easier access to specific sites within cells or tissues. Additionally they boast low toxicity, and relatively short half-lives in the body [5]. Harnessing these attributes could revolutionize cancer diagnosis and treatment [6]. Notably, Anti-Angiogenic Peptides (AAPs) exhibit promise by efficiently inhibiting tumor angiogenesis, a pivotal process fueling cancer growth [7]. Angiogenesis is regarded as a hallmark of cancer, outlining the fundamental traits of cancer cells [8, 9].

Angiogenesis, the formation of new blood vessels from existing ones, plays a pivotal role in tumor development by providing nourishment and oxygen to cancer cells [10]. Angiogenesis arises from an imbalance between various endogenous factors, some promoters and others inhibitors, that regulate this crucial process [11]. Proangiogenic factors include vascular endothelial growth factor (VEGF) [12], basic fibroblast growth factor (bFGF) [13], angiogenin (ANG) [14], transforming growth factor (TGF)[15], tumor necrosis factor (TNF) [16], platelet-derived growth factor (PDGF) [17], placental growth factor (PGF) [18], interleukin-8 (IL-8) [19], hepatocyte growth factor [20] and epidermal growth factor (EGF) [21], among others. On the other hand, there are endogenous direct angiogenic inhibitors, which are proteins or protein fragments naturally produced by the extracellular matrix which limit angiogenesis. In addition, exogenous indirect angiogenic inhibitors negatively regulate the expression or action of proangiogenic agents such as VEGF or epidermal growth factor receptor (EGFR). These intricate mechanisms illustrate the com-

plexity of angiogenesis and underscore the need to thoroughly understand these processes to develop more effective therapeutic approaches in the fight against cancer [11]. AAPs, by impeding angiogenesis, hold promise as a therapeutic avenue. They curtail the blood supply to cancer cells, limiting their proliferation. This presents hope for more effective and less invasive cancer treatments [7].

Methods of obtaining these peptides vary, involving extraction from natural sources or synthesis from mimotopes. Both approaches necessitate extensive in vitro and in vivo evaluation, consuming considerable time and resources[22]. To expedite this process, researchers frequently employ preliminary in silico studies for drug discovery [23]. Databases, web-available tools, and computational software are indispensable bioinformatics resources for uncovering novel drugs. Despite abundant peptide data in public databases, the identification of AAPs has progressed slowly due to labor-intensive experimental efforts[24] [24]. Noteworthy databases include the Benchmark Dataset (also called as AntiangioPred DB), and StarPep DB, housing vital data for angiogenesis research [25, 26].

Machine learning (ML) emerges as a cost-effective screening tool for peptide-based drugs, given its ability to handle vast datasets [27]. However, the application of network science to comprehensively explore AAPs' feature space remains uncharted. This study aims to address this gap, unraveling the defining features that confer antiangiogenic properties to therapeutic peptides. Chemical space is the fundamental foundation of cheminformatics, being an essential concept for drug discovery and beyond [28]. In addition, chemical space is closely related to computational chemogenomics, a discipline that pursues the prediction (followed by experimental validation) of the intersection between chemical space and those biologically significant molecules [29].

To understand chemical space fully, explore related concepts like similarity, diversity, and graphical representation via similarity networks. Chemical space is an abstract representation where each point signifies a molecule or chemical compound [18, 20], in our case, focusing on peptides. These points are distributed according to the similarities and differences between the molecules. This opens the door to comparing and analyzing chemical compounds based on aspects such as their structure, reactivity, properties, and much more [30].

Chemical similarity, on the other hand, refers to the degree to which two molecules

are similar in terms of structure and physico-chemical properties. Various metrics and approaches can be used to quantify this similarity [31]. In the same context of chemical space, chemical diversity is of fundamental importance. It involves representing a wide variety of compounds within the chemical space. This aspect plays a crucial role in drug discovery and molecule optimization, as it seeks to ensure that the molecules selected for experimental testing are as diverse as possible [32]. Chemical diversity plays an important role in minimizing the risk of bias and allowing a more thorough exploration of the chemical space in search of new substances with the desired properties [33]. On the other hand, similarity networks are presented as graphical representations of similarity relationships between chemical compounds within chemical space [34]. In these networks, compounds are represented as nodes, and connections between nodes indicate significant similarities.

These connections are based on similarity calculations and can be used to identify groups of similar molecules or to visualize patterns in chemical data sets [35]. Graphical representations of chemical space and similarity networks are essential for visually understanding the relationships between chemical compounds. In these representations, molecules can be shown as points in a multidimensional space, and connections between similar molecules are represented by lines or arcs in similarity networks [32]. These visual representations make it easier for scientists to explore chemical space's structure and diversity and identify possible directions for research and compound design.

Our study delves into AAPs chemical space using StarPep DB and StarPep toolbox [26, 36]. These tools enable visual analysis of chemical space networks (CSNs) and half-space proximate networks (HSPNs). CSNs have been proposed as coordinate-free representations to analyze and visualize chemical space without reducing its dimensionality [34]. The similarity between peptides is using alignment-free metrics derived from their sequence descriptors [37]. Nonetheless, we acknowledge the significance of three-dimensional structure. Integrating structural data could enhance our understanding of biological activities. In contrast, HSPNs offer an advantage over CSNs with fewer connections, significantly reducing processing time and computational load. This is because this specific network considers only a subset of the relationships between nodes instead of all possible connections [38]. Another relevant distinction between the two networks lies in the use of a similarity threshold or cut-off point (t). In the case of CSN, this threshold is a mandatory requirement, while it is an option in HSPN. HSPNs were selected for

use in this work due to their demonstrated superiority over CSNs in previous research, specifically in studies addressing peptides with other activities, such as antiparasitic and hemolytic properties [38, 39]. Accordingly, our research delineated the chemical space of AAPs by employing HSPNs and examined the effect of t on these networks. To achieve this, we compared networks constructed with an optimal value of t and those constructed without a value of t ($t = 0.00$), keeping the other parameters involved in this construction process constant.

The study focused on the pioneering construction and visualization of HSPNs designed specifically for AAPs. This novel approach incorporated six distinct (dis)similarity metrics: angular separation (As), Bhattacharyya (Bh), Chebyshev (Ch), Euclidean (Eu), Manhattan (Ma), and Soergel (So). These metrics allowed us to capture unique and complementary information, in contrast to previous studies in network science that were limited to using only Eu distance as the default similarity metric. In addition, metadata networks (MENTs) were established to explore and discover general patterns associated with AAPs, thus revealing biological characteristics and diverse relationships between peptides. Subsequently, a scaffold extraction process was carried out to generate simplified representations of the chemical space, thus allowing the optimal selection of HPSNs. Through this simplification, an alignment-free motif identification method discovered possible potential motifs of AAPs. SEA subsequently validated these motifs [40], using an external database as a reference.

1.1 Scope of Research

The present research constitutes a fundamental step towards a more precise analysis and interpretation of the information linked to the reported peptide sequences, with the purpose of providing researchers with a solid starting point for the future development of anti-angiogenic drugs as a therapeutic option against cancer. In the context of the search for alternative approaches to cancer treatment, antiangiogenic drugs based on antiangiogenic peptides and proteins have attracted increasing interest. Although databases exist that compile these sequences, conventional data analysis requires a considerable investment of resources and time. Moreover, it does not guarantee that researchers have specific and relevant information for antiangiogenic activity. Therefore, this study aims to simplify the initial stage of the process, bypassing manual scrutiny and experimental

validation. To this end, we propose approaching the analysis computationally, employing network science and visual mining. This was achieved through network analysis, scaffold extraction, and similarity search to represent the chemical space of antiangiogenic peptides through HSPNs and subsequent motif search and enrichment.

1.2 Objectives

1.2.1 Overall Objective

To deepen the understanding of the chemical space of antiangiogenic peptides (AAPs) through an innovative approach integrating computational methods such as interactive visual mining and network science to gain a more comprehensive and detailed perspective on the characteristics of AAPs.

1.2.2 Specific Objectives

- To use Metadata Networks (METNs) to analyze and describe general properties of antiangiogenic peptides.
- To create HSPNs, or Half-Space Proximal Networks, which depict the chemical space of antiangiogenic peptides.
- To analyze the impact of threshold values (t) and compare the use of various metrics to determine the optimal representation of the chemical space of AAPs.
- To offer scaffolds of hemolytic peptides that can accurately represent the entire chemical space without overrepresenting.
- Examine the physicochemical parameters characterizing HSPNs of the most optimal selected metrics.
- Identify and enrich new motifs that possess potential antiangiogenic activity.

Chapter 2

Background Information

2.1 Peptides in Cancer Treatments

Cancer is a well-known disease around the world, which is considered to be the second leading cause of death according to the World Health Organization, counting up to 9.6 million deaths in the year 2018 [1]. Cancer is when cells acquire some characteristics due to abnormal growth, which can be caused either by an inherited genetic mutation or by various environmental factors [2]. Among these characteristics, we can find the generation of own and response to weak signals not identified by non-cancer cells, non-response towards antiproliferative signals, resistance towards apoptotic signals, limitless replication capabilities, angiogenesis, metastasis, and tissue invasion.

The difference between cancer and normal cells can easily be observed, as shown in **Figure 2.1**, where cancer cells present a higher negative charge, a fluid outer membrane, and a greater surface area when compared to non-cancer cells [41]. It is inferred that when cancer progresses, the membrane fluidity of cells increases, which increases the micro-villi, conceiving a higher surface area in those cells.

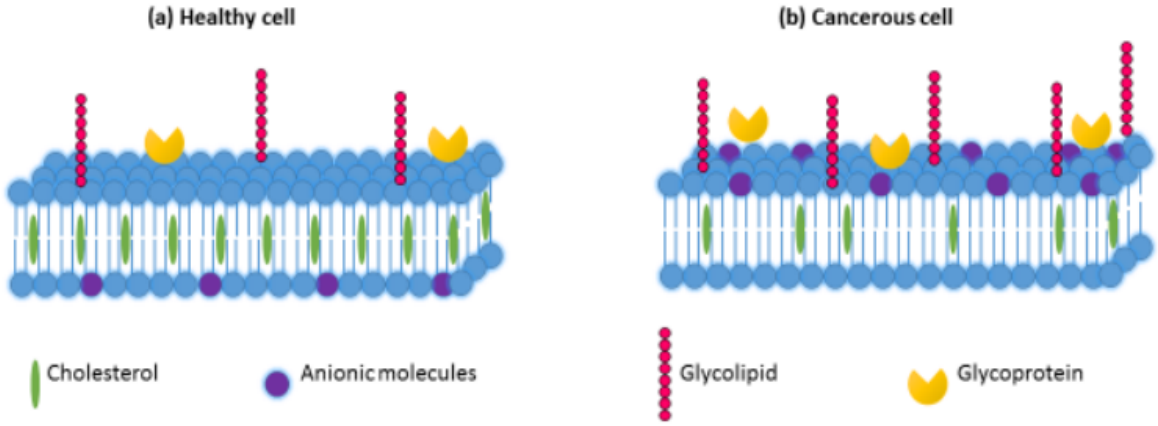

Figure 2.1: Differentiation between a) a healthy cell, and b) a cancer cell. Reprinted from [41].

The most commonly used cancer treatments are localized radiotherapy, surgery, or even both in some cases. However, in the scenario where metastasis has occurred or can-

cer is in an advanced stage, the preferred treatment is chemotherapy [42]. This approach is also used in non-local treatments to reduce tumor size. There are certain disadvantages to using chemotherapy as a main treatment. The main ones are that clinically used drugs do not have the capability of differentiating between cancer and healthy cells and that cancer cells can generate multi-drug resistance (MDR). This causes the known adverse side effects in cancer patients [43–45]. Concerning this matter, there is a need in the pharmaceutical industry to develop a new alternative to anticancer agents that may have a different mode of action in going against the high resistance of cancer cells without the added toxicity that it represents to healthy ones [6]. In this context, peptides could offer a promising alternative to conventional anticancer agents due to their advantageous characteristics, such as low toxicity and specificity against cancer cells. Also, these peptides could rapidly penetrate tumor tissues due to their small size. This allows the use of peptides in diagnosis, prognosis, and cancer treatment [46]. Recent evidence indicates that peptides with anticancer and anti-angiogenic properties are playing a significant role in cancer therapy. As a result, they are increasingly recognized as promising therapeutics for the future [7].

2.1.1 Antiangiogenic Peptides

Angiogenesis, the formation of new blood vessels from existing ones, is frequently observed in pathological conditions such as cancer. It provides essential oxygen and nutrients necessary for cancer growth and progression. The dysregulation between antiangiogenic and proangiogenic factors contributes to this phenomenon. This results in the progression of the disease. Some of the factors that are involved in this process are vascular endothelial growth factor (VEGF), platelet-derived growth factor (PDGF), fibroblast growth factor (FGF), and angiopoietins that interact with the cell-extracellular matrix. This is observed in **Figure 2.2** [47].

Many anti-angiogenic drugs are widely available; however, due to their high long-term toxicity and drug resistance, they can have some hazardous health effects [48]. Peptides stand out among anti-angiogenic proteins due to their unique benefits in cancer treatment. They are considered potential candidates for inhibiting angiogenesis, offering a less toxic therapeutic and effective option for diseases characterized by abnormal blood vessel formation, such as cancer [49]. Therefore, there has been an increased interest

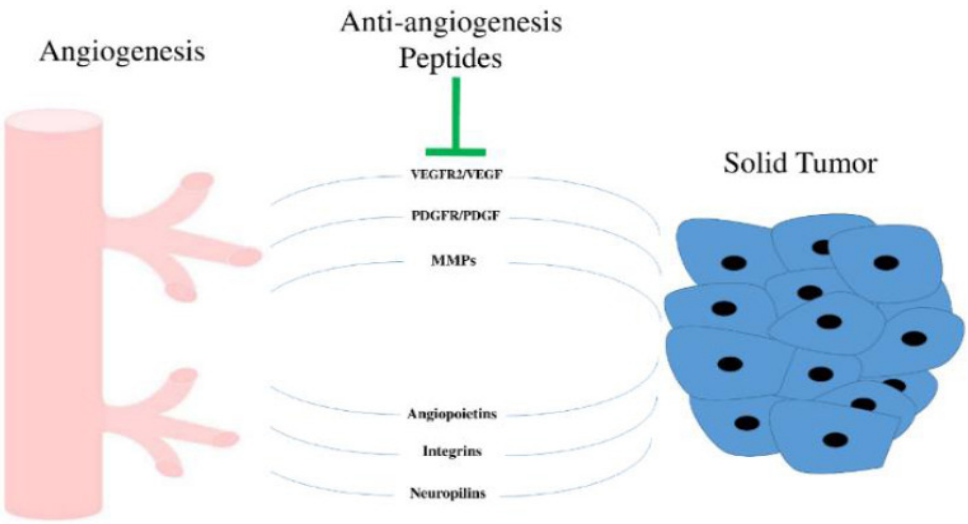

Figure 2.2: Enzymes and factors involved in the angiogenesis process. Reprinted from [41].

among scientists in the cancer treatment field to study and develop peptides that possess antiangiogenic properties to overcome the issues of their parent proteins. To do so, it is necessary to perform a pre-analysis on potential antiangiogenic peptide-based drugs, typically carried out *in silico* to avoid wasting resources, most of them using computational software [7, 23]. However, there are studies in the computational detection of antiangiogenic peptides [50]. This research will use computational tools and machine learning potentiated programs such as Startpep Toolbox to detect potential AAPs [50].

2.2 Computational Methods for Detection of AAPs

Computational methods have become a reliable alternative to experimental methods to save time and resources. Through this, studying and analyzing possible AAPs and AAPs-based drugs before their introduction in the global market is possible [25, 51]. Before getting into the computational methods used in this research, it is necessary to understand some concepts like machine learning and chemical similarity networks.

2.2.1 Machine Learning

Machine learning, also called ML, lies as a subgroup of artificial intelligence, and it is the main course of action applied for drug detection and discovery. It relies on algorithms and the creation of mathematical models to generate training data sets, enabling automated predictions of test sets [52, 53].

In this context, it's important to draw a comparison. Training data sets consist of both labeled and unlabeled data utilized as samples, whereas test sets comprise unknown data sets that require analysis. It is possible to determine the type of ML model by the training data's nature. Among these, supervised, semi-supervised, unsupervised, transfer, and reinforcement learning can be found [52, 53]. Supervised models correspond to labeled training data, as unsupervised models belong to unlabeled training models. Hence, finding patterns to predict the possible outcomes is necessary when using unsupervised machine learning models. Semi-supervised machine learning combines parts of both supervised and unsupervised models instead, and it contains labeled and unlabeled data sets, with the amount of unlabeled data being the bigger one [51]. In either case, the training data is used as feedback for reinforcement learning, considering that data is constantly changing and transferred to other domains. Supervised learning is the model commonly used in therapeutic peptide prediction as AAPs. [51, 54]

Another common characteristic of machine learning models is that they can classify or even regress the training data over the test sets; hence, the model performance will depend on the quantity and quality of the training data [51, 53]. to predict Therapeutic peptides models there is a common usage of classifiers of supervised learning like Random Forest, and Support Vector Machine techniques [55].

Random Forest uses classification or regression algorithms and is based on decision trees. Otherwise, Support Vector Machine classifies unlabeled data and performs a binary classification using a linear hyperplane, maximizing the separation between classes [56, 57]. Support Vector Machine utilizes a core function to create a feature space that facilitates linear separation, incorporating radial, polynomial, or Gaussian functions, particularly in cases where the space is non-linear. [57].

2.2.2 Chemical Similarity Networks

Chemical space is the term used to represent all synthetically and natural molecules. Nonetheless, considering the vastness of molecules within the chemical realm, selective portions of this domain focus on specific activities of interest. This targeted approach involves delving into the biologically significant chemical sphere, which predominantly encompasses compounds pivotal to biological systems [58, 59].

This chemical space is conceptualized as a multi-dimensional framework. Within this

framework, numerical features or computational vectors define molecules and encapsulate their physicochemical attributes, known as molecular physicochemical descriptors [59]. Here, each molecular descriptor is used to represent one dimension. Coordinate-based maps require a dimension reduction technique for visualizing 2D and 3D maps [34, 60]. To do so, there are coordinate-free chemical spaces, which is the case for Half Space Proximal Networks (HSPN). These types of chemical spaces are commonly used to visualize the chemical space with lower complexity [31, 61].

2.2.3 Half Space Proximal Network (HSPN)

A known disadvantage of typical chemical networks is the requirement of high RAM capacity in the device being used [34]. Then, a Half-Space Proximal Network (HSPN) is an alternative type of network that requires low RAM since it builds a lower-density network [62]. It works using the Half-Space Proximal test, which is explained starting from an initial point u. First, the initial point u and its nearest neighbor v are connected by an edge, followed by the addition of an imaginary orthogonal line in the middle of the edge which divides the space into two half-planes. The half-plane farthest to u is called the forbidden area. Then, the initial point u is connected to the nearest point of the non-forbidden area, and the process is repeated until the set of points of the non-forbidden area is empty. All points are somehow connected to the core node in the same region as shown in **Figure 2.3** [62].

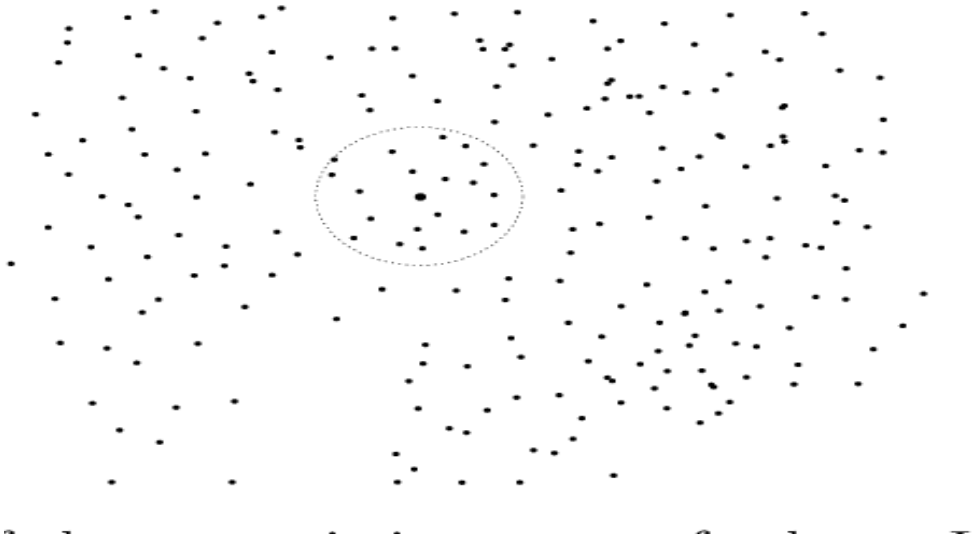

Figure 2.3: Example of connection using HSPN, where all nodes within the range would be connected to the central node. Reprinted from [62].

2.2.4 Network characterization parameters.

The parameters needed to characterize these networks are similarity threshold, density, node degree, centrality, clustering, and metrics [62].

Similarity Threshold

The similarity threshold is important in network science and defines its topology and appearance. Also, it establishes the lower limit value of similarity between node pairs connected by an edge [37, 63]. It can be understood that they are connected if two or more nodes have an equal or greater similarity value than the previously established one.

Node Degree

Node degree, also called vertex degree, is the number of edges bonded to a node, and it is used to represent the number of nodes to which each edge is attached [64].

Density

Network density refers to the proportion of edges within the network compared to the total potential edges. It typically relies on the similarity threshold value and plays a crucial role in shaping the network's properties. It provides insights into the level of connectivity and intensity of interactions among the elements within the network [62]. Furthermore, network density is defined by the following equation

$$\rho = \frac{2m_t}{n(n-1)} \tag{2.2.1}$$

Here, m_t is the number of edges, t is the threshold value, and n is the number of nodes in the network[34].

Clustering

Clustering has great importance in unsupervised learning models. It bases its function on dividing the graph data into different communities in accordance with the similarity between each node [34, 37]. Accordingly, similar nodes residing in the same community

and nodes from distinct communities differ. In a network, the modularity typically measures how well the nodes are classified into communities [65]. This value can be positive or negative with a maximum value of one and is defined by the following [66].

$$Q = \frac{1}{2m_t} \sum_{uv} (a_{uv} - \frac{k_u k_v}{2m_t}) \delta(c_u, c_v) \tag{2.2.2}$$

Here, a_{uv} represents the weight of the edge which is similarity value between a node u and a node v, the sum of the weight of edges joined to node u is represented by k_u, c_u shows the community of u, and finally, $\delta(c_u, c_v)$ is defined as the following.

$$\delta(c_u, c_v) = \begin{cases} 1 & \text{if} \quad c_u = c_v \\ 0 & \text{if} \quad c_u \neq c_v \end{cases} \tag{2.2.3}$$

Modularity describes the number of edges connecting nodes intra-community minus the expected number of edges randomly determined in an identical network [66]. It increases as density decreases; hence, community structures are better resolved, which is why modularity is a parameter that must be optimized as much as possible to acquire the best partition of the network.

Louvain Clustering is an algorithm that has been demonstrated to have the best accuracy and computing time, and it is widely applied for modularity optimization. This algorithm starts assigning all the available nodes in different communities and typically consists of two phases [67]. In the first phase (I), one node is moved to the community of its nearest neighbor, and its new modularity parameter is computed. When modularity increases due to movement, the node relocates to the respective community; otherwise, it stays within its initial community. This iterative process continues for all nodes until no further enhancement in modularity is achieved. In the second phase (II), a new network is built based on the previously obtained communities in the (I) phase. Ultimately, the process is repeated until there are no changes in the modularity parameter [67].

Additionally, it is possible to measure the network's ability to cluster together using the "average clustering coefficient (ACC)." This coefficient represents the network's capability to connect to nodes sharing the same neighbor, and it is a measurement of the neighborhood connectivity [37].

Centrality

Centrality is one of the essential measurements in network science, and its nodes rank according to how representative they are in their network. There are different methods to calculate centrality, and the most used are the harmonic centrality, community hub-bridge centrality, weighted degree centrality, and betweenness [62, 68].

Harmonic centrality is considered as a global centrality measurement, and it is based on the distance between two nodes. The harmonic centrality for node u is defined as

$$C_H(u) = \sum_{v \neq u} \frac{1}{d(u,v)} \tag{2.2.4}$$

here $d(u, v)$ represents the distance from node u to node v [69].

This centrality is considered in network analysis because of its sensitivity to network structure, its ability to capture the relative importance of nodes and its effectiveness in identifying important nodes in biological networks.

Community hub-bridge centrality is a local centrality measurement based on where the node is located in the community. Also, nodes can be considered hubs or bridges in this type of centrality. Local hub nodes connect various internal nodes, while bridge nodes are located at the boundary of a community and act as the attachment between two neighboring communities [34, 70]. The hub-bridge centrality for node u is defined as

$$C_{HB}(u) = k_u^{in} * CS(u) + k_u^{ex} * NC(u) \tag{2.2.5}$$

where k_u^{in}, and k_u^{ex} are internal and external strength, respectively, defined as in Weighted degree. $CS(u)$ is the community size of u, and $NC(u)$ is the number of neighboring communities directly attached to u by other nodes from its community.

Weighted degree centrality establishes the similarity between a node pair, and it is known as the weight of the edge. The internal and external strength gives it by.

$$k_u^{in} = \sum_{v \in c_u} a_{uv} k_u^{ex} = \sum_{v \notin c_u} a_{uv} \tag{2.2.6}$$

Here k_u^{in} denotes the internal strength, k_u^{ex} represents the external strength, and a_{uv} corresponds the similarity value between u and v [34].

Betweenness centrality This centrality primarily involves assessing the prevalence of short path lengths. Specifically, the betweenness centrality of node u quantifies the quantity of shortest paths between pairs of nodes, excluding node u itself, that traverse through it. This centrality provides insight into the extent to which a node serves as a bridge or intermediary within the network, facilitating efficient communication between disparate nodes.

$$C_B(u) = \frac{1}{(N-1)(N-2)} \sum_{x \neq u, x \neq v, v \neq u} \frac{SP_{xv}(u)}{SP_{xv}} \tag{2.2.7}$$

Here, N is the number of total nodes, $(N-1)(N-2)$ is the number of node pairs excluding node u,$SP_{xv}(u)$ represents the number of shortest paths between the nodes x and v that cross node u, finally SP_{xv} denotes the total number of shortest paths between nodes x and v [71].

2.2.5 Metrics

The model used to analyze similarity networks employs eight distinct metrics using an optimal cut-off point selected for each metric [62]. The metrics are Manhattan, Euclidean, Soergel, Chebyshev, Bhattacharyya, and Angular Separation:

Manhattan (*Ma*) or taxicab is a geometry where the Euclidean geometry is replaced by a new one where the distance between two points is the sum of the absolute difference of the corresponding Cartesian coordinates [72].

Euclidean metric (*Eu*) represents the length between two points in a line segment; this happens in the Euclidean space. It is calculated using Cartesian coordinates of the points and mainly uses the Pythagorean theorem. In this metric, the distance between two objects not considered points is the smallest distance among a pair of points from two objects [72].

Soergel metric (*So*) is a distance metric used to measure the similarity between two binary points (vectors). It is defined as the number of positions in which the two vectors differ. Is considered to be simple and efficient to compute and has proved to be effective in various applications, including bio-computation [73]

Chebyshev metric (*Ch*) is defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension. It goes as $d(x, y) = max|x_i - y_i|$ where x and y are in the vector space, while x_i and y_i are elements

of x and y, respectively [74].

Bhattacharyya metric (*Bh*) is considered a similarity measure between two probability distributions. It is defined as $Bh(P,Q) = \sqrt{\sum_x P(x)Q(x)}$, where all the possible values of a random variable X are taken over, and P and Q are the two probability distributions [75].

Angular Separation metric (*As*) is a measure of the distance between two points in angular space, and it is often used in computational approaches to measure the distance between objects in an image or network. It is the angle between two sight-lines, or between two point objects viewed from a node (observer), measured in degrees or radians [76].

2.3 Physicochemical descriptors

The molecular configuration of a compound is intrinsically linked to its chemical behavior, the prediction of which is achieved through the abstraction of its structure in terms of chemical similarity parameters, known as "descriptors"[77]. These descriptors are valuable for anticipating the pharmacological properties of drug candidates and for predicting reactivity, toxicity, and other essential chemical aspects. Molecular descriptor research seeks to establish quantitative relationships between structure and activity to categorize similar peptides in diverse contexts, avoiding synthesizing and evaluating overly similar compounds that would waste resources [78].

Maximizing or minimizing the structural diversity of peptides is essential for searching or refining essential compounds. The utility of a molecular descriptor in peptide library design is that subtle variations in this descriptor between two molecules should reflect equally subtle biological differences. Molecular descriptors are numerical representations derived from molecular features that allow a mathematical approach to molecules, which is an essential step in converting molecular features into quantifiable data [79]. These descriptors, defined by specific algorithms or experimental protocols, encapsulate various aspects of a molecule's chemical information.

It is important to clarify that none of these descriptors are directly based on the three-dimensional structure; instead, they are exclusively based on the sequence of the peptides.

- Peptide Sequence Length: This descriptor indicates the number of amino acids that comprise a peptide. The sequence length can vary significantly among different

peptides and influence their three-dimensional structure, biological function and stability [80].

- Net Charge : The net charge of a peptide refers to the difference between the total number of positive and negative charges in its structure. This property is crucial for understanding the interaction of the peptide with other molecules, such as cell membranes, and can influence its ability to bind to specific receptors[81].
- Isoelectric Point: The pH at which a peptide has a net charge equal to zero. At this point, the peptide exists in its most neutral form and can have a higher solubility. The isoelectric point is important in determining the optimal conditions for peptide separation and purification by techniques such as electrophoresis [81].
- Molecular Weight: The total mass of a peptide, measured in atomic mass units. Molecular weight is a fundamental property that affects many peptide characteristics, such as its behavior in solution, its ability to cross biological membranes and its interaction with other molecules [80].
- Boman Index: The Boman index shows the degree of discrimination between membrane-interacting peptides and protein-interacting peptides. The Boman index is defined as the sum of the free energies of the side chains for transfer from cyclohexane to water and divided by the total number of residues in the peptide. A more hydrophobic peptide tends to have a negative ratio, while a more hydrophilic peptide tends to have a more positive ratio [82].
- Hydrophobicity : This descriptor indicates its affinity for hydrophobic or hydrophilic environments. Higher hydrophobicity may influence the ability of the peptide to interact with biological membranes and other proteins. This parameter is important for peptide stabilization, but the solvent and solubility in which the sequence is found must be considered [83] .
- Aliphatic Index: The aliphatic index provides information on the proportion of aliphatic amino acids in the peptide sequence. Aliphatic amino acids contribute to the thermal and structural stability of the peptide and can influence its ability to fold correctly. The aliphatic index is the relative volume occupied by a protein's aliphatic amino acid side chains. Amino acids belonging to this group are alanine (A), valine (V), isoleucine (I), and leucine (L) [77].
- Instability Index: This index indicates the structural stability of a peptide. Peptides

with a high instability index may be prone to denaturation or degradation under specific conditions, which may affect their functionality and half-life in the body [84].

- AvgGRAVY: This descriptor quantifies the relative hydrophobicity of a peptide by calculating the average value of the hydrophobicity of its amino acids. A negative AvgGRAVY value indicates that the peptide is more hydrophilic, while a positive value indicates higher hydrophobicity. This property may influence the solubility and interaction of the peptide with other molecules in aqueous environments [85].

Chapter 3

Experimental Procedure

The following steps are constituents of the whole workflow: (i) visual mining: metadata network generation (METNs), (ii) HSPNs production and analysis: Chemical space representation of AAPs obtained from StarPepDB, (iii) scaffold extraction and exploration: construction of representative subsets using the best HSPN candidates, (iv) motif finding and enrichment: making a comparison with data reported in the literature. **Figure 3.1** provides a comprehensive summary of the entire methodology, offering an overview of the research.

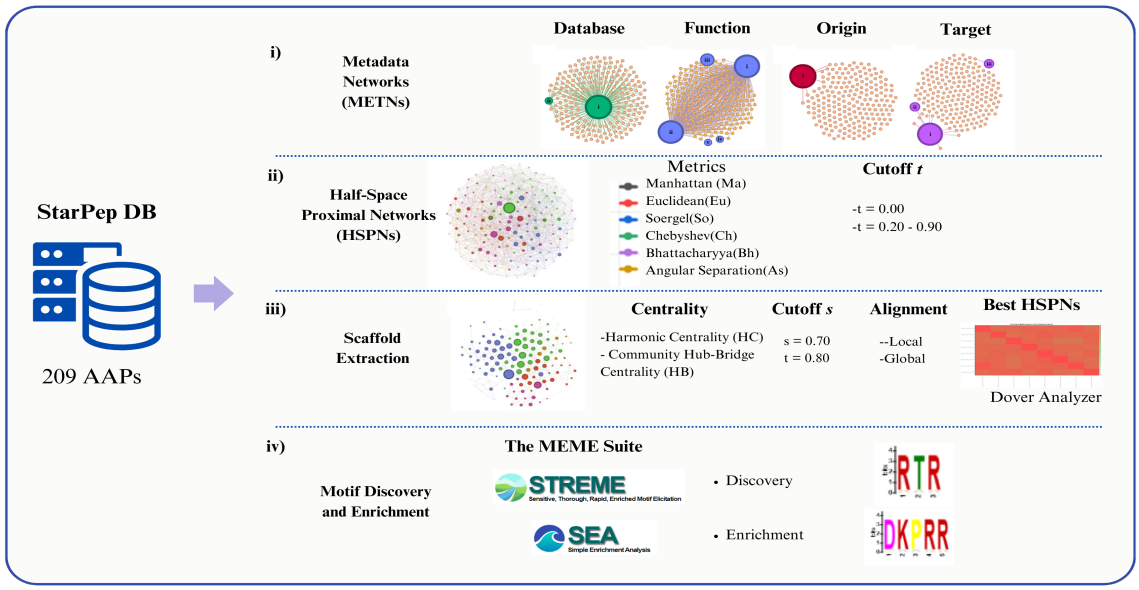

Figure 3.1: Overview of the experimental section: (i) Visual mining: generating Metadata Networks (METNs) comprising Database, Function, Origin, and Target, (ii) production and analysis of HSPNs: depicting the chemical space of AAPs derived from StarPep DB, (iii) scaffold extraction and exploration: creating representative subsets using top HSPN candidates, (iv) motif search and enrichment: conducting a comparison with data reported in the literature.

3.1 Databases, Analysis Tools

3.1.1 StarPep DB and StarPep Toolbox

The StarPep DB graph database, implemented in Java, stores information on 45,120 bioactive peptides, including their functions and metadata, from multiple databases and other resources [26]. For our research, we selected a subset of data consisting of 209 AAPs. These data were used for the construction of HSPNs and METNs and the discovery of new anti-angiogenic motifs. On the other hand, we employed the StarPep toolbox [36], a software tool that facilitates a better understanding of the integrated data. This tool played a pivotal role in facilitating the construction and visualization of networks, enabling interactive exploration, and supporting the exportation of AAPs.

3.1.2 AntiAngioPred DB

The AntiAngioPred database [25] played a key role in motif enrichment. The authors organized this dataset as follows: (i) positive data set, initially composed of 257 AAPs obtained from various research articles and patents, reduced to 135 AAPs after excluding those sharing more than 70% sequence identity; (ii) negative data set, consisting of 135 Non-Antiangiogenic Peptides (NAAPs), consisting of random protein sequences extracted from the Swiss-Prot database [86]; (iii) independent dataset, consisting of 28 AAPs and 28 NAAPs, obtained by extracting 20% of the positive and negative datasets respectively; and finally, (iv) random dataset, comprising a total of 675 NAAPs distributed in subsets named 'Random1', 'Random2', 'Random3', 'Random4', and 'Random5'. These subsets were created using the same procedure to develop the negative data set.

3.1.3 BIG_ANTIAN_DB

To facilitate the management of these sets, we created a schema (**Figure 3.2**) to better organize the collected AAPs and NAAPs. On the left side of the schematic, there is exclusively positive data. Of the 257 AAPs reported, we identified 202 unique sequences that we named MAIN_DB_POSITIVE. Subsequently, we divided this set into two groups according to the criterion proposed by the dataset authors (70% sequence identity): sequences showing higher redundancy were grouped into the subset MAIN30%_DB_POSITIVE (67

AAPs), whereas those showing higher diversity were grouped into MAIN70%_DB_POSITIVE (135 AAPs). From the latter subset, we created two additional subsets. The first subset, called IND_DB, is an independent set containing 28 AAPs, and the second subset, called TR_DB, is a training set comprising 107 AAPs, both representing 20% and 80%, respectively, of the MAIN70%_DB_POSITIVE set.

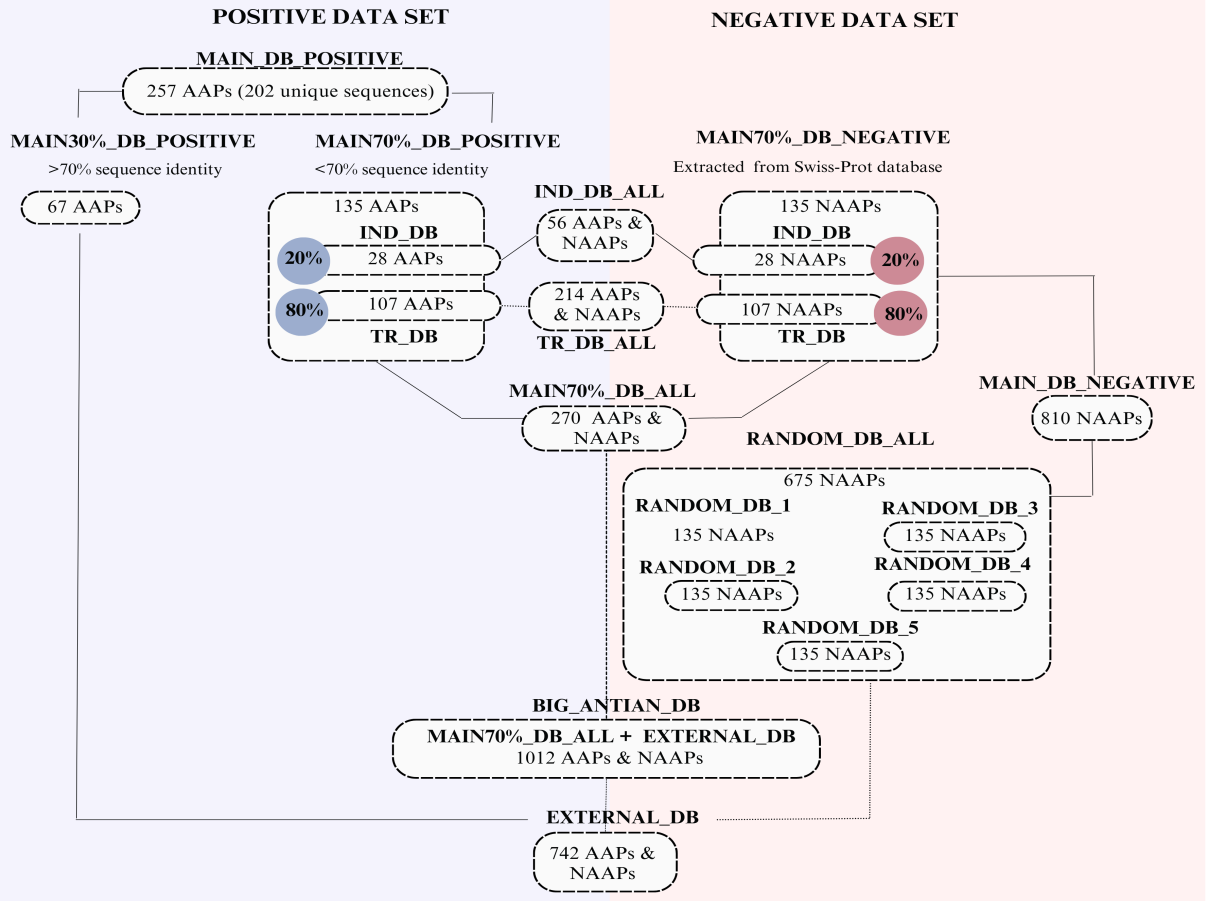

Figure 3.2: Databases generated from the AntiAngioPred database [25]. AAPs=Antiangiogenic Peptides, NAAPs = Non Antiangiogenic Peptides.

The negative data is on the right side of the schematic, with subsets analogous to those of the positive data. The difference lies in the addition of all NAAPs from the five random subsets, forming the set RANDOM_DB_ALL (675 NAAPs). Combining RANDOM_DB_ALL with MAIN70%_DB_NEGATIVE (135 NAAPs) created the set MAIN_DB_NEGATIVE (810 NAAPs). In addition, subgroups that share AAPs and NAAPs were created. The first, called IND_DB_ALL, contains 56 AAPs and NAAPs and comprises the peptides from the independent data of the positive and negative data. The second, TR_DB_ALL, contains 214 AAPs and NAAPs and represents the training data from the positive and negative sets. Combining all the data from MAIN70%_DB_POSITIVE

and MAIN70%_DB_NEGATIVE yielded MAIN70%_ALL, which comprises 270 AAPs and NAAPs. Next, we pooled the data from MAIN30%_DB_POSITIVE with RANDOM_DB_ALL to form an external dataset named EXTERNAL_DB (742 AAPs and NAAPs). Finally, the set encompassing all positive and negative data was assigned the name BIG_ANTIAN_DB (1012 AAPs and NAAPs).

3.1.4 Gephi 0.10.1

For generating different types of networks, Gephi 0.10.1 was considered one of the most efficient alternatives in terms of visualization and exploration [87]. This software, characterized by its free and open-source nature, has been used for the retrospective calculation of significant measures from the networks. These metrics encompass a range of quantitative measures, including average degree, diameter, radius, density, modularity, clustering coefficient, average clustering coefficient (ACC), average path length, and counts for edges and nodes. This tool is used for calculating measurements and visualizing networks that have been previously generated using the StartPep Tool Box [36].

3.1.5 The MEME Suite

The MEME Suite, a comprehensive set of software tools for motif-based sequence analysis, offers versatile applications across peptides, proteins, DNA, and RNA [88]. With access to extensive proteomic and genomic databases, this suite includes 13 tools, such as Sensitive, Thorough, Rapid, Enriched Motif Elicitation (STREME) [89] for motif discovery and Simple Enrichment Analysis (SEA) for enrichment [40]. The web version (https://meme-suite.org/meme/doc/streme.html) facilitates de novo motif identification, improving the accuracy and versatility of discovery, even in large datasets. STREME facilitated both motif discovery and the generation of statistical estimates of the relevance of each anti-angiogenic motif identified. On the other hand, SEA identifies known motifs in input sequence sets and performs differential enrichment analysis when additional control sequences are present [40]. Motif enrichment analysis evaluates the significant presence of known motifs in DNA, RNA and protein sequences [88]. Both tools provide accurate statistical estimates in discovering and enriching motifs linked to antiangiogenic activity. They also can handle different alphabets, such as DNA, RNA and proteins, allowing a customized definition according to specific needs. The sequences of each cluster obtained

from the best-selected HSPNs were used for the discovery process. Both newly discovered and previously identified motifs were used for enrichment, together with sequences from the AntiAngioPred and BIG_ANTIAN_DB datasets. This was done to satisfy SEA requirements, including control and input sequences.

3.2 Network Generation and Analysis

3.2.1 Metadata Network (METNs)

MENTs provide a clear and detailed view of the interactions within complex systems by identifying meaningful groups within data sets, thus providing a simplified understanding of how the components of that system interact [90]. This parameter could be derived from database source, function, origin, or target attributes. Consequently, four distinct METNs were established, each grounded in various metadata attributes. This analysis reveals hierarchical structures in the data, especially the connection between nodes representing "peptides" and their corresponding "metadata" nodes. These connections illustrate specific relationships, such as the association of a peptide with its data base source. In addition, different peptides may be linked to the same metadata. These networks are depicted as unweighted pseudo-bipartite graphs, denoted as $F = (V, E)$, wherein $V(F)$ represents the set of nodes or vertices, encompassing two distinct classes: AAPs and metadata information. On the other hand, the set of edges, denoted as $E(F)$, defines the connections within the graph [91]. In this instance, the peptide classification within the set V(F) encompassed a compilation of 209 AAPs procured from the StarPep toolbox [36].

It is imperative to acknowledge that METNs do not conform to the classification of fully bipartite graphs. This deviation arises from the fundamental constraint within bipartite graphs, wherein nodes originating from the same category are precluded from establishing adjacencies [30, 32]. METNs, however, can establish connections within the metadata class, provided that a hierarchical relationship between nodes is established.

Through the construction of METNs, valuable insights were garnered to formulate a comprehensive overview of databases that exhibit pronounced concentrations of AAPs. This methodology facilitated an examination of the redundancy prevalent among these peptides across various databases. Furthermore, it enabled the identification of peptides

exclusively documented within distinct databases. The function of interest within the database, namely the category of AAPs, was pre-selected. Subsequently, one of the four specific parameters we intend to work with is chosen to set up the network. The graphical representation clearly distinguishes two components: the peptides in question and the focus parameter. These elements are differentiated by assigning distinctive colors to each one. The METNs were built using StarPep toolbox [36] and further improve the visualization of the network, the Gephi 0.10.1 platform was used [87]. In this context, an exhaustive customization is applied to the network to highlight those parameters of primary interest.

METNs Visualization

To customize our network, we utilized the StarPep DB, StarPep toolbox along with the Gephi 0.10.1 software [26, 36, 87]. In the context of the four generated metadata networks, we maintained a consistent color scheme for the peptides while assigning distinct colors to each metadata category. Specifically, orange was designated for peptides, while shades of green, blue, red, and purple were employed for the Database, Function, Origin, and Target metadata categories, respectively.

Within the StarPep framework, we computed the Betweenness centrality, a metric that aided us in adjusting the sizes of nodes to reflect the significance of the metadata [92]. Concurrently, we employed the modularity optimization technique to compute clusters, contributing to identifying densely interconnected groups within the network. For improved visual representation, we employed two different layout methodologies: the first being Force Atlas 1, which optimized the arrangement of nodes based on attractive and repulsive forces, leading to a clear spatial organization[93]; the second was Noverlap, a technique used to address the issue of node overlap, particularly for peptides [94]. This approach to network visualization ensured a comprehensive and aesthetically refined depiction of the intricate relationships encompassed within the data.

3.2.2 Half Space Proximal Networks (HSPNs)

HSPNs, or Half-Space Proximal Networks, serve as networks indicating similarity or correlation, connecting nodes if their similarity coefficient meets or exceeds a predetermined t. In a manner akin to metadata networks, HSPN construction follows a structured for-

mat denoted as $G = (V, E)$. Here, $V(G)$ represents the assembly of individual AAPs, each functioning as a distinct node. These nodes are defined by vectors comprising sequence-based molecular descriptors (MODs) values. Edges, on the other hand, establish connections between these nodes. $E(G)$ signifies the compilation of these edges, serving as links between nodes based on two explicit criteria.

Firstly, a similarity matrix M of size nn is crafted, where $M_{i,j}$ signifies the similarity score between nodes V_i (G) and V_j (G). Similarity scores range from 1 denoting utmost similarity to 0 indicating the lowest. These values originate from peptide feature vectors utilized to calculate dissimilarity measures for node pairs, which are subsequently normalized through min-max normalization. Following this, by applying the Half-Space Proximal (HSP) test [95], an intricately connected yet sparse network HSPN emerges. Despite keeping the number of nodes constant, the network has fewer links, which significantly improves the system's efficiency, thus considering it an advantage over CSN [34]. The resulting HSPN displays robust interconnections while maintaining a sparse structure.

Secondly, another criterion comes into play: further reduction in graph density can be achieved by selectively removing edges with similarity values falling below a predetermined t value. As a result, an enhanced depiction of the chemical space occupied by AAPs is obtained by assessing the structural features of the graph. It's imperative to note that including a t value during HSPN creation remains an optional choice, emphasizing the adaptability of this approach.

The construction of HSPN commenced with extracting pertinent data from the StarPep DB [26]. Specifically, we acquired a set of 209 peptides demonstrating antiangiogenic activity from this database. Subsequently, measures were employed to mitigate the redundancy within the peptide sequences. This was achieved by employing the Smith-Waterman local alignment algorithm in combination with the Blossum 62 substitution matrix developed by Henikoff & Henikoff [96]. Moreover, a rigorous standard of 98% sequence identity was enforced to guarantee a successful reduction in sequence redundancy.

For feature selection, an unsupervised technique was employed to calculate the peptide sequence's Mutual Dependence. To further eliminate extraneous attributes associated with angiogenic peptides, the Shannon Entropy was employed as a relevance criterion [97], with a predetermined threshold set at 10%. Conversely, in the pursuit of excluding

duplicative features, the Spearman correlation coefficient was adopted as a metric [98]. In this instance, a threshold of 0.8% was established. Subsequently, all retained peptide features were chosen to facilitate the generation of networks. To provide a more in-depth understanding of the selection criteria for constructing HSPN, one can refer to the comprehensive analysis conducted by Aguilera and colleagues [34]. This work thoroughly elucidates the detailed characteristics considered in the construction of HSPN.

Selecting an appropriate (dis)similarity metric became imperative to enhance the modeling and visualization of the chemical space mapping. In this study, six distinct metrics were employed: *As, Bh, Ch, Eu, Ma, and So* (**Table 3.1**). This array of metrics is included from prior research, underscoring the significance of using diverse measures. This approach was founded on the idea that various metrics can effectively capture unique information, recognizing that distance metric does not necessarily coincide uniformly with the commonly used Euclidean metric. [53].

A systematic variation of the t value was conduct to assess the performance characteristics of the HSPNs. Specifically, 16 discrete threshold points were investigated for each metric. This range spanned from the initial point of 0.00, followed by increments of 0.05, progressing through the values of 0.20 to 0.90. This selection of thresholds was deliberate, driven by extracting comprehensive insights into the overarching parameters governing the behavior of the HSPNs. Notably, the chosen threshold values were motivated by previous investigations, wherein it was observed that the parameter fluctuations remained relatively subdued within the spectrum spanning from 0.00 to 0.45 [38, 39]. This discernment informed the upper limit of the chosen threshold range. This rigorous threshold variation scheme yielded a collection of 128 distinct HSPNs, fostering a comprehensive evaluation of the network's behavior.

Table 3.1: (Dis)Similarity Metrics used to Build HSPNs.

Measure	Formula	Range1	Average	Range
Angular Separation (As)	$d_{xy} = 1 - Cos_{xy}$ where, $Cos_{xy} = \frac{XY}{\|\|X\|\|\|\|Y\|\|} = \frac{\sum_{j=1}^{h} x_i y_j}{\sqrt{\sum_{j=1}^{h} x_j^2 \sum_{j=1}^{h} y_j^2}}$	[0,2]		
Bhattacharyya (Bh)	$d_{xy} = \sqrt{\sum_{j=1}^{h} (\sqrt{x_j} - \sqrt{y_j})^2}$	[0,∞)	$d = \frac{d_{xy}}{\sqrt{n}}$	[0,∞)
Chebyshev (Ch)	$d_{xy} = \max\left\{\|x_j - y_j\|\right\}$	[0,∞)	$d = \frac{d_{xy}}{n^{1/p}}$	[0,∞)
Euclidean (Eu)	$d_{xy} = \left(\sum_{j=1}^{h} \|x_j - y_j\|^2\right)^{\frac{1}{2}}$	[0,∞)	$d = \frac{d_{xy}}{n^{1/p}}$	[0,∞)
Manhattan (Ma)	$d_{xy} = \frac{1}{n} \sum_{j=1}^{h} \|x_j - y_j\|$	[0,∞)		
Soergel (So)	$d_{xy} = \frac{1}{n} \sum_{j=1}^{h} \frac{\|x_j - y_j\|}{\max\{x_j, y_j\}}$	[0,1]	$d = \frac{d_{xy}}{n}$	$[0, \frac{1}{n}]$

(a) The variables x_i and y_j represent the values of peptide descriptors j for peptides m and n respectively. Peptides m and n are denoted by feature vectors X and Y. The value h represents the number of peptide features. "Range" refers to the span and not the order, defined as Range = $(\max x_i - \min y_j)$

HSPNs Visualization

The procedural framework involved the application of the Louvain method [67] in conjunction with the utilization of a specific centrality measure, namely the Community Hub-Bridge (HB) centrality [99]. This centrality measure was systematically applied to each node encompassed within the StarPep DB[26]. Subsequently, these nodes were designated distinct colors corresponding to the specific clusters they were affiliated with. Furthermore, in pursuit of enhanced visualization, a proportional scaling approach was employed to configure the node sizes in direct correlation with their respective HB centralities. Incorporating a Bezier interpolator further refined this scaling process [100]. As a culminating step, a Fruchterman Reingold Layout was executed, resulting in an optimized spatial arrangement of the nodes [101].

The resultant visual representation was then extracted in a GraphML file format, facilitating seamless integration into the Gephi 0.10.1 software environment for advanced analysis and visualization [87]. Within this software, the visualization of the comprehensive network involved retrieving specific parameters in the past tense. Included within these global network parameters are various elements, including the number of edges, modularity, density, average clustering coefficient (ACC), number and size of clusters or communities, singletons GC (nodes disconnected from the giant component), single-

tons D0 (nodes of degree zero), diameter, average path length, mean degree, and the probability distribution of degrees (expressed as the probability of k). This procedural methodology was executed across eight distinct metrics, incorporating diverse cut-off points. Subsequently, the optimal cut-off point was meticulously selected for each metric, thereby enabling the subsequent execution of scaffold extraction based on these carefully determined cut-off values.

3.3 HSPNs Scaffold Extraction

For this processing, we took advantage of a data mining tool available through the StarPep toolbox [36], which facilitates the mining of sub-networks by extracting scaffolds. The scaffold extraction process was implemented with the objective of achieving a minimum representation within the chemical space of the AAPs. In this context, optimal cutoff points (t) and without cutoff points ($t = 0.00$) were evaluated for each of the six metrics analyzed. The two variations of t for each metric were considered, totaling 12 configurations, which served as the main basis for scaffold construction.

Other parameters involved in this process included the centrality measure, which varied between the calculation of Harmonic Centrality (HC) [102] and Community Hub-Bridge Centrality (HB) [99] for each node. Since the underlying objective at this stage was to identify the most fundamental and distinctive AAPs, redundancy elimination was carried out. This was guided by a criterion in which peptide sequences that showed a percentage of identity exceeding a particular cut-off value (r) were excluded. In this study, the particular r values ranged between 70% and 80%. It is important to note that the cutoff values, referred to as "t" and "r", are different; "t" was used in the construction of HSPNs, while "r" was used in the construction of scaffolds.

Furthermore, the alignment approach underwent variation, encompassing global alignment using the Needleman-Wunsch method (G)[103] and local alignment via the Smith-Waterman technique (L)[104]. In summary, four fundamental parameters were modified with two variations each for each of the six metrics: the t-value, the centrality metric, the percent identity (r) threshold, and the sequence alignment mode. All computational iterations were performed using the Blosum 62 substitution matrix of Henikoff & Henikoff [96]. Overall, a total of 96 scaffold extraction experiments were meticulously carried out.

A specific notation was used to store this information. This notation goes as follows:

the type of network being constructed is placed, followed by the cut point t, the notation of the metric used, the type of alignment, G for global and L for local, followed by the cut point r and finally the number of nodes left after performing the whole procedure. For example, a name was: HSPN_00_As_HB_G_0.7_149.

Subsequently, the Dover Analyzer [105] was employed, an application specifically designed to facilitate the analysis of peptide sequences' overlap, diversity, and redundancy. A thorough examination was conducted of all the scaffolds that were obtained. An input set was introduced, which in this case represented a set of scaffolds to be analyzed to carry out the analysis with this tool. This grouping brings together scaffolds that share the same centrality measure, the same alignment type and the same r value. The variation among the scaffolds in this grouping lay in the different metrics (*As, Bh, Ch and Eu*), with their respective optimal values of t and without cut-off ($t = 0.00$). The examination was based on three parameters: identical overleap, similarity overleap with a threshold of 80%, and diversity ratio. Analyzing the most representative data sets for each metric makes it feasible to determine the optimal metrics with their respective t-value for the subsequent studies proposed in this work.

3.4 Physicochemical Descriptors

Molecular physicochemical descriptor calculations were performed for each cluster using the StarPep toolbox [36]. For this purpose, the previously generated fasta documents for each cluster were loaded and the StarPep molecular feature tool was used. Properties relative to each cluster were calculated, including peptide sequence length, net charge, isoelectric point, molecular weight, Boman index, average hydrophobicity, aliphatic index, instability index, and AvgGRAVY. The procedure involved calculating each property for each AAP and then determining the property's average for each group. This process was carried out for the As metric with 6 groups, the Eu metric with six groups and the Ch metric with eight groups, using the cut-off point chosen from previous analyses.

Subsequently, the analysis of each descriptor was carried out in general for the metrics studied. The average of the values of the descriptors of each cluster belonging to each metric was recalculated in order to evaluate the trend of the descriptor values with respect to the metrics.

3.5 Motif Discovery and Enrichment

Motif Discovery

STREME [89], a tool from the MEME 5.5.2 suite [88], was used to discover motifs using the alignment-free method. This algorithm is notable for its versatility in facilitating pattern discovery in large datasets containing hundreds of thousands of sequences. It has the unique ability to recognize both short and long motifs, whose length can vary from 3 to 30 positions. In addition, it can perform differential analysis by comparing patterns in different sequence data sets. In this case, the generated patterns had a minimum length of 3 positions and a maximum length of 6.

Motif Enrichment

Motif enrichment was performed using MEME Suite 5.5.2 [88] as well; however, on this occasion, we chose to use the Sequence Enrichment Analysis (SEA) tool [40]. This analysis aimed to investigate the existence and relevance of motifs in the sequences of each metric cluster studied. After eliminating the redundancy of the discovered anti-angiogenic motifs, these motifs were grouped with those previously reported in the literature [25, 106–114], thus forming a single motif dataset. This set was analyzed using the BIG_ANTIAN_DB and AntiangioP+red databases [43].

Here, control sequences were provided using the full set of peptides considered as negative in AntiangioPred, consisting of the 135 negative peptides extracted from proteins in the Swiss-Prot database, together with those randomly generated under the same criteria. In total, 810 control peptide sequences from BIG_ANTIAN_DB were provided to SEA. The input sequences included the totality of positive peptides reported in this database, totaling 202 positive peptides.

Several parameters were taken into account during the enrichment process, and one of them was establishing the E-value threshold, set at less than or equal to 10. Consequently, post-analysis, the motifs displaying statistical significance superior to the specified enrichment E-value were acquired. The background model utilized was the Markov model [61], which is the default for this particular tool. Furthermore, central alignment was employed for the sequence alignment for site positional diagrams. Finally, motifs that showed sta-

tistical significance in all three data sets were retained.

Chapter 4

Results and Discussion

4.1 Metadata Networks (METNs)

4.1.1 Database METN

The primary data source for the AAPs was predominantly the SATPdb database [115], followed by the Cancer PPD database [116]. These two databases were identified as the most central nodes, as illustrated in **Figure 4.1 A**. A total of 209 nodes were interconnected with the SATPdb, signifying that SATPdb exhibited the highest betweenness centrality and node degree values, especially when contrasted with Cancer PPD, which displayed connections with merely 14 nodes. Within the SATPdb repository, the AAPs under consideration were confined to a subset of 1099 anticancer peptides[25]. This repository encompassed peptide sequences that were documented to span between 2 to 50 amino acids in length. In contrast, Cancer PPD was curated manually, drawing data from a compilation of published research articles, patents, and assorted data sources. Among the 3491 anticancer peptides cataloged within Cancer PPD, those of specific interest were of paramount focus [116].

Database MENT proves advantageous when the objective is to conduct searches encompassing the most significant databases pertaining to the antiangiogenic activity of peptides. This resource effectively steers users towards the databases with the highest frequency of peptide reporting in this context.

4.1.2 Function METN

Function MENTs accentuated the supplementary functionalities linked to the peptides under scrutiny. Predominantly and conspicuously emphasizing the foremost antiangiogenic activity, which emerged as the most salient and central node, alongside the interconnected anticancer activity that shared an identical linkage with these 209 nodes. Subsequently, there was the antitumor activity, displaying connections with 14 peptides. Conclusively, the functions encompassing drug administration and tumor localization were represented,

each connected to a distinct node.The distribution of these peptides, illustrating the delineated pattern, is visually presented in **Figure 4.1 B**.

Peptides exhibiting anti-angiogenic activity hold potential applications across various biomedical research domains, particularly in the context of cancer treatment, as previously mentioned. As observed within this METN, the correlation between antiangiogenic peptides and the diverse activities demonstrated can be outlined as follows:

- *Anticancer activity:* Anticancer activity is closely related to inhibiting angiogenesis in the context of peptides [117]. Research in anti-angiogenic peptides offers promising therapeutic approaches to combat cancer by targeting one of the key biological processes underpinning cancer cell growth and progression. It is clear that a significant portion of the PAAs reported in Starpep DB exhibit anticancer properties, as reflected by the numerous interconnections observed. Within this data set is the peptide starPep_23641, also known as scospondistatin, which consists of a 19 amino acid sequence: GPWEDCSVSCGGGGGEQLRSR. This peptide is derived from proteins containing SCO-spondin type I thrombospondin motifs and has been shown to have the ability to inhibit both endothelial cell proliferation and migration [118]. Another peptide belonging to this category is TSWSQCSKTCGTGISTRV, which is composed of 18 amino acids and is also known as Cyrostatin. This particular peptide is sourced from proteins within the CCN protein family. Through in vitro studies, it has been observed to possess the capacity to effectively impede both the proliferation and migration of human umbilical vein endothelial cells[119]. Although only two examples are presented, accompanied by their respective details, it is clear that many more peptides possess this function.

- *Antitumor Activity*: Angiogenesis, the process of generating new blood vessels, assumes a critical role in supplying nutrients and oxygen to rapidly proliferating tumor cells [120]. As a result, the inhibitory effect of AAPs on angiogenesis can moderately curtail the growth and dissemination of tumors. Among these peptides is the peptide starPep_08183, whose sequence is AAVPIVNLKDELLFPSWEALFSGSE, which is composed of 25 amino acids derived from human endostatin. Another peptide belonging to this metadata is starPep_11149, with 28 amino acids: LGQSAASAHHAYIVLAIENSFMTASKKK, derived from lyophilized recombinant

human endostatin, where the antiangiogenic activity of endostatin is determined by its N-terminal region [121]. In addition, IVRRADRAAVP is an 11 amino acid peptide identified in StarPep DB as starPep_10474, derived from the amino-terminal end of one of the most effective negative regulatory proteins in neovascularization, endostatin. Endostatin shows great potential for the treatment of tumors and diabetic retinopathy. However, its short half-life and poor stability have hindered its widespread use. However, PEG conjugation has failed to provide additional bioactivity to the modified protein [122].

- *Drug Delivery Activity:* AAPs possess the capacity to serve as carriers for drug delivery, courtesy of their targeted recognition of angiogenic blood vessels. This characteristic renders them promising contenders for the advancement of precision-targeted drug delivery systems [123]. This metadata contains a unique peptide, the cyclic decapeptide CTTHWGFTLC (starPep_17101), which acts as an inhibitor of the matrix metalloproteinases MMP-2 and MMP-9. This peptide not only suppresses tumor cell and endothelial cell migration in vitro but also inhibits tumor growth by harboring tumor vasculature in vivo [124].

- *Tumor Homing:* AAPs can selectively engage with the vascular network of tumors, thereby presenting themselves as viable contrast agents for enhancing tumor localization through imaging modalities such as Magnetic Resonance Imaging (MRI) or ultrasound [24]. According to StarPep DB, reference is made to the same peptide mentioned in the drug delivery activity.

4.1.3 Origin METN

This Origin METN enable the discernment of the provenance of antiangiogenic peptides. As depicted in **Figure 4.2 A** , the analysis distinctly highlights the category originating from synthetic constructs, characterized by a node degree of 5, signifying its interconnection with five nodes. This delineates that five antiangiogenic peptides are derived from synthetic constructs. Furthermore, it is noteworthy that most of the antiangiogenic peptides depicted in the figure lack an accompanying metadata node that specifies their origin.

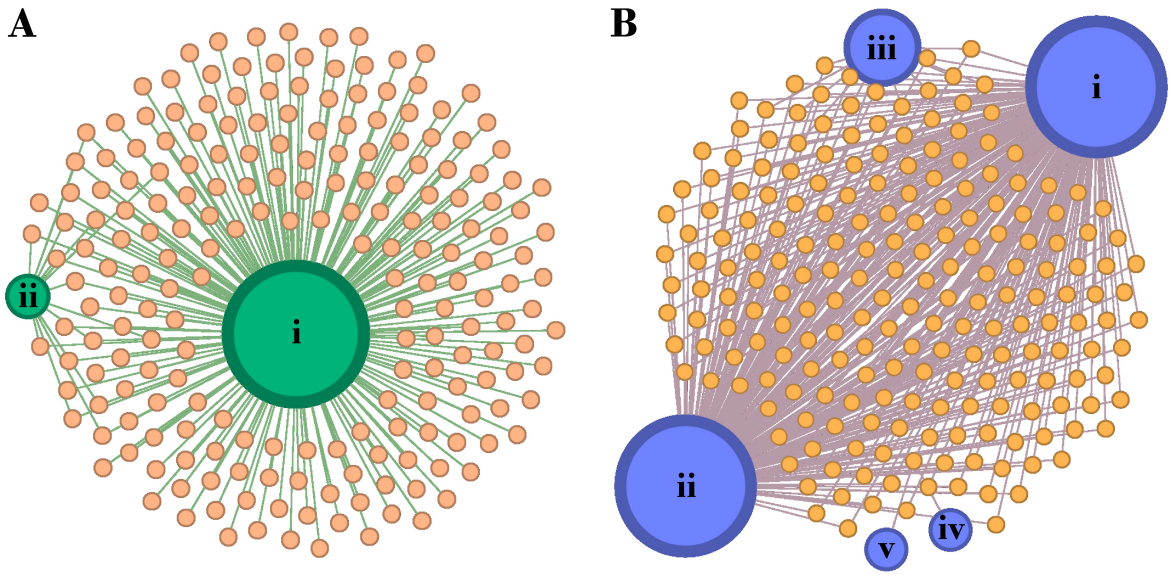

Figure 4.1: Metadata networks (METNs) for (A) Database which describes the database from which the antiangiogenic peptides of StarPep DB have been recovered. The databases are listed according to their range of betweenness centrality: i) SATDPdb database and ii) CancerPPDand. Metadata networks (METNs) for (B) Function which describes the functions associated with antiangiogenic peptides: i) Antiangiogenic, ii) Anticancer, iii) Antitumor, iv) Drug delivery and v) Tumor homing. These networks were visualized in Gephi using Force Atlas layout and edited with Inkscape.

4.1.4 Target METN

The METN Target, akin to its preceding network counterpart, functions as an illustrative platform for the hierarchical taxonomic categorizations. Specifically, it elucidates the distinct species or cellular types within which the antiangiogenic potential of peptides has been assessed. Within **Figure 4.2 B**, the visualization accentuates the prominence of three pivotal nodes: pancreatic cancer, Cancer, and skin cancer. Notably, discernible data pertains to the utilization of these AAPs in combating pancreatic cancer, where a specific tally of 8 peptides has been identified for this therapeutic intent. Additionally, the network manifests a peptide intended for general cancer treatment alongside another designed for addressing skin cancer.

Pancreatic cancer is characterized by its high vascularity, relying on neovascularization for its proliferation and dissemination. The potential implementation of antiangiogenic agents holds promise in addressing this malignancy. Several investigations have concentrated their efforts in this realm [125]. These antiangiogenic interventions have demonstrated the capability to impede the progression and dissemination of pancreatic

cancer, thereby enhancing patient prognoses. This effect is achieved through inhibiting neovascularization, which is pivotal for the growth of new blood vessels[124]. Among these peptides is starPep_08183, whose sequence consists of 25 amino acids: AAVPIVNLKDELLFPSWEALFSGSE. Also, there is starPep_10026, with the same number of amino acids: GSEGPLKPGARIFSFDSFDGKDVLRHPT, and starPep_13195 whose amino acid sequence is TFRAFLSSRLQDLYSIVRRADRAAV. All of them were synthesized from human endostatin [121].

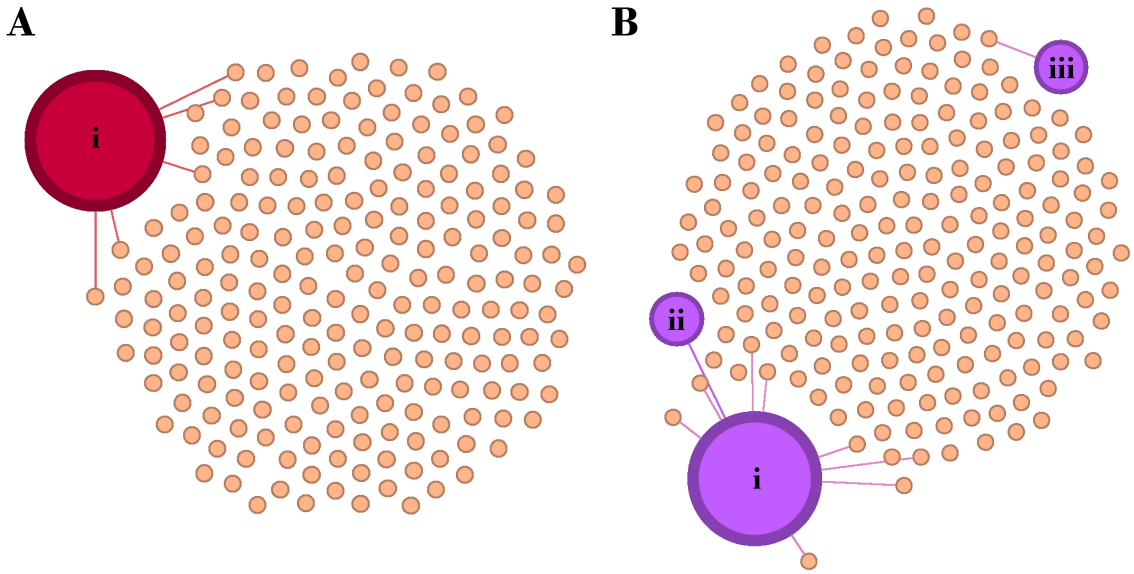

Figure 4.2: Metadata networks (METNs) for (A) Origin which describes the origin from which the antiangiogenic peptides, particularly i) Synthetic constructs and Metadata networks (METNs) for (B) Target which represents the specific target associated with antiangiogenic peptides: i) Pancreatic cancer, ii) Cancer, and iii)Skin cancer. These networks were visualized in Gephi using Force Atlas layout and edited with Inkscape.

4.2 Half-Space Proximal Networks (HSPNs)

A total of 176 entities exhibiting a less than 98% similarity were excluded from the initial set of 209 AAPs retrieved from StarPepDB [26] using a local Smith-Waterman alignment technique [104]. Before establishing the 176-node HSPNs, a judicious threshold for similarity was selected. This phase bears significant importance as it delineates the topological structure and network parameters. Subsequently, pivotal metrics such as edge quantity, modularity, density, ACC, community count, and singleton count were scrutinized.

Consequently, the optimal threshold for constructing the HSPN was ascertained based

on the observed variations in the parameters as mentioned earlier. Thus, the criteria considered in each parameter and the analysis made to choose the best cut-off point are better explained below.

- **Modularity**

  To ascertain the construction of a good HSPN, the objective is to maximize the modularity function. This metric is an effective tool for assessing the quality of the community partitioning, shedding light on both the precision of delineation and the count of individual communities. Essentially, modularity quantifies the proportion of edges that connect nodes within the same community, subtracted by the anticipated value of such connections in a network possessing identical vertex degrees and community structure, but with edges randomly distributed [34].

  The data illustrated in **Figure 4.3 A** shows that the initial similarity cutoff values reflect a relatively modest level. However, notable distinctions emerge: the *Eu* metric exhibits the highest modularity values, while the *Ch* metric registers the lowest values within this initial range. Upon a comprehensive examination of various metrics, a general trend materializes: modularity values consistently remain low within the span of t=0.00 to t=0.50.

  Upon closer scrutiny, a pattern emerges wherein modularity demonstrates an upward trajectory commencing at t=0.50. Remarkably, despite its initially low modularity, the *Ch* metric displays the most rapid increase as t escalates. Conversely,*So, Eu,* and *Ma* metrics exhibit akin tendencies, escalating their modularity at a brisker pace compared to the *Bh* metric. Lastly, the angular separation metric manifests the slowest modularity augmentation rate concerning similarity cutoff changes (t). Consequently, the prudent selection of the modularity parameter is of paramount significance. An optimal choice safeguards against the emergence of widely dispersed networks featuring a profusion of communities, some of which might be artifacts devoid of substantial, informative value.

- **Density**

  This measure delineates the proportion between the count of existing edges within the network and the theoretical count of all conceivable edges [126]. As per earlier investigations, it has been established that an inversely proportional correlation exists between density and the similarity threshold (t) [34]. Applying the refined set

of optimized features results in decreased density levels. Thus, the consideration of an optimal density would involve an intermediary value. This approach is adopted to avoid the loss of valuable network information in the case of excessively low density or the potential obfuscation of pertinent information in the event of an overly high density.

The trend of density concerning changes in the similarity cutoff (t) exhibits remarkable similarity across the various metrics. This trend maintains a near-constant pattern until reaching t=0.45, after which a gradual reduction becomes evident. However, As metric departs from this general pattern, displaying a decrease commencing at t of around 0.60, with the reduction occurring more gradually. A distinctive behavior is notable in the Ch metric case, where density diminution initiates at approximately 0.30, progressing at a notably swifter rate compared to the other metrics. This density behavior is evidenced in **Figure 4.3 B**. To achieve the construction of a precise HSPN, the pursuit of low-density values, around 0.20, is paramount.

- **Average Clustering Coefficient (ACC)**

  The clustering coefficient is the ability to establish a connection between two nodes with a common neighbor. The average clustering coefficient (ACC) is a comprehensive measure of neighborhood interconnectivity. A specific study demonstrated that the ACC curve's apex corresponds to the most favorable clustering outcome. This pinnacle serves as a dependable indicator for identifying the optimal value of the threshold parameter (t)[63].

  Within the scope of this study, a distinct dichotomy in the behavior of the ACC parameter is discernible. Specifically, this dichotomy is evident across all metrics except for the As metric. The ACC remains consistently low across the entire range of t values for the former group. Conversely, the As metric maintains a similarly low ACC value until t=0.7. However, as illustrated in **Figure 4.3 C**, a noteworthy deviation emerges as the metric's ACC attains its zenith at t=0.90.

- **Communities and singletons (*D0*)**

  The computation of the communities or clusters involved the application of the Louvain method. Subsequently, the tally of singletons, denoted as D0 (nodes possessing a degree of zero), and the count of singletons GC (nodes disengaged from the gi-

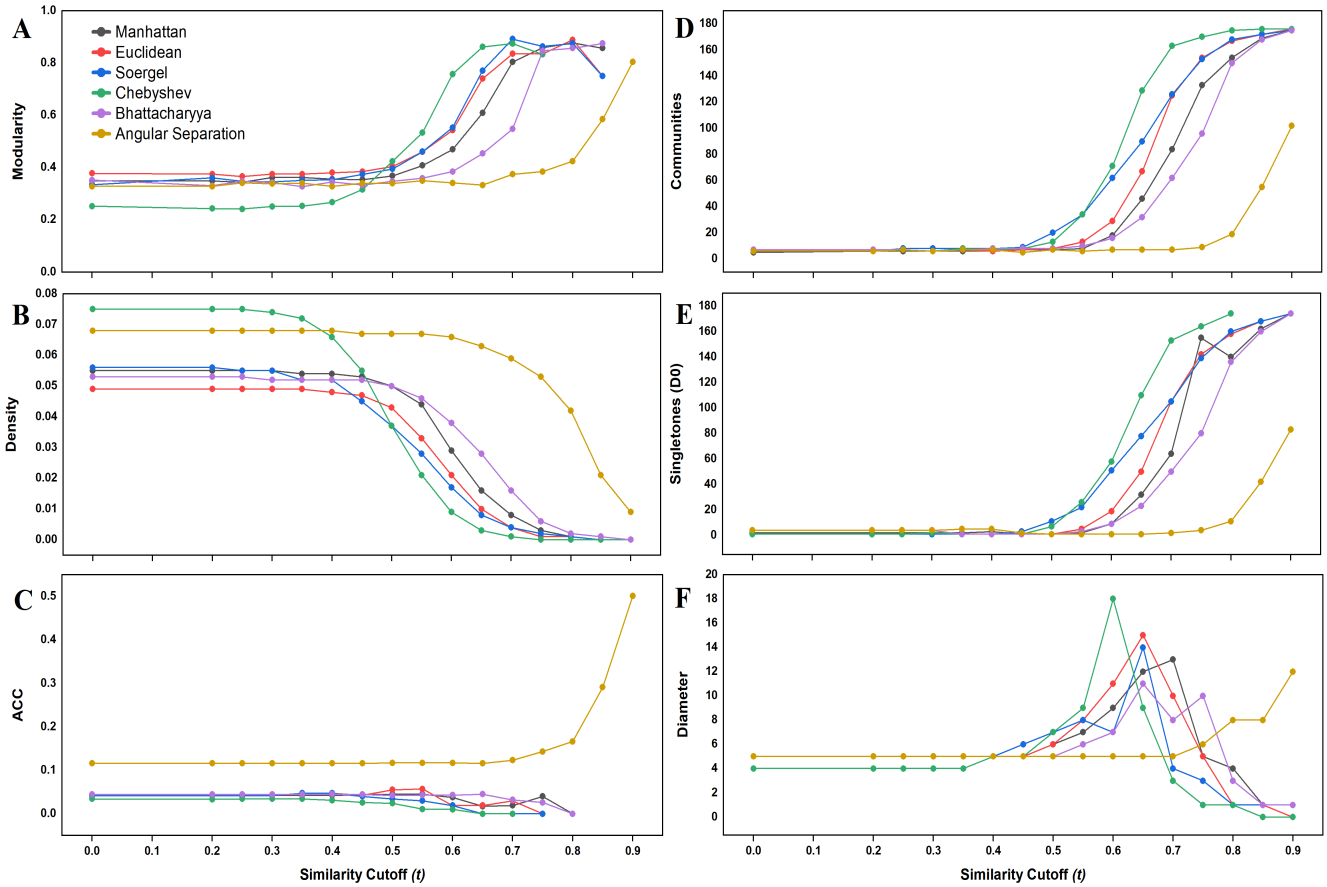

Figure 4.3: Global network parameters of HSPNs established using various metrics and similarity cutoff values t. ACC = Average Clustering Coefficient. This figure was created with Origin and edited with Inkscape

ant component) were computed. This computation served as the basis for selecting networks featuring the most reasonable values of these parameters. Concerning the HSPN, evaluations were conducted for two distinct similarity cutoff values: $t = 0.00$ and the optimal similarity cutoff. The assignment of $t = 0.00$ implies a scenario wherein all nodes are interconnected. In essence, the HSPNs possess the minimum spanning tree as a subgraph; this signifies that neither *D0* nor *DC* persists at this particular t value. The progressive rise in modularity is concurrent with the reduction in density, signifying an enhanced resolution of community structures. A low density network contains many singletons(i.e., unconnected nodes) and is not very informative [126].

The dynamics of the number of communities and singletons (*D0*) related to the similarity cutoff exhibit uniform trends across all metrics. A distinctive pattern emerges, characterized by a substantial surge in both parameters from the thresh-

old of $t = 0.55$, with the exception of As metric as observed in**Figure 4.3 D, E** . Notably, the behavior of the As distance diverges from this norm, with these parameters maintaining lower values up until t = 0.8, after which an increase becomes apparent.

A judicious equilibrium must be sought when determining the optimal value of t concerning these parameters. This involves identifying a t value that balances atypical peptides (singletons) and the count of communities. Such a balance is integral, as reducing network edges leads to an escalation in isolated nodes, thereby concentrating unique elements within the communities.

The distribution of the communities is visualized more precisely in **Figure 4.4**, focusing especially on the Euclidean metric without a cut-off point. In this case, the HSPN consists of six different clusters, each represented with a different color, as seen in the image with each cluster isolated on the right side.

In addition, in the development of the HPSN, additional parameters related to the overall network were determined. These parameters included the measurement of the graph diameter, which indicates the maximum distance between any pair of nodes within the network, as seen in **Figure 4.3F**. Average path length (APL) and average slope were also evaluated.

Tt was essential to identify points of convergence between these general network parameters to determine the optimal value of t for each network metric. Initially, a selection of the best value of t for each parameter was carried out. Several considerations were considered for this choice, such as low-density networks, ideally with less than 20 clusters, and a balanced number of singletons, preferably between 15 and 30. In addition, the corresponding values of ACC and modularity were kept high. The full description of the HSPN parameters for each metric, including their optimal cut-off values t and no cut-off point t=0.00, is comprehensively documented in **Table 4.1**. In total, 12 HSPNs were meticulously constructed.

- **Degree distribution**

  The probability distribution of k, also known as the degree distribution, has been calculated for each metric at their respective optimal cut-off points and also without a cut-off point, as shown in **Figure 4.5** . The left side shows the representations

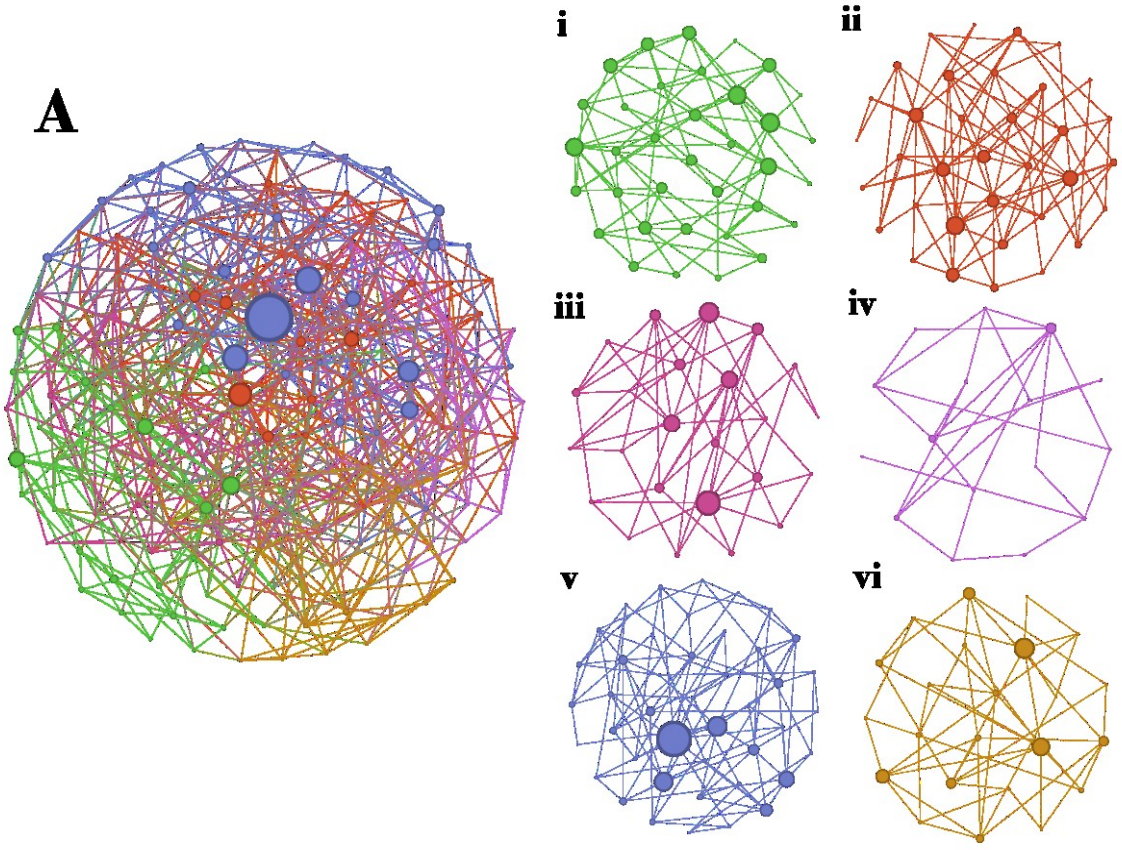

Figure 4.4: A) Graphical representation of HSPN of Euclidean Metric with t = 0.00 showcasing its respective clusters: i) Cluster 1, ii) Cluster 2, iii) Cluster 3, iv) Cluster 4, v) Cluster 5, vi) Cluster 6. Node colors signify distinct peptide communities, and the Hub-Bridge centrality value calculated the size of the node. Layout used: Fruchterman-Reingold[56]. Networks were created with StarPep toolbox [36], visualized in Gephi [87] and edited with Inkscape

corresponding to t=0.00, while the right side shows the representations for the best cutoff point.

In the former (t=0.00), a right-skewed bell-shaped distribution is observed, where the highest probability is between 5¡k¡15 for all metrics, and in addition, a probability of 0 is observed for singletons (when k=0). On the other hand, in the plots with the optimal t-value, a discernible pattern emerges: the degree distribution is predominantly concentrated among the lowest node degrees.

In this case, the probability of k is significantly higher for singletons compared to the t=0.00 representations. For example, for *So* and *Bh*, the probability of k is 0.28, while for *As*, it is 0.23. They are followed by *Ch* with 0.14, *Eu* with a value of 0.10 and finally *Ma* with 0.05. All metrics for the best value of t share that the

Table 4.1: Global network parameters of HSPNs alongside their optimal t values accompanied by their respective networks at $t = 0.00$

No.	Metrics	Similarity Cutoff (t)	Edges	Modularity	Density	ACC	Clusters (no D0)	Singletons (D0)
1	Manhattan (Ma)	0.00	841	0.349	0.055	0.042	5	2
2		0.60	453	0.469	0.029	0.038	18	9
3	Euclidean (Eu)	0.00	757	0.378	0.049	0.044	7	1
4		0.60	328	0.543	0.021	0.019	29	19
5	Soergel (So)	0.00	861	0.334	0.056	0.041	7	1
6		0.60	262	0.553	0.017	0.019	62	51
7	Chebyshev (Ch)	0.00	1161	0.252	0.075	0.034	7	1
8		0.55	316	0.534	0.021	0.011	34	26
9	Bhattacharyya (Bh)	0.00	816	0.352	0.053	0.044	7	4
10		0.70	254	0.548	0.016	0.032	62	50
11	Angular Separation(As)	0.00	1048	0.328	0.068	0.116	6	4
12		0.85	320	0.585	0.021	0.291	55	42

highest concentration of degree distribution occurs at values of k less than about 12. k less than approximately 12. From 12 onwards, the degree distribution decays completely, again reaching zero probability. Therefore, by focusing exclusively on the probability distribution parameter of k, differences can be observed between the HSPNs with and without a cutoff point, as well as differences between each metric.

The selected HSPNs with their best cutoff point and without cutoff point are visually represented on the **Figure 4.6** and **Figure 4.7**.

4.3 HSPNs Scaffold Extraction

To simplify the data analysis, 16 scaffolds were selected for each of the six metrics, with the best t and no cutoff point (t=0.00), with two variations in centrality measure, two types of alignment algorithm and two in r-value. This resulted in a total of 96 scaffolds. The purpose of this experiment was to facilitate the comparison and analysis of structural diversity and redundancy between different HSPN representations by simplifying the number of metrics to work with.

Initially, the metrics were analyzed and evaluated based on the data from the various HSPN global network parameters and the reports in Section 3.2. It was observed that *So*

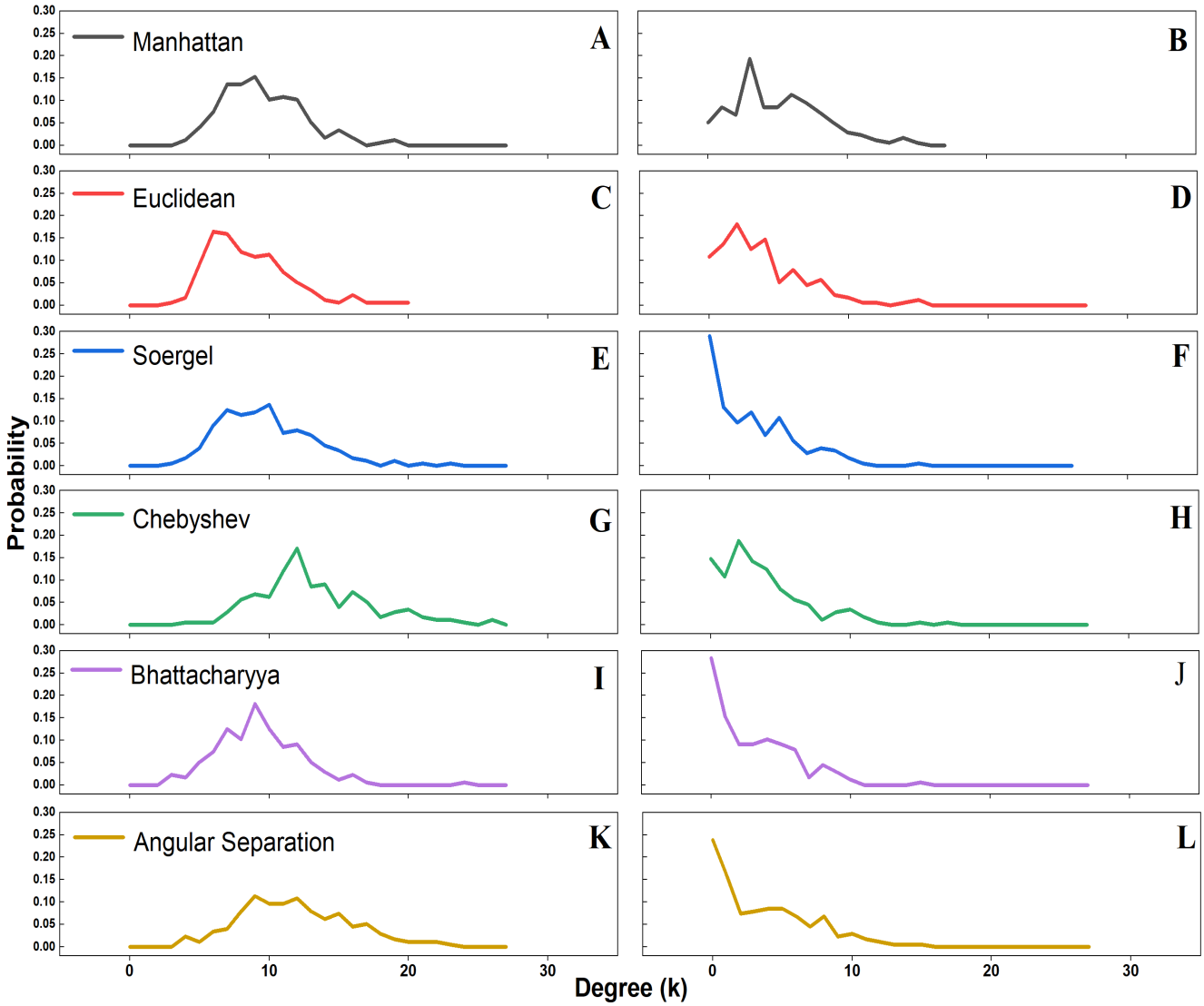

Figure 4.5: Probability of k or degree distribution of the HSPNs with cutoff t = 0.00 (left) and with the best cutoff t (right). Where A) and B) represent Manhattan metric for t=0.00 and t= 0.60 respectively, C) and D) represent Euclidean Metric for t=0.00 and t=0.60 respectively, E) and F) represent Soergel Metric for t=0.00 and t=0.60 respectively, G) and H) represents Chebyshev Metric for t=0.00 and t=0.55 respectively, I) and J) represents Bhattacharyya Metric for t=0.00 and t=0.70 respectively, and finally K) and L) represents Angular Separation Metric for t=0.00 and t=0.85 respectively. This figure was created with Origin and edited with Inkscape

and *Ma* showed similarities to each other and also closely resembled *Eu*. Therefore, only *Eu* was considered among the three metrics.

Then, an analysis was performed using Dover Analyzer based on the *As, Bh, Ch* and *Eu* metrics. A total of 8 analyses were carried out using this tool. It is important to mention that there are differences between the clusters that served as input and differences between the scaffolds that make up each cluster. For the first, the differences lie in three

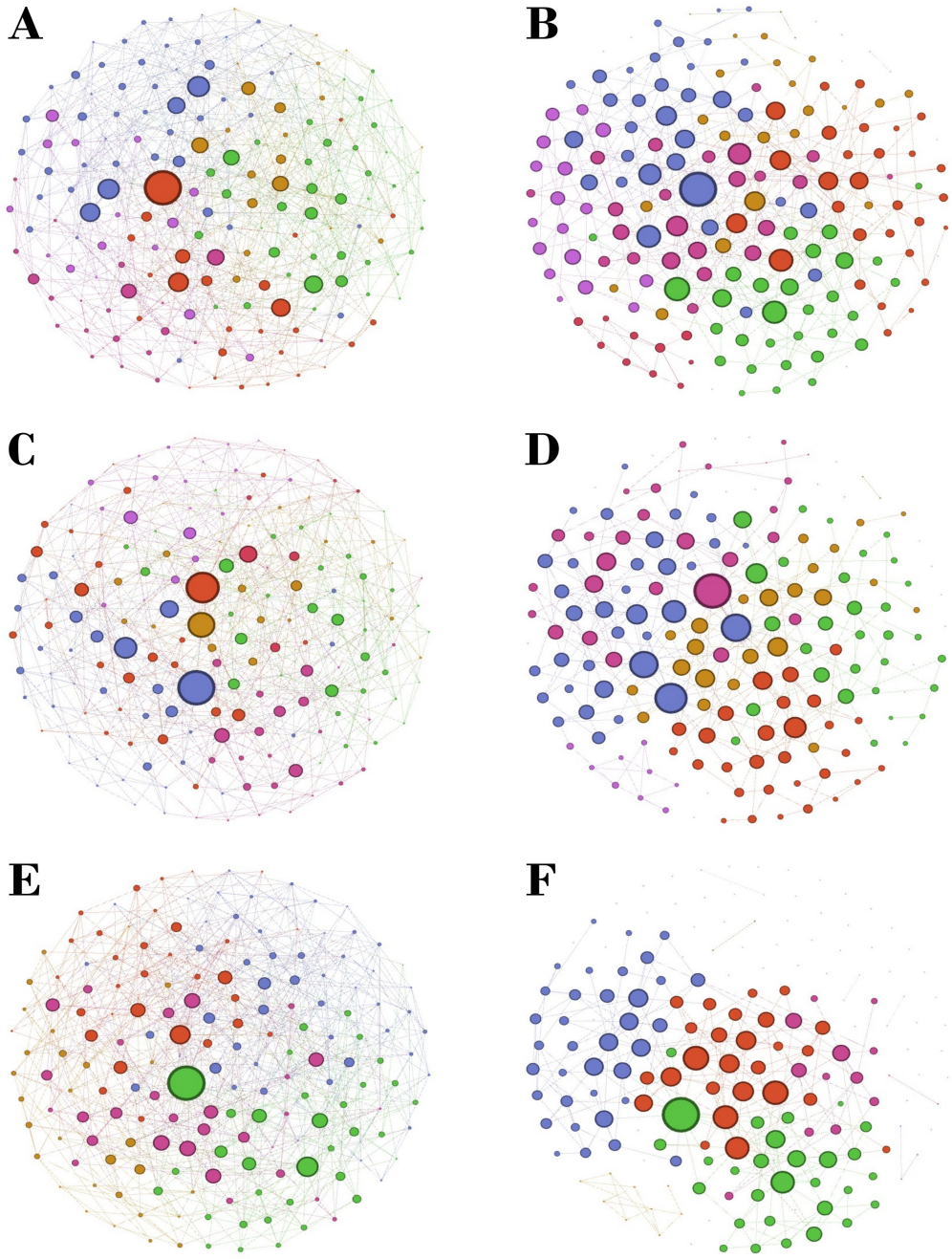

Figure 4.6: Graphical illustration of HSPNs at $t = 0.00$ (on the left), alongside networks displaying the optimal t value for each metric (on the right).) Represents Manhattan metric with t=0.00 (B) Represents Manhattan metric with t=0.60,(C) Represents Euclidean metric with t=0.00 (D)Represents Euclidean metric with t=0.60 (E) Represents Soergel metric with t=0.00, and (F) Represents Soergel metric with t=0.60. Node colors represent communities of peptides, and the node size represents the Hub-Bridge centrality value. Layout used: Fruchterman-Reingold. Networks were created with StarPep toolbox, visualized in Gephi and edited with Inkscape

variations: different measures of centrality, different types of alignment, and different values of r, and each of them has two variations. The differences for the second were the

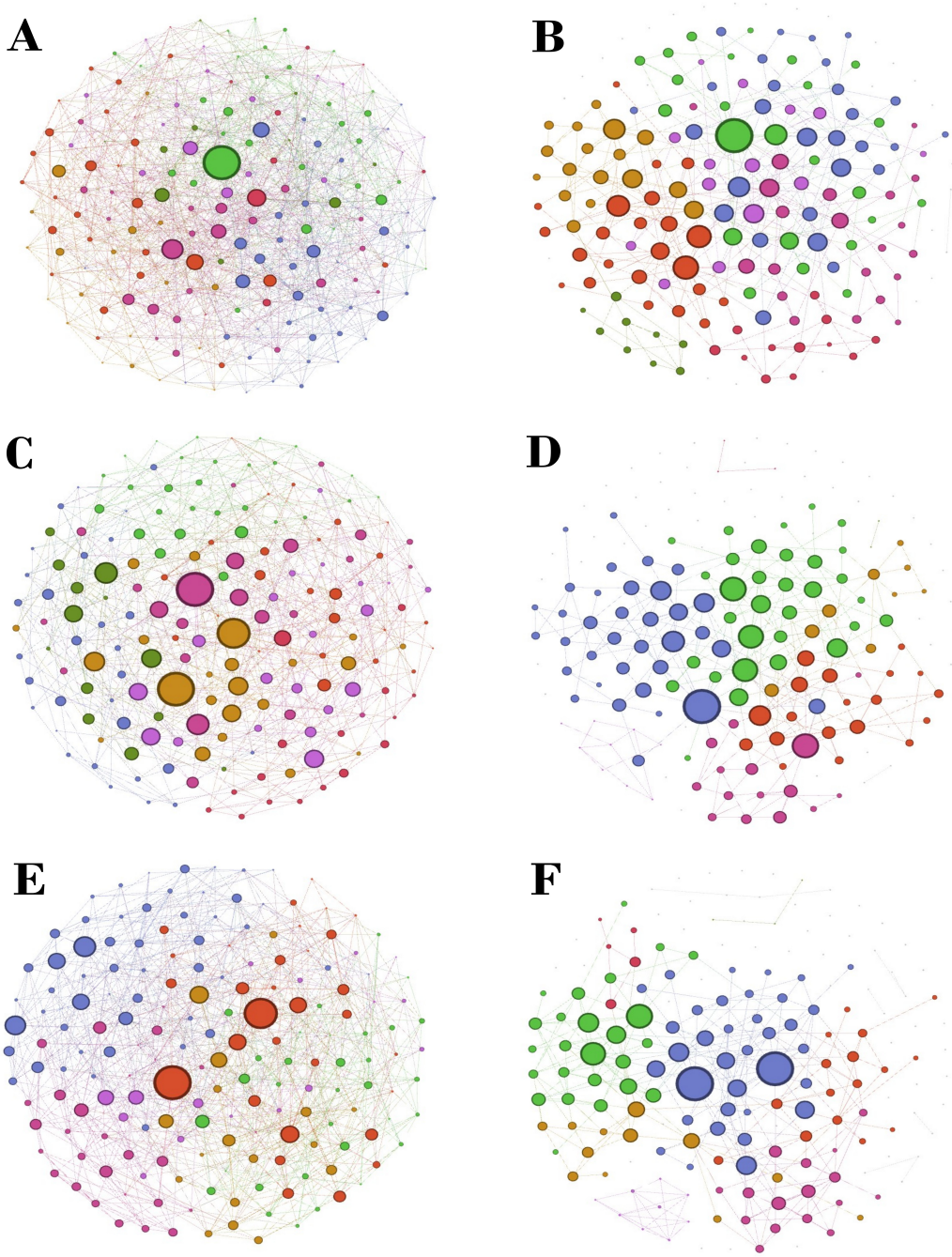

Figure 4.7: Graphical illustration of HSPNs at $t = 0.00$ (on the left), alongside networks displaying the optimal t value for each metric (on the right). (A) Represents Chebyshev metric with t=0.00 (B) Represents Chebyshev metric with t=0.55,(C) Represents Bhattacharyya metric with t=0.00 (D)Represents Bhattacharyya metric with t=0.70 (E) Represents Angular Separation metric with t=0.00, and (F) Represents Angular Separation metric with t=0.85. Node colors represent communities of peptides, and the node size represents the Hub-Bridge centrality value. Layout used: Fruchterman-Reingold. Networks were created with the StarPep toolbox, visualized in Gephi and edited with Inkscape

four different metrics with their corresponding best t and no t.

The results were visualized in heat maps for Similarity overlap, Identical Overlap, and an additional plot comparing all metrics with their cluster numbers and identity threshold. An example of one of the 8 analyses can be observed in **Figure 4.8**. A notable observation is that the heat maps comparing scaffolds were predominantly colored in red, indicating a high level of similarity between the representations of the scaffolds being compared. This analysis suggests that in future research focusing on HSPN and AAP, exhaustive exploration of network topology to determine an optimal similarity threshold (t) may not be necessary. Rather, a t value of zero appears to suffice.

Moreover, this experiment unveiled another pertinent finding: the *Eu* and *Bh* metrics demonstrate significant similarities between them, with Eu displaying the largest differences compared to A*s* and *Ch*. Consequently, in the subsequent sections of this study, only the metrics of *As*, *Ch*, and *Eu* will be considered, using a cutoff point with a t value equal to zero.

4.4 Physicochemical Descriptors

Before approaching motif discovery, conducting a detailed analysis of each group was crucial to considering their structural and physicochemical characteristics, commonly referred to as descriptors. By examining the disparities between groups based on these global peptide descriptors, we can identify the key parameters associated with each metric, as summarized in **Table 4.2**, which presents the average of each parameter for each metric. This comparative analysis sheds light on several significant observations.

For the *As* metric, Cluster 2 presented peptides with the longest sequence length, followed by Cluster 1, while Cluster 6 exhibited the shortest length. As for the net charge, Cluster 1 was the only cluster with negative values. Cluster 3 showed the opposite trend, with the highest positive charge, and the other clusters oscillated in similar positive ranges between 0.5 and 1.5. This observation is consistent because antiangiogenic peptides usually have a net positive charge, facilitating interaction with cell membranes as they contain many anionic phospholipids [78]. In relation to the isoelectric point, relative homogeneity is observed between each group, with Group 3 having the highest value for this parameter, 10.4. However, the ranges for this parameter oscillate between 6.5 and 8.5 for the rest of the clusters. Regarding molecular weight, Clusters 3, 4, 5, and 6 have

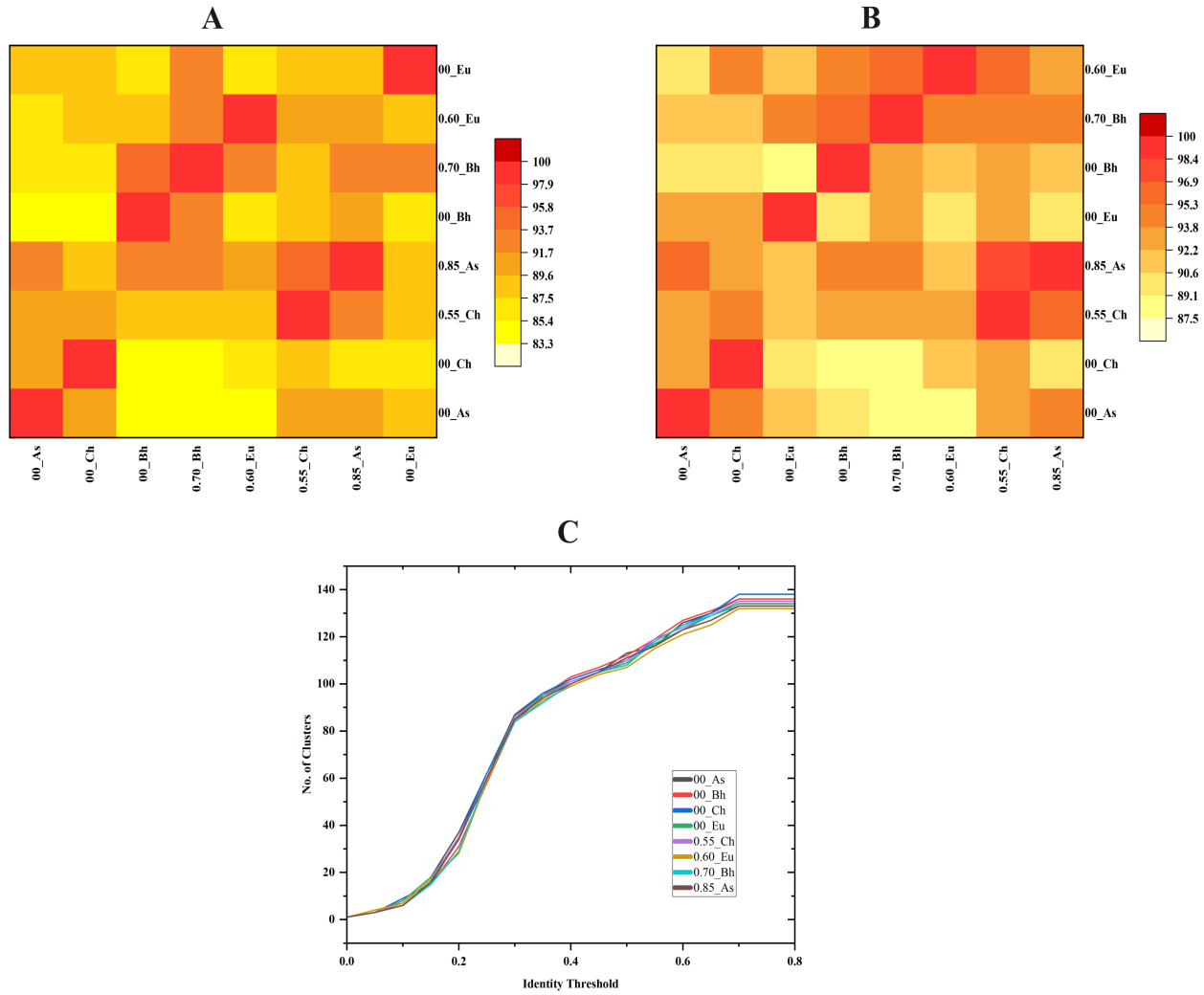

Figure 4.8: Representation of parameters considered for analyzing scaffolds extraction in Dover Analyzer. A) Similarity overlap, B) Identical Overlap, C) Diversity Ratio. Created with OriginPro2023[77] and edited with Inkscape.

very similar values, with Cluster 2 having the highest value.

The highest Boman index corresponds to Cluster 3, which agrees with the average hydrophilicity since both parameters are proportionally related. As for the aliphatic index, the lowest is for Cluster 3, and the highest is for Cluster 2. This parameter, associated with the relative volume occupied by the aliphatic side chains, is considered a positive factor for increasing the thermostability of the peptides. In contrast, the instability index, inversely proportional to the aliphatic index, has the highest value in Cluster 3. This information on physicochemical parameters provides insights into similarities and communities within the chemical space. Apparently, certain clusters may appear similar, but a closer analysis of

Table 4.2: Averages of the descriptors calculated for the HSPNs obtained from *As*, *Ch* and *Eu* metrics using StarPep toolbox.

Metric	C	Length	Net Charge	Iso. Point	Mol. Weight	Boman Index	Avg. Hydr.	Alip. Index	Inst. Index	Avg. Gravy
As	1	22.903	-0.323	6.604	2547.484	1.459	0.892	96.071	46.885	-0.151
	2	24.353	1.706	9.080	2756.601	2.233	0.935	83.838	45.539	-0.441
	3	18.448	4.552	10.384	2170.194	4.718	1.051	18.3766	69.984	-1.771
	4	18.922	0.882	8.174	1999.703	1.840	0.893	49.121	59.924	-0.465
	5	16.875	0.937	8.557	1978.488	2.327	0.931	61.363	42.977	-0.796
Ch	1	16.233	0.033	7.077	1782.121	1.249	0.877	72.734	55.609	-0.095
	2	17.869	2.130	9.400	2062.737	2.796	0.961	66.617	42.464	-0.689
	3	18.900	6.300	11.616	2229.473	5.172	1.062	17.560	96.745	-1.901
	4	24.440	0.960	8.304	2679.371	2.464	0.931	36.058	64.396	-0.798
	5	20.103	0.718	7.969	2338.203	1.926	0.914	89.789	39.338	-0.389
	6	15.000	1.100	8.233	1685.515	1.662	0.875	46.981	31.811	-0.362
	7	21.111	-0.111	7.271	2257.266	0.716	0.839	91.368	51.170	0.084
	8	18.364	-0.636	6.293	2050.198	2.498	0.945	57.889	32.591	-0.889
Eu	1	28.333	0.667	8.215	3135.529	2.011	0.908	57.955	57.012	-0.524
	2	19.371	4.771	10.780	2262.840	4.370	1.032	25.642	71.150	-1.536
	3	17.953	0.046	6.883	1944.575	1.077	0.861	75.151	44.635	0.018
	4	14.400	-0.429	6.978	1634.433	2.465	0.943	51.787	47.538	-0.918
	5	16.250	1.062	8.680	1876.940	0.812	0.839	110.096	41.951	0.338
	6	22.130	1.869	9.051	2539.277	2.267	0.933	92.430	52.552	-0.499

their chemical properties reveals significant differences, thus justifying their classification as separate clusters **Figure 4.9**

In the examination of descriptors for Metric *Ch*, notable variations emerge (**Figure 4.10**), particularly in the context of peptide chain lengths across clusters. Cluster 4 stands out with the most extended peptide chains, while Cluster 6 exhibits the shortest ones. Regarding net charge, three distinct cases warrant attention: Cluster 1's charge hovers close to the null boundary but is not exactly 0, whereas Clusters 7 and 8 present negative values. Cluster 3 boasts the highest isoelectric point, a value directly proportional to the net charge. The isoelectric point represents the pH at which the net charge of a protein becomes zero. Moreover, Cluster 4 claims the highest molecular weight among the clusters.

Assessing thermostability through the aliphatic index reveals that Cluster 7 exhibits superior stability. The Grand Average of the Hydropathicity Index (Gravy), along with the Average Hydrophilicity, predicts the hydrophobicity and hydrophilicity of the peptides. Notably, all clusters, except Cluster 7, demonstrate a negative GRAVY value,

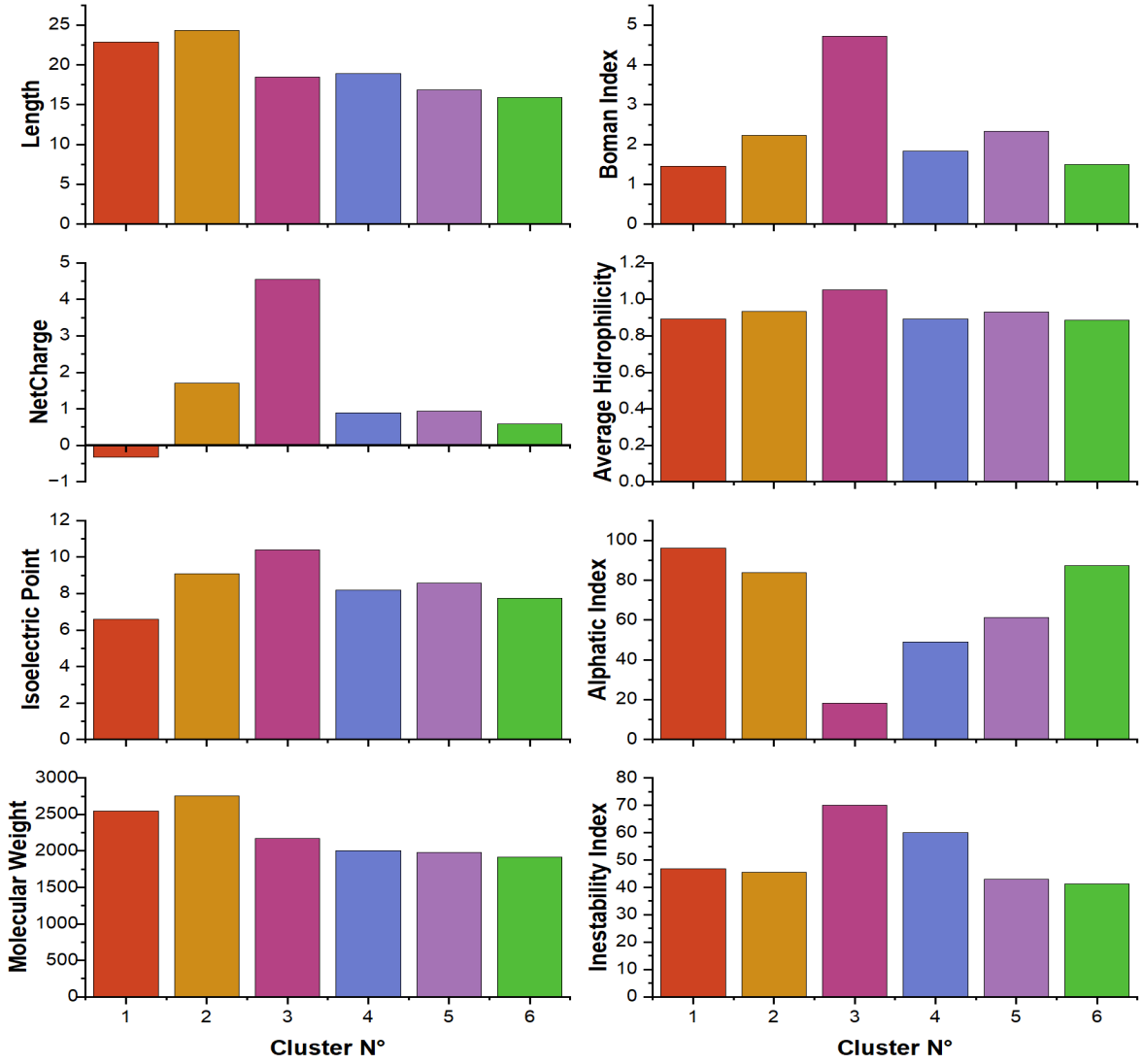

Figure 4.9: Representation of the average values of various molecular descriptors computed for peptides within the clusters of the Angular Separation Metric in the HSPN, where no cutoff is applied using StarPep toolbox. Each color represents a different cluster. This figure was created with OriginPro2023 and edited with Inkscape

signifying a hydrophilic nature. This aligns with the hydrophilicity assessment, where Cluster 7 emerges as the most hydrophobic and the least hydrophilic.

Examining the Boman index, Cluster 7 boasts the lowest value, while Cluster 3 has the highest, establishing an inverse relationship with both the Aliphatic Index and Average Hydrophilicity. Finally, the dataset with the lowest Instability Index is Cluster 6, contrasting with Cluster 3, which presents the highest instability index among the clusters.

Similar trends are observed in line with the analogous analysis of the *Eu* metric, mirroring the approaches taken for *As* and *Ch* (**Figure 4.11**). Cluster 1 stands out

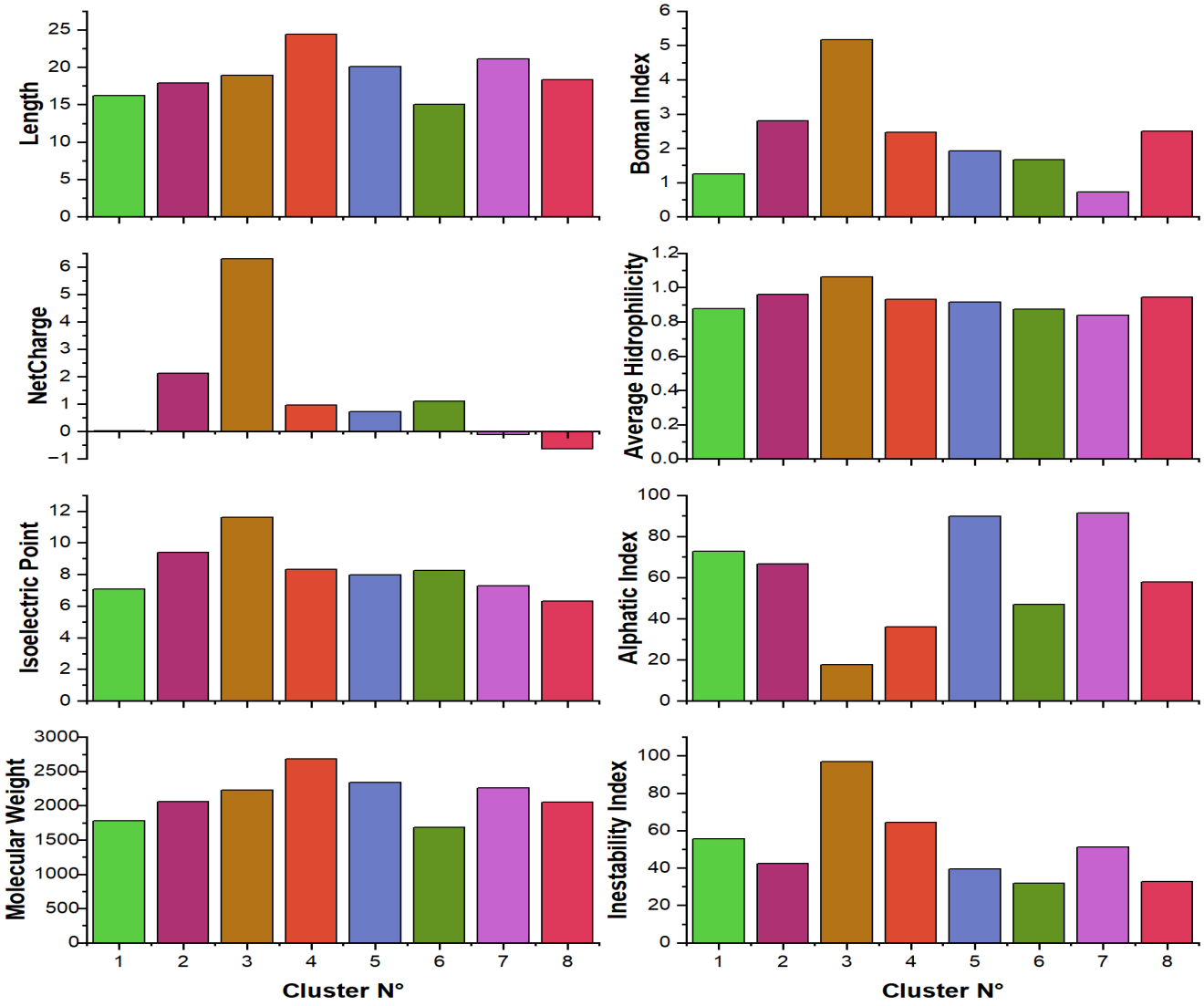

Figure 4.10: Representation of the average values of various molecular descriptors computed for peptides within the clusters of the Chebyshev Metric in the HSPN, where no cutoff is applied using StarPep toolbox. Each color represents a different cluster. This figure was created with OriginPro2023[77] and edited with Inkscape[72]

for having peptides with the lengthiest amino acid chains. Turning to net charge, an interesting observation is the prevalence of negative charges in Cluster 3. Notably, the isoelectric point follows a similar pattern among specific clusters, specifically Cluster 3 and others, drawing attention due to their uniformly low values.

Examining molecular weight, Cluster 1 peptides exhibit the highest values, indicating a distinctive characteristic within this cluster. The Boman Index reaches its lowest values in Clusters 3 and 5, which interestingly are the only sets featuring positive values for GRAVY. Furthermore, Clusters 3 and 5 also boast the lowest values for Average

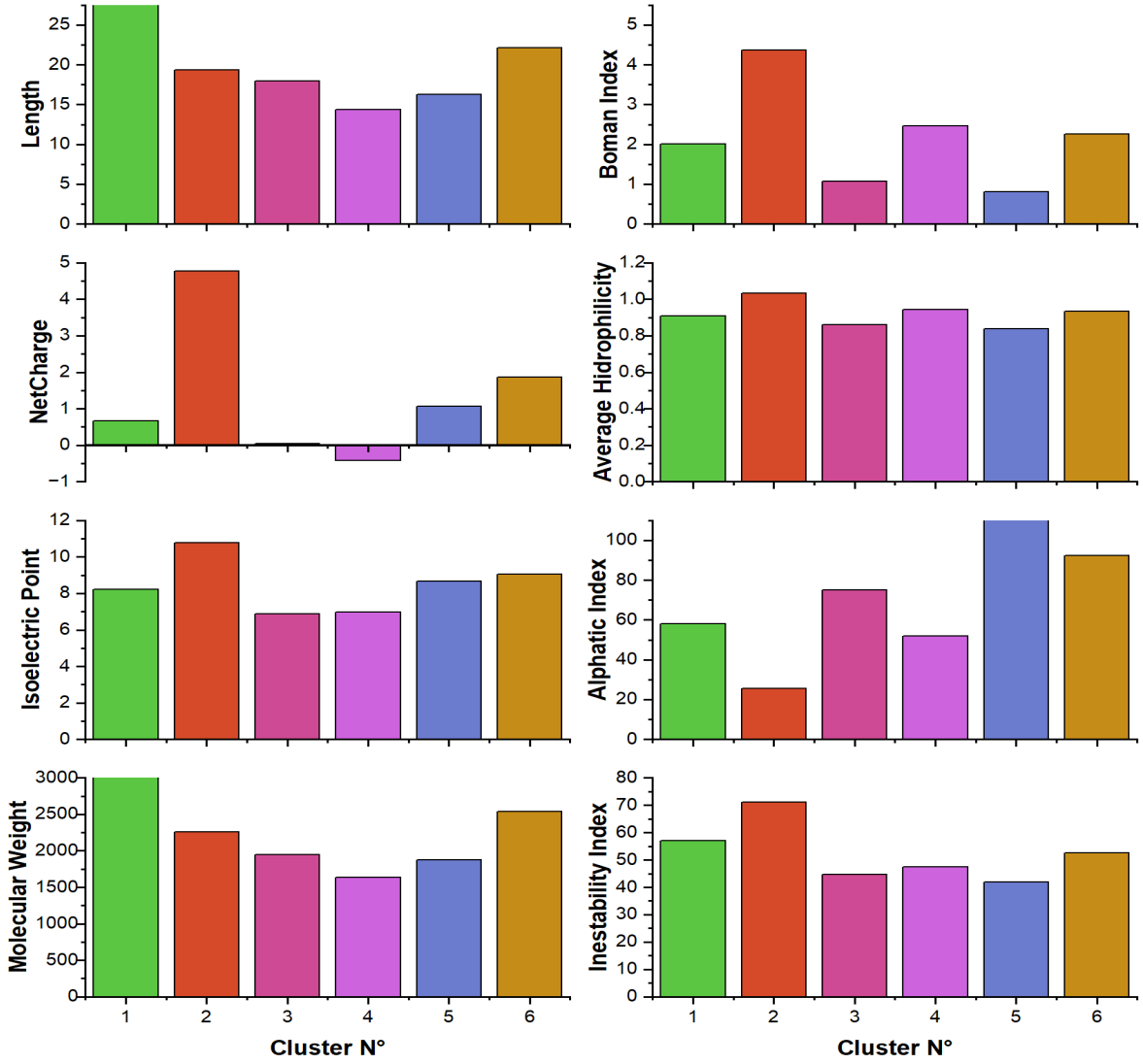

Figure 4.11: Representation of the average values of various molecular descriptors computed for peptides within the clusters of the Euclidean Metric in the HSPN, where no cutoff is applied using StarPep toolbox. Each color represents a different cluster. This figure was created with OriginPro2023 and edited with Inkscape

Hydrophilicity, suggesting their designation as the most hydrophobic, while Cluster 2 emerges as the most hydrophilic.

The Aliphatic Index, with its highest value in Cluster 5, implies superior thermal stability for this cluster, a finding consistent with the comparison to the instability index, which highlights Cluster 2 as the most unstable among the groups. These considerations provide insights into the distinct biochemical characteristics and potential functional implications of the peptide clusters within the context of the *Eu* metric

Significant variations were identified in each metric and cluster analyzed. For example,

it was observed that peptide length tends to be longer in *Eu* compared to *Ch* and *As*. The length of peptides may influence their ability to interact with proteins and their structural stability. Longer peptides may have more potential binding sites and more stable secondary structures, which could increase their affinity for target proteins and their resistance to adverse environmental conditions. In contrast, shorter peptides may have more specific but less extensive interactions, and more flexible or disordered structures, which may affect their stability and susceptibility to degradation.

As for the net charge, a wide variability was recorded, with some clusters exhibiting positive, negative, or near-zero net charges. The *Ch* metric showed greater variability in net charges, with some clusters with more prominent negative values. About the isoelectric point, it was observed that some clusters presented higher isoelectric points, indicating a higher propensity to positive charges. In comparison, others showed lower points, suggesting a tendency to negative charges. In addition, differences in molecular weights between clusters were detected, which may influence their bioavailability and ability to cross cell membranes. The Boman index variability suggests different antimicrobial activity levels among the clusters.

Also, the average hydrophilicity and aliphatic index showed diversity, which may affect the solubility and structural stability of the peptides. This analysis provides detailed insight into the physiochemical properties of peptides in different metric contexts, essential to better understand their structure, function, and potential applications in various biomedical and therapeutic areas. Finally, variability in instability index values was observed, which may indicate differences in the structural and biological stability of the peptides, with a higher index associated with greater susceptibility to enzymatic degradation.

After examining the descriptors overall for each metric, rather than analyzing them by cluster, the results reveal remarkable consistency among the distance metrics employed in the analysis. Minimal variability is observed between the values of each metric, suggesting that they provide similar insight into the physicochemical descriptors. Therefore, it can be concluded that the trend in relation to the descriptors is conserved across all metrics studied (**Figure 4.12**).

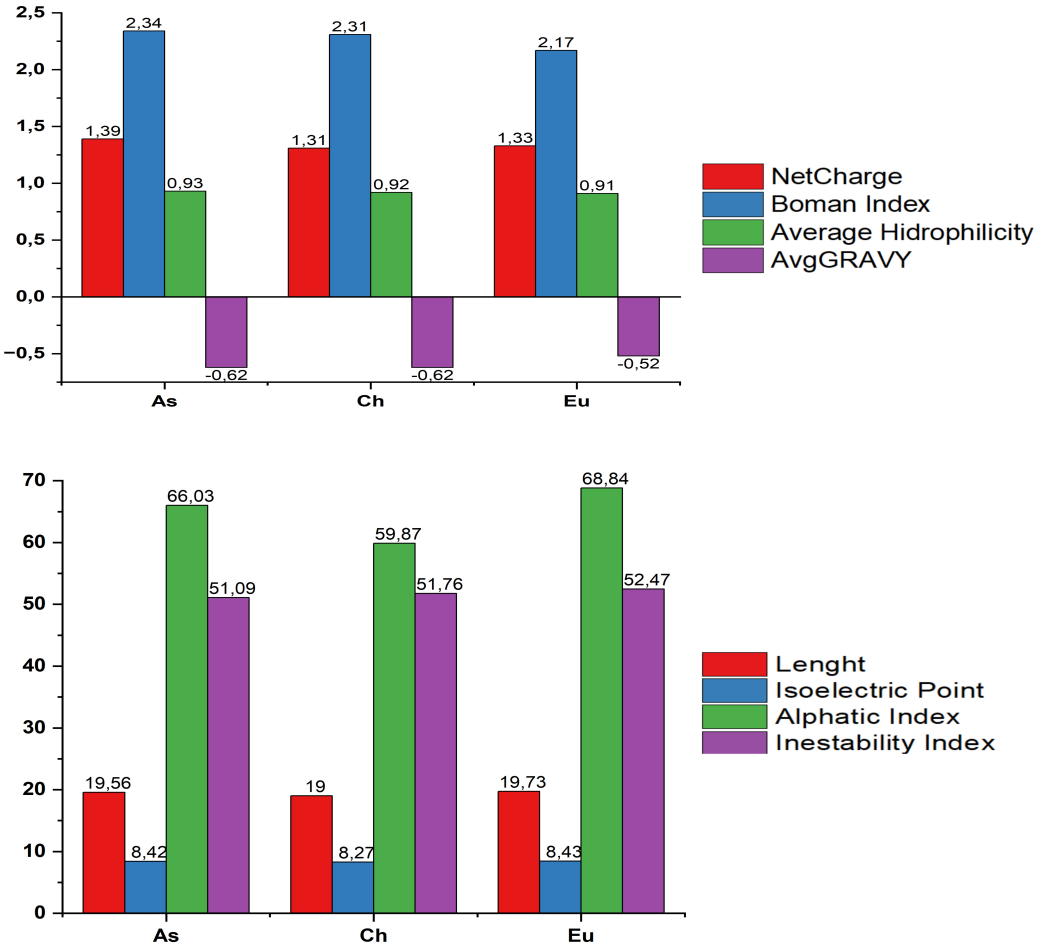

Figure 4.12: Average of the physicochemical descriptors of the Angular Separation, Chebyshev and Euclidean metrics. This figure was created with OriginPro2023 and edited with Inkscape

4.5 Motif Discovery

After completing the community analysis for each metric, the corresponding motif set was identified using MEME Suite's STREME. STREME has previously been reported to be the most accurate and sensitive algorithm among its state-of-the-art competing partners for motif discovery [127]. STREME shows an advantage over previously reported algorithms [128, 129] as STREME uses a position weighting matrix (PWM) to efficiently count position matches of a candidate motif against a Markov model derived from the input sequences.

In this process, STREME generated a control set by combining the input sequences

while preserving the lower-order statistics. In this way, motif discovery focused on the most relevant motifs. Statistical analysis was performed by comparing the occurrence of these motifs in the primary sequences with that of the control set.

The input sequences consisted of the AAP sequences from each community for each metric, with a total of 20 sets of sequences entered into STREME. Thirty motifs were identified for the *Eu* metric with $t = 0.00$, 40 motifs for *Ch* with $t = 0.00$, and 28 motifs for *As* with $t = 0.00$, with a total of 98 motifs discovered. Subsequently, an intermetric analysis of the motifs was carried out, eliminating duplicates and condensing those that incorporated some motifs to others. As a result of this analysis, 74 unique motifs were obtained (**Table 4.3**)

Importantly, only the GRG fragment was repeated among the metrics in *As*, *Ch* and *Eu*. The identification of a motif in all three metrics increases confidence in its significance, since the consistency in the results between different metrics suggests a robustness in its presence, which reinforces its credibility. This consistency indicates that the motif is independent of the method of analysis used, which increases confidence in its significance. Likewise, the motifs GVQTR and QKR were repeated in both *Ch* and *As* metrics. Several similar, although not identical, motifs were identified. For example, the PEAP motif was found in the *Eu* metric, while PEAPF was in the *As* metric. Similarly, the LKK motif was found in the *Eu* metric, while LKKF was observed in the *Ch* metric.

Three motifs were identified without regard to community diversity, using the 209 AAPs as input sequences. The StarPep DB fasta file was extracted and used as input file in STREME. These motifs were referred to as HSPN anti-angiogenic motifs.(**Table 4.4**)

In previous studies documented in the literature, certain motifs have been identified. In particular, AntiAngioPred reported the discovery of 22 motifs using MERCI software[25]. This method is based on the condition that the motif must be present in at least 10% of the total peptides in the positive data set, and its maximum length is five spaces. Other work has also reported the discovery of 10 additional motifs, which are listed in **Table 4.5**

Among the motifs reported in the literature, those discovered using the metrics and motifs found without considering community diversity, there is a single match of a three-peptide fragment. The motif reported is NGR [112] , while the one discovered in the *Eu* metric is also NGRE.

Table 4.3: Motifs identified by STREME utilizing community data from HSPNs generated using Euclidean (*Eu*), Chebyshev (*Ch*), and Angular Separation (*As*) metrics without applying a cutoff.

No.	Metric	Motif	Cluster	Cluster Size	Matches (+) Seq.	Matches (-) Seq	Sites (%)	Score	Match Threshold
1	Eu	CSSR	1	24	6	0	25.00	0.011	0.935
2		ETWRTE			4	0	16.70	0.055	3.254
3		QRPS	2	35	14	0	40.00	0.000	6.785
4		RTR	3	43	11	0	25.60	0.000	7.645
5		PFL			14	1	32.60	0.000	6.809
6		RDI			5	0	11.60	0.028	9.722
7		TGAL	4	35	10	0	28.60	0.000	1.212
8		EEDP			6	0	17.10	0.012	9.615
9		HRI			6	0	17.10	0.012	9.066
10		KNWP			5	0	14.30	0.027	10.190
11		TVT	5	16	6	0	37.50	0.009	7.257
12		FST			3	0	18.80	0.110	11.252
13		IKRY			3	0	18.80	0.110	1.000
14		VRRA	6	23	7	0	30.40	0.005	8.962
15	Ch	STRI	1	30	8	0	26.70	0.002	10.703
16		WSSCS			8	0	26.70	0.002	5.064
17		ING	2	23	7	0	30.40	0.005	7.979
18		ADRAA			4	0	17.40	0.054	1.717
19		PSDK	3	20	14	0	70.00	0.000	7.208
20		RRP			7	1	35.00	0.046	6.939
21		APC	4	25	7	1	28.00	0.049	7.972
22		APF	5	39	13	0	33.30	0.000	7.180
23		GRELCL			11	0	28.20	0.000	10.134
24		IIEK			11	0	28.20	0.000	7.272
25		FGN	6	10	3	0	30.00	0.110	4.238
26		ISNA			3	0	30.00	0.110	12.109
27		GCG			2	0	20.00	0.240	9.018
28		GYCS			2	0	20.00	0.240	11.078
29		HGKG			2	0	20.00	0.240	13.746
30		GAR	7	18	7	0	38.90	0.004	3.305
31		ATC			6	0	33.30	0.010	8.374
32		LHLV			5	0	27.80	0.023	10.337
33		EGL			7	1	38.90	0.044	7.152
34		TGKI	8	11	3	1	27.30	0.110	1.866
35		HPH			5	1	45.50	0.160	3.685
36		PEN			5	0	45.50	0.160	9.291
37		DDDD			1	0	9.10	0.500	14.139
38	As	SPT	1	31	14	0	45.20	0.000	6.270
39		KVI			14	1	45.20	0.000	6.503
40		LAS	2	17	11	1	64.70	0.001	6.234
41		SLD			7	0	41.20	0.004	6.853
42		IVL			5	0	29.40	0.022	4.749
43		TFE			5	0	29.40	0.022	11.016
44		GRK			7	1	41.20	0.043	7.247
45		WSDC	3	29	19	0	65.50	0.000	6.965
46		CSASCG	4	51	19	0	37.30	0.500	1.051

No.	Metric	Motif	Cluster	Cluster Size	Matches (+) Seq.	Matches (-) Seq	Sites (%)	Score	Match Threshold
47		HGHK	5	16	6	0	37.50	0.009	8.869
48		NPR			4	0	25.00	0.051	9.797
49		KIKSST			2	0	12.50	0.240	21.045
50		RSRQVR			2	0	12.50	0.240	18.755
51		EVCH	6	32	8	0	25.00	0.002	11.110
52		MPF			8	0	25.00	0.002	10.129
53		RIST			6	0	18.80	0.012	4.826
54		NDYSYW			4	0	12.50	0.057	17.595
55	Eu	[R][KC]GRG[T]	2	35	19	0	54.30	0.000	8.872
	As		3	29	18	0	62.10	0.000	7.261
	Ch		3	20	14	0	70.00	0.000	1.897
	Ch		2	23	5	0	21.70	0.025	2.548
56	Ch	GVQTR	4	25	13	0	52.00	0.000	8.252
	As		4	51	15	0	29.40	0.500	8.329
57	Eu	PEAP[F]	6	23	9	0	39.10	0.001	8.663
	As		1	31	5	0	16.10	0.026	5.429
58	Ch	QKR	2	23	11	0	47.80	0.000	5.965
	As		3	29	4	0	13.80	0.056	9.558
59	Eu	SPL[S]	1	24	12	0	50.00	0.000	7.521
	Ch		8	11	3	0	27.30	0.110	8.034
60	Eu	LKK[F]	5	16	8	0	50.00	0.001	6.995
	Ch		1	30	10	0	33.30	0.000	7.845
61	Eu	CGG[G][V]	3	43	14	0	32.60	0.000	5.843
			1	24	7	0	29.20	0.005	3.737
62	Ch	[SG]PW[E][R][C]	4	25	22	2	88.00	0.000	5.914
	As		4	51	19	0	37.30	0.220	5.029
	Eu		1	24	5	0	20.80	0.025	1.481
63	Ch	VQK[I]	5	39	10	0	25.60	0.001	8.041
	Eu		6	26	13	0	56.50	0.000	7.720
64	Eu	WSPCS[V]	2	35	16	0	45.70	0.000	11.578
			3	43	14	0	32.60	0.000	8.143
65	Eu	YCNI[N][Z]	5	16	9	0	56.30	0.000	7.263
	Ch		1	30	3	0	10.00	0.120	0.707
	As		6	32	8	0	25.00	0.002	10.281
66	As	[D][K]PRR	3	29	12	3	41.40	0.051	7.714
	Eu		2	35	8	0	22.90	0.003	7.821
67	Ch	KRRR[E][K]	3	20	10	1	50.00	0.004	7.494
	Eu		2	35	8	0	22.90	0.003	5.480
68	Eu	[N][P]ASP	4	35	5	0	14.30	0.027	8.840
	Ch		5	39	5	0	12.80	0.027	8.455
69	Eu	[K][E][I]CLD	6	23	12	0	52.20	0.000	9.346
	As		1	31	9	0	29.00	0.001	6.843
70	Eu	NGRE	6	23	9	0	39.10	0.001	10.905

No.	Metric	Motif	Cluster	Cluster Size	Matches (+) Seq.	Matches (-) Seq	Sites (%)	Score	Match Threshold
71	Ch	[V]TCG[D][G][V]	1	30	6	0	20.00	0.012	1.717
			4	25	14	0	56.00	0.000	8.624
72	Ch	[P][W][S]QCS[V]	4	25	12	1	48.00	0.001	8.110
			2	23	7	0	30.40	0.005	1.135
	As		5	16	6	0	37.50	0.009	9.738
73	As	[S][P]SGG[P]	1	31	9	0	29.00	0.001	5.230
	Ch		7	18	4	0	22.20	0.052	2.818
74	Ch	[R]EKQR	3	20	10	0	50.00	0.000	8.238
	As		3	29	11	0	37.90	0.000	7.032

Table 4.4: Motifs discovered by STREME regardless of community diversity.

No.	Motif	Matches (+) Seq.	Matches (-) Seq	Sites (%)	Score	P-Value	Match Threshold
1	TCGGG	33	1	17.2	2.0E-09	0.16	8.618
2	WSPCS	32	0	16.3	2.3E-10	0.29	7.856
3	REKQRP	16	0	8.6	1.5E-05	0.29	15.055

To conclude this stage, a single list of non-redundant anti-angiogenic motifs was created, where motifs discovered using the metrics, those found using all StarPep DB AAPs, and motifs from the literature were merged in preparation for further enrichment. In total, the list comprises 105 motifs.

4.6 Motif Enrichment

The enrichment process allows discerning the most relevant and reliable motifs discovered, ensuring that the occurrences within the chemical space of the anti-angiogenic peptides have true functional and potential significance.

The validation or enrichment process using the SEA algorithm involved two stages. To ensure the reliability of the validation process, the external data set used was subjected to an overlap analysis to obtain non-redundant data and allow a more fruitful analysis. In the first stage, it was carried out with E-values ≤ 10, enriching 82 motifs. Subsequently, the validation process was refined, starting from the 82 previously enriched motifs, by another processing in SEA, this time considering E-values ≤ 0.01, which finally resulted in 53 enriched motifs. By comparing the occurrence of motifs in the primary sequences with those of the control set, SEA provides information on the enrichment and significance of motifs, which can be observed in **Table 4.6**.

Table 4.5: . Motifs reported in the literature for antiangiogenic peptides. 'X' represents a gap.

No.	Year	Motif	Reference
1	1999	HWGF	[106]
2	2023	HGR	[107]
3	2008	HHQK	[108]
4		LVFF	
5	2013	RGD	[109]
6	2009	RTS	[110]
7	2003	YH	[111]
8	2006	NGR	[112]
9	2022	IQ	[113]
10	2018	NITY	[114]
11	2015	CGXG	[25]
12		TC	
13		SC	
14		SPXS	
15		WXSXC	
16		WSXC	
17		SXTXC	
18		SXCXS	
19		CSXT	
20		CXSXT	
21		TXC	
22		SXC	
23		CXGXG	
24		TR	
25		SXTXG	
26		SXPXS	
27		SP	
28		RT	
29		PXW	
30		PXC	
31		CXN	
32		CG	

Table 4.6: Motif enrichment by SEA - Second stage

No.	Motif	P-Value	E-value	TP	Enrichment Ratio	Score Threshold
1	TC	6.52E-23	5.35E-21	63 / 182 (34.6%)	7.98	2.85
2	WSPCSV	1.04E-21	8.52E-20	30 / 182 (16.5%)	123.66	5.75
3	PWSQCSV	5.20E-21	4.27E-19	29 / 182 (15.9%)	119.67	3.14
4	WSSCS	2.60E-20	2.14E-18	28 / 182 (15.4%)	115.68	9.34
5	CSASCG	1.63E-17	1.34E-15	24 / 182 (13.2%)	99.73	12.18
6	RTR	4.23E-17	3.47E-15	65 / 182 (35.7%)	4.88	4.2
7	SC	2.92E-16	2.40E-14	46 / 182 (25.3%)	7.21	2.85
8	WSXC	3.30E-16	2.71E-14	24 / 182 (13.2%)	49.86	9.11
9	XPWERC	5.71E-16	4.68E-14	32 / 182 (17.6%)	14.63	1.73
10	CSSR	8.64E-16	7.08E-14	43 / 182 (23.6%)	7.63	2.44
11	PXC	1.12E-15	9.19E-14	100 / 182 (54.9%)	3.03	1.67
12	WXSXC	1.31E-15	1.07E-13	53 / 182 (29.1%)	5.52	0.69
13	SXC	2.92E-15	2.39E-13	97 / 182 (53.3%)	3.03	1.67
14	CGXG	9.01E-15	7.39E-13	27 / 182 (14.8%)	18.62	7.61
15	WSDC	2.52E-14	2.07E-12	24 / 182 (13.2%)	24.93	6.06
16	CXN	3.07E-14	2.51E-12	81 / 182 (44.5%)	3.27	2.22
17	VTCGDGV	5.14E-14	4.21E-12	19 / 182 (10.4%)	79.78	6.59
18	SPSGGP	1.43E-13	1.17E-11	24 / 182 (13.2%)	19.95	2.34
19	REKQR	2.57E-13	2.11E-11	18 / 182 (9.9%)	75.79	6.55
20	GCG	2.15E-12	1.76E-10	26 / 182 (14.3%)	11.97	6.36
21	CGGGV	2.20E-12	1.80E-10	21 / 182 (11.5%)	21.94	6.36
22	HGKG	3.35E-12	2.75E-10	56 / 182 (30.8%)	3.85	1.52
23	CXGXG	4.83E-12	3.96E-10	35 / 182 (19.2%)	6.53	5.28
24	SXTXG	9.65E-12	7.91E-10	20 / 182 (11.0%)	20.94	6.18
25	SP	1.38E-11	1.13E-09	39 / 182 (21.4%)	5.32	8.19
26	QRPS	4.69E-11	3.85E-09	20 / 182 (11.0%)	16.75	3.01
27	ATC	4.21E-10	3.45E-08	61 / 182 (33.5%)	3.02	0.09
28	SPXS	1.56E-09	1.28E-07	49 / 182 (26.9%)	3.38	4.02
29	CSXT	2.46E-09	2.02E-07	53 / 182 (29.1%)	3.12	0.52
30	GYCS	3.18E-09	2.61E-07	22 / 182 (12.1%)	8.34	3.38
31	SXTXC	3.18E-09	2.61E-07	22 / 182 (12.1%)	8.34	5.99
32	SPLS	3.59E-09	2.95E-07	26 / 182 (14.3%)	6.34	2.55
33	SXCXS	4.81E-09	3.94E-07	27 / 182 (14.8%)	5.88	5.58
34	CG	7.62E-09	6.25E-07	41 / 182 (22.5%)	3.64	3.23
35	SPT	8.84E-09	7.25E-07	46 / 182 (25.3%)	3.29	4.48
36	APC	1.24E-08	1.01E-06	39 / 182 (21.4%)	3.71	3.89
37	DKPRR	3.45E-08	2.83E-06	24 / 182 (13.2%)	5.87	0
38	YCNINZ	5.91E-08	4.85E-06	14 / 182 (7.7%)	14.96	2.28
39	PXW	1.20E-07	9.84E-06	29 / 182 (15.9%)	4.27	3.81
40	TXC	2.37E-07	1.94E-05	37 / 182 (20.3%)	3.3	3.14
41	GVQTR	3.69E-07	3.03E-05	16 / 182 (8.8%)	8.48	5.33
42	PSDK	1.04E-06	8.52E-05	22 / 182 (12.1%)	4.83	2.81
43	QKR	1.25E-06	1.03E-04	41 / 182 (22.5%)	2.79	4.43
44	RIST	1.39E-06	1.14E-04	49 / 182 (26.9%)	2.49	0.93
45	CXSXT	1.57E-06	1.29E-04	44 / 182 (24.2%)	2.64	1.2
46	RRP	3.26E-06	2.67E-04	12 / 182 (6.6%)	10.37	6.7
47	RT	3.41E-06	2.80E-04	35 / 182 (19.2%)	2.93	2.39
48	RTS	4.30E-06	3.53E-04	16 / 182 (8.8%)	6.16	5.43
49	SXPXS	4.48E-06	3.67E-04	10 / 182 (5.5%)	14.63	7.74
50	ETWRTE	1.59E-05	1.30E-03	10 / 182 (5.5%)	10.97	2.91
51	KNWP	2.04E-05	1.67E-03	41 / 182 (22.5%)	2.39	0.89
52	SLD	6.40E-05	5.25E-03	13 / 182 (7.1%)	5.58	5.21
53	RXGRGT	8.29E-05	6.80E-03	15 / 182 (8.2%)	4.56	3.89

Identifying novel anti-angiogenic peptide motifs through computational tools provides a solid foundation for developing more precise and targeted cancer therapies. These peptides can selectively target angiogenic pathways that promote tumor growth, potentially limiting vascularization and nutrient delivery to cancer cells. Diversifying anti-angiogenic peptide motifs could help overcome the drug resistance observed in some cancer treatments. By targeting multiple checkpoints in angiogenesis, these peptides could be less likely to generate resistance from cancer cells. The ability to identify and use different motifs of anti-angiogenic peptides may pave the way for more personalized treatments tailored to the specific characteristics of each cancer type and patient.

A comparison of the motifs discovered and validated with the scientific literature was carried out. Among them, the RTS motif stands out with an E-value of 0.000353, being one of the highest statistically among the validated ones, which positions it as one of the most significant anti-angiogenic motifs.

RTS disintegrins, present in snake venom, have demonstrated their ability to specifically disrupt the interaction between integrin $\alpha1\beta1$ and collagens IV and I in vitro, leading to inhibition of angiogenesis in vivo. Extensive research on these snake venom disintegrins reveals their selectivity to block $\alpha1\beta1$ integrin function in laboratory settings and living organisms, making them promising candidates for antiangiogenic drugs [130].

The therapeutic potential of snake venom disintegrins, including RTS, has been highlighted in the blockade of specific integrin receptors involved in tumor neovascularization. Selective inhibition of $\alpha1\beta1$ integrin function by integrin-targeted disintegrins such as $\alpha5\beta1$, $\alpha v\beta5$ and $\alpha v\beta3$, all containing the RGD motif, has been studied. The use of synthetic peptides and disintegrins, together with recombinant $\alpha1\beta1$ integrin-specific ligands, such as jerdostatin, opens new possibilities to explore KTS/RTS disintegrins in various applications [131]. The importance of these findings is confirmed when considering alternative cancer therapies with antiangiogenic peptides.

In addition, another important motif is CG with an E-value of 0.000000261. This motif is present in DNA rich in unmethylated CpG motifs and plays a crucial role in facilitating the induction of immune responses against co-administered antigens. CpG motifs, being free of methylation, are recognized by the immune system and trigger robust responses against foreign agents. For this reason, CpGs are considered among the most promising adjuvants to date for enhancing the efficacy of immunotherapeutic interventions

and vaccine development. Their ability to stimulate significant immune responses makes them crucial components in formulating therapeutic and preventive strategies against infectious diseases and other immune-related disorders [132].

On the other hand, the WSPCSV motif (E-value: 8.52E-20)has emerged as a prominent element in the analysis of the humoral immune response to TRAP (Thrombospondin-Related Protein). It is also one of the motifs enriched in this section. This motif is shared between TRAP and thrombospondin and other proteins that exhibit cell adhesion properties, such as propidin. In addition, the presence of these motifs has been observed in proteins of sporozoites of different Plasmodium species and in the cytoplasmic protein HL6 of the parasite Eimenriatenella [133].

The remarkable conservation of the WSPCSV motif in the Trap gene of several TRAP isolates suggests that it plays a crucial functional role [133]. Given its presence in proteins associated with cell adhesion and its conservation across diverse species, including parasites, it is likely that this motif is involved in cell interaction and other biological activities essential for the function and pathogenicity of these organisms [133].

The PXW motif, enriched by our method with an E-value of 0.00000984 is present in the C-terminal domain (HP35) of chicken villin, highlighted by the sequence Pro62-Xxx63-Trp64-Lys65. It has been observed that the arrangement of Trp64 on the side chains of Pro62 and Lys65 is essential for the interaction with F-actin and the structural stability of HP. The conservation of this sequence suggests its importance in the proper folding of the C-terminal domain, with Pro62 as a crucial requirement. These Pro-Trp interactions are fundamental to the structure and function of chicken villin [134].

The SPXS motif with a E-value of 0.000000128, is present in the intracellular region of LRP6/5, is critical for Wnt/b-catenin signaling. Phosphorylation of the PPPSPxS motifs in this region inhibits GSK3b, a crucial step in the Wnt signaling pathway. Synthetic peptides containing this motif inhibit GSK3b in vitro when phosphorylated. The intracellular region of LRP6/5 acts as a direct inhibitor of GSK3b by recruiting and inhibiting it, suggesting a novel activation mechanism in this signaling pathway. Experiments with LRP6/5 mutants confirm the importance of the SPxS motif in Wnt/b-catenin signal transduction. In summary, the SPxS motif in the intracellular region of LRP6/5 is crucial in regulating the Wnt/b-catenin signaling pathway by specifically inhibiting GSK3b activity[135].

The 53 motifs identified mark a significant advance in preparing future pharmacological studies focused on peptides with antiangiogenic properties. However, it is worth noting that motifs with a higher E-value are statistically more significant. Among them are: PSDK, QKR, RIST, CXSXT, RRP, RT, RTS, SXPXS, ETWRTE, KNWP, SLD, and RXGRGT. These motifs, having a higher E-value, suggest a higher probability of random sequence occurrence, potentially making them more statistically and biologically relevant for future research and pharmacological applications in angiogenesis inhibition.

Chapter 5

Conclusions

5.1 Conclusions

The research has provided insightful knowledge that enriches our understanding of antiangiogenic peptides by creatiating of METNs. Special attention has been paid to the main databases from which these peptides originate, SATDPdb being the most prominent. In addition, an exhaustive evaluation of their functionalities has been carried out, highlighting their antiangiogenic activity and their connection with antitumor and anticancer functions. Likewise, the origin of these peptides has been studied in depth, observing that most of them come from synthetic constructs.

Effective chemical space representation of the 209 stored StarPep DB AAPs was achieved by constructing HSPNs. Sixteen HSPNs were generated for each metric, resulting in a complete set of 96 HSPNs. For ease of visualization and analysis, two HSPNs per metric were selected: one corresponding to the optimal cutoff point and one without a cutoff point, totaling 12 visualizations. These provide a representative and detailed perspective of the chemical space explored by the AAPs. For a more accurate identification of representative peptides, the use of HB centrality is recommended. After analysis of the global network parameters and scaffold analysis with Dover Analyzer, it was determined that the *As*, *Ch* and *Eu* metrics best represent the set of AAPs. Furthermore, it was highlighted that, in terms of representation, there are no significant differences between the optimal cut-off points and no cut-off points for each metric, evidencing the similarity in the information provided by both.

A thorough analysis of structural and physicochemical descriptors of *As*, *Ch*, and *Eu* metrics highlighted distinctive antiangiogenic peptides (AAPs) patterns, providing detailed insight into the biochemical diversity among clusters and their potential relevance in therapeutic applications.

Using the most representative HSPNs allowed the discovery and enrichment of motifs in antiangiogenic peptides (AAPs). Through the STREME algorithm, 74 unique motifs were identified, which were complemented by 32 motifs previously reported in the lit-

erature and validated and enriched through the SEA algorithm, resulting in 82 motifs initially enriched using the AntiangioPred and BIG_ANTIAN_DB databases. Ultimately, through reverse validation, 53 potential antiangiogenic motifs are reported. These identified motifs provide valuable information, contributing significantly to the understanding and discovery of AAPs, and establishing solid foundations for alternative cancer therapies by inhibiting angiogenesis in tumors.

It is suggested that further research should delve deeper into the relationship between the identified motifs and the three-dimensional structure of peptides, using techniques such as docking or molecular dynamics simulation. This would allow a more complete understanding of how motifs fold and adapt in specific biological contexts, thus improving the accuracy of biological activity predictions and facilitating the rational design of new peptides with therapeutic applications in cancer and other angiogenesis-related diseases.

References

(1) Torre, L. A.; Bray, F.; Siegel, R. L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A. *CA Cancer J Clin* **2015**, *65*, 87–108.

(2) Zugazagoitia, J.; Guedes, C.; Ponce, S.; Ferrer, I.; Molina-Pinelo, S.; Paz-Ares, L. *Clinical Therapeutics* **2016**, *38*, 1551–1566.

(3) Hoppe, C.; Buntzel, J.; von Weikersthal, L. F.; Junghans, C.; Zomorodbakhsch, B.; Stoll, C.; Prott, F. J.; Fuxius, S.; Micke, O.; Richter, A.; Sallmann, D.; Hubner, J. *In Vivo* **2023**, *37*, 106–114.

(4) Ramadhani, D.; Maharani, R.; Gazzali, A. M.; Muchtaridi, M. *Molecules* **2022**, *27*, DOI: 10.3390/molecules27144428.

(5) Charoenkwan, P.; Chiangjong, W.; Lee, V. S.; Nantasenamat, C.; Hasan, M. M.; Shoombuatong, W. *Scientific Reports* **2021**, *11*, 1–13.

(6) Marqus, S.; Pirogova, E.; Piva, T. J. *Journal of Biomedical Science 2017 24:1* **2017**, *24*, 1–15.

(7) Shoari, A.; Khodabakhsh, F.; Ahangari Cohan, R.; Salimian, M.; Karami, E. *Research in Pharmaceutical Sciences* **2021**, *16*, 559–574.

(8) Hanahan, D.; Weinberg, R. A. *Cell* **2000**, *100*, 57–70.

(9) Fouad, Y. A.; Aanei, C. *American Journal of Cancer Research* **2017**, *7*, 1016.

(10) Ramjiawan, R. R.; Griffioen, A. W.; Duda, D. G. *Angiogenesis* **2017**, *20*, 185–204.

(11) Vafopoulou, P.; Kourti, M. *Journal of Cancer Metastasis and Treatment* **2022**, *8*, DOI: 10.20517/2394-4722.2022.08.

(12) Ferrara, N. *Arteriosclerosis, Thrombosis, and Vascular Biology* **2009**, *29*, 789–791.

(13) Montesano, R.; Vassalli, J. D.; Baird, A.; Guillemin, R.; Orci, L. *Proceedings of the National Academy of Sciences of the United States of America* **1986**, *83*, 7297–7301.

(14) Li, S.; Hu, G. F. *International Journal of Biochemistry and Molecular Biology* **2010**, *1*, 26–35.

(15) Elliott, R. L.; Blobe, G. C. *Journal of Clinical Oncology* **2005**, *23*, 2078–2093.

(16) Chu, W. M. *Cancer Letters* **2013**, *328*, 222–225.

(17) Farooqi, A. A.; Siddik, Z. H. *Cell Biochemistry and Function* **2015**, *33*, 257–265.

(18) Dewerchin, M.; Carmeliet, P. *Expert Opinion on Therapeutic Targets* **2014**, *18*, 1339–1354.

(19) Zhang, N.; Wang, L.; Liang, Y.; Zhao, Y. M.; Xue, Q.; Wu, W. Z.; Sun, H. C.; Fan, J.; Tang, Z. Y. *Journal of Cancer Research and Clinical Oncology* **2009**, *135*, 1447–1453.

(20) Kaga, T.; Kawano, H.; Sakaguchi, M.; Nakazawa, T.; Taniyama, Y.; Morishita, R. *Vascular Pharmacology* **2012**, *57*, 3–9.

(21) Pache, J. C. Epidermal Growth Factors, 2006.

(22) Oguntade, A. S.; Al-Amodi, F.; Alrumayh, A.; Alobaida, M.; Bwalya, M. *Journal of the Egyptian National Cancer Institute* **2021**, *33*, DOI: 10.1186/s43046-021-00072-6.

(23) V. Rosca, E.; E. Koskimaki, J.; G. Rivera, C.; B. Pandey, N.; P. Tamiz, A.; S. Popel, A. *Current Pharmaceutical Biotechnology* **2011**, *12*, 1101–1116.

(24) Li, X.; Cai, H.; Wu, X.; Li, L.; Wu, H.; Tian, R. *Frontiers in Chemistry* **2020**, *8*, 1–19.

(25) Ramaprasad, A. S. E.; Singh, S.; Gajendra, R. P.; Venkatesan, S. *PLoS ONE* **2015**, *10*, 7–12.

(26) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Beltran, J. A.; Tellez Ibarra, R.; Guillen-Ramirez, H. A.; Brizuela, C. A. *Bioinformatics* **2019**, *35*, ed. by Wren, J., 4739–4747.

(27) Mahesh, B. *International Journal of Science and Research* **2020**, *9*, 381–386.

(28) Medina-Franco, J. L.; Sánchez-Cruz, N.; López-López, E.; Díaz-Eufracio, B. I. *Journal of Computer-Aided Molecular Design* **2022**, *36*, 341–354.

(29) Capecchi, A.; Reymond, J. L. *Medicine in Drug Discovery* **2021**, *9*, 100081.

(30) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K. R.; Anatole Von Lilienfeld, O. *New Journal of Physics* **2013**, *15*, 0–16.

(31) Maggiora, G. M.; Bajorath, J. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 795–802.

(32) Reymond, J. L. *Accounts of Chemical Research* **2015**, *48*, 722–730.

(33) Tran, T. D.; Ogbourne, S. M.; Brooks, P. R.; Sánchez-Cruz, N.; Medina-Franco, J. L.; Quinn, R. J. *International Journal of Molecular Sciences* **2020**, *21*, 1–35.

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

(35) Wang, F.; Jin, R.; Agrawal, G.; Piontkivska, H. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE* **2007**, 1318–1322.

(36) Aguilera-mendoza, L.; Ayala-ruano, S.; Martinez-rios, F.; Chavez, E.; Jacas, C. R. G.; Brizuela, C. A.; Marrero-ponce, Y. **2023**, 2–5.

(37) Zwierzyna, M.; Vogt, M.; Maggiora, G. M.; Bajorath, J. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 113–125.

(38) Ayala-Ruano, S.; Marrero-Ponce, Y.; Aguilera-Mendoza, L.; Pérez, N.; Agüero-Chapin, G.; Antunes, A.; Aguilar, A. C. *ACS Omega* **2022**, *7*, 46012–46036.

(39) Castillo-mendieta, K.; Agüero-chapin, G.; Marquez, E. A.; Perez-castillo, Y.; Stephen, J **2023**, DOI: 10.20944/preprints202303.0322.v1.

(40) Bailey, T. L.; Grant, C. E. *bioRxiv* **2021**, 2021.08.23.457422.

(41) Kunda, N. K. *Drug Discovery Today* **2020**, *25*, 238–247.

(42) Siegel, R. L.; Miller, K. D.; Jemal, A. *CA: A Cancer Journal for Clinicians* **2018**, *68*, 7–30.

(43) Huang, W.; Seo, J.; Willingham, S. B.; Czyzewski, A. M.; Gonzalgo, M. L.; Weissman, I. L.; Barron, A. E. *PLoS ONE* **2014**, *9*, ed. by Afarinkia, K., e90397.

(44) Hosseinzadeh, E.; Banaee, N.; Nedaie, H. A. *Current Cancer Therapy Reviews* **2017**, *13*, 17–27.

(45) Amit, D.; Hochberg, A. *Journal of Translational Medicine* **2010**, *8*, 134.

(46) Craik, D. J.; Fairlie, D. P.; Liras, S.; Price, D. *Chem Biol Drug Des* **2013**, *81*, 136–47.

(47) Carmeliet, P. *Nature 2005 438:7070* **2005**, *438*, 932–936.

(48) Vicari, D.; Foy, K. C.; Liotta, E. M.; Kaumaya, P. T. *Journal of Biological Chemistry* **2011**, *286*, 13612–13625.

(49) Nakamura, T.; Matsumoto, K. *Biochemical and Biophysical Research Communications* **2005**, *333*, 289–291.

(50) Zahiri, J.; Khorsand, B.; Yousefi, A. A.; Kargar, M.; Shirali Hossein Zade, R.; Mahdevar, G. *Journal of Translational Medicine* **2019**, *17*, 1–6.

(51) Basith, S.; Manavalan, B.; Hwan Shin, T.; Lee, G. *Medicinal Research Reviews* **2020**, *40*, 1276–1314.

(52) Zhang, X.-D., *Chapter 6 Machine Learning*; 13, 2017; Vol. 45, pp 40–48.

(53) Zhou, H.; Li, S.; Zeng, X.; Zhang, M.; Tang, L.; Li, Q.; Chen, D.; Meng, X.; Hong, X. *Chinese Chemical Letters* **2020**, *31*, 1382–1386.

(54) Zhang, W.; Yu, L.; Ji, T.; Wang, C. *Frontiers in Chemistry* **2020**, *8*, 1–7.

(55) Attique, M.; Farooq, M. S.; Khelifi, A.; Abid, A. *IEEE Access* **2020**, *8*, 148570–148594.

(56) Joseph, S.; Karnik, S.; Nilawe, P.; Jayaraman, V. K.; Idicula-Thomas, S. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2012**, *9*, 1535–1538.

(57) Wei, G.; Wang, Y.; Huang, X.; Hou, H.; Zhou, S. *Small Methods* **2018**, *2*, 1–16.

(58) Dobson, C. M. *Nature* **2004**, *432*, 824–828.

(59) Kitano, H. *Nature* **2002**, *420*, 206–210.

(60) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Advanced Drug Delivery Reviews* **1997**, *23*, 3–25.

(61) Li, W.; Tan, S.; Xing, Y.; Liu, Q.; Li, S.; Chen, Q.; Yu, M.; Wang, F.; Hong, Z. *Molecular Pharmaceutics* **2018**, *15*, 1505–1514.

(62) Talamantes, A.; Chavez, E. *Pattern Recognition Letters* **2022**, *156*, 88–95.

(63) Zahoránszky-Kohalmi, G.; Bologa, C. G.; Oprea, T. I. *Journal of Cheminformatics* **2016**, *8*, 1–17.

(64) Newman, M., *Networks*; 1; Oxford University Press: 2010; Vol. 15, pp 583–605.

(65) Newman, M. E. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **2004**, *70*, 9.

(66) Newman, M. E. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 8577–8582.

(67) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, *2008*, P10008.

(68) Lü, L.; Chen, D.; Ren, X. L.; Zhang, Q. M.; Zhang, Y. C.; Zhou, T. *Physics Reports* **2016**, *650*, 1–63.

(69) Boldi, P.; Vigna, S. *Internet Mathematics* **2014**, *10*, 222–262.

(70) Csermely, P.; Korcsmáros, T.; Kiss, H. J.; London, G.; Nussinov, R. *Pharmacology and Therapeutics* **2013**, *138*, 333–408.

(71) Pfeiffer, J. J.; Neville, J. **2011**.

(72) Kesilmiş, Z. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* **2022**, *44*, 469–492.

(73) Bajusz, D.; Rácz, A.; Héberger, K. *Journal of cheminformatics* **2015**, *7*, 1–13.

(74) Klove, T.; Lin, T.-T.; Tsai, S.-C.; Tzeng, W.-G. *IEEE Transactions on Information Theory* **2010**, *56*, 2611–2617.

(75) Aherne, F. J.; Thacker, N. A.; Rockett, P. I. *Kybernetika* **1998**, *34*, 363–368.

(76) Lau, B. K.; Dillon, O.; Vinod, S. K.; O'Brien, R. T.; Reynolds, T. *Medical Physics* **2023**.

(77) Atsushi, I. *Journal of Biochemistry* **1980**, *88*, 1895–1898.

(78) Consonni, V.; Todeschini, R., *Molecular descriptors*, 2010; Vol. 8, pp 29–102.

(79) Tyunina, E. Y.; Badelin, V. G. *Russian Journal of Bioorganic Chemistry* **2009**, *35*, 453–460.

(80) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. *Journal of Medicinal Chemistry* **1998**, *41*, 2481–2491.

(81) Of Astrobiology, E., *Isoelectric Point*, 2011.

(82) Boman, H. G. *Journal of Internal Medicine* **2003**, *254*, 197–215.

(83) Cid, H.; Bunster, M.; Canales, M.; Gazitúa, F. *Protein Engineering, Design and Selection* **1992**, *5*, 373–375.

(84) *Computational and Structural Biotechnology Journal* **2023**, *21*, 3234–3247.

(85) Kyte, J.; Doolittle, R. F. *Journal of Molecular Biology* **1982**, *157*, 105–132.

(86) *Nucleic Acids Research* **2014**, *42*, 191–198.

(87) Bastian, M.; Heymann, S.; Jacomy, M *International AAAI Conference on Weblogs and Social Media* **2009**, 361–362.

(88) Bailey, T. L.; Johnson, J.; Grant, C. E.; Noble, W. S. *Nucleic Acids Research* **2015**, *43*, W39–W49.

(89) Bailey, T. L. *Bioinformatics* **2021**, *37*, ed. by Birol, I., 2834–2840.

(90) Peel, L.; Larremore, D. B.; Clauset, A. *Science Advances* **2017**, *3*, DOI: 10.1126/sciadv.1602548.

(91) Diestel, R., *Graph theory*, Fifth Edit; 9783319530031; Springer Berlin Heidelberg: 2017, pp 49–64.

(92) Brandes, U. *Journal of Mathematical Sociology* **2010**, 37–41.

(93) Gibson, H.; Vickers, P. **2017**.

(94) Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. *PLoS ONE* **2014**, *9*, 1–12.

(95) Chavez, E.; Dobrev, S.; Kranakis, E.; Opatrny, J.; Stacho, L.; Tejeda, H.; Urnitia, J. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2006**, *3974 LNCS*, 235–245.

(96) Henikoff, S.; Henikoff, J. G. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89*, 10915–10919.

(97) Godden, J. W.; Stahura, F. L.; Bajorath, J. *Journal of Chemical Information and Computer Sciences* **2000**, *40*, 796–800.

(98) Schober, P.; Schwarte, L. A. *Anesthesia and Analgesia* **2018**, *126*, 1763–1768.

(99) Blöcker, C.; Nieves, J. C.; Rosvall, M. *Applied Network Science* **2022**, *7*, DOI: 10.1007/s41109-022-00477-9.

(100) Shimagaki, K.; Barton, J. P. **2023**, *107*, 1–14.

(101) Fruchterman, T. M. J.; Reingold, E. M. *Force-Directed Placement, in: Software-Practice and Experience* **1991**, *21*, 1129–1164.

(102) Ortega, J. M. E.; Eballe, R. G. *Advances and Applications in Discrete Mathematics* **2022**, *31*, 13–33.

(103) Needleman, S. B.; Wunsch, C. D. *Journal of Molecular Biology* **1970**, *48*, 443–453.

(104) Xia, Z.; Cui, Y.; Zhang, A.; Tang, T.; Peng, L.; Huang, C.; Yang, C.; Liao, X. *Interdisciplinary Sciences – Computational Life Sciences* **2022**, *14*, 1–14.

(105) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Tellez-Ibarra, R.; Llorente-Quesada, M. T.; Salgado, J.; Barigye, S. J.; Liu, J. *Bioinformatics* **2015**, *31*, 2553–2559.

(106) Koivunen, E.; Arap, W.; Rajotte, D.; Lahdenranta, J.; Pasqualini, R. *Journal of Nuclear Medicine* **1999**, *40*, 883–888.

(107) Robles, J. P.; Zamora, M.; Siqueiros-Marquez, L.; Adan-Castro, E.; Ramirez-Hernandez, G.; Nuñez, F. F.; Lopez-Casillas, F.; Millar, R. P.; Bertsch, T.; Martínez de la Escalera, G.; Triebel, J.; Clapp, C. *Angiogenesis* **2022**, *25*, 57–70.

(108) Patel, N. S.; Quadros, A.; Brem, S.; Wotoczek-Obadia, M.; Mathura, V. S.; Laporte, V.; Mullan, M.; Paris, D. *Amyloid* **2008**, *15*, 5–19.

(109) Son, H. N.; Nam, J. O.; Kim, S.; Kim, I. S. *Biochimica et Biophysica Acta - Molecular Cell Research* **2013**, *1833*, 2378–2388.

(110) Ma, D.; Gao, L.; An, S.; Song, Y.; Wu, J.; Xu, X.; Lai, R. *Toxicon* **2010**, *55*, 45–51.

(111) Nam, J. O.; Kim, J. E.; Jeong, H. W.; Lee, S. J.; Lee, B. H.; Choi, J. Y.; Park, R. W.; Park, J. Y.; Kim, I. S. *Journal of Biological Chemistry* **2003**, *278*, 25902–25909.

(112) Meng, J.; Nan, M.; Yan, Z.; Han, W.; Zhang, Y. *Journal of Biochemistry* **2006**, *140*, 299–304.

(113) Kumar, D.; Patel, S. A.; Khan, R.; Chawla, S.; Mohapatra, N.; Dixit, M. *Molecular Cancer Research* **2022**, *20*, 77–91.

(114) Cao, X. W.; Yang, X. Z.; Du, X.; Fu, L. Y.; Zhang, T. Z.; Shan, H. W.; Zhao, J.; Wang, F. J. *Journal of Drug Targeting* **2018**, *26*, 777–792.

(115) Singh, S.; Chaudhary, K.; Dhanda, S. K.; Bhalla, S.; Usmani, S. S.; Gautam, A.; Tuknait, A.; Agrawal, P.; Mathur, D.; Raghava, G. P. *Nucleic Acids Research* **2015**, *44*, D1119–D1126.

(116) Tyagi, A.; Tuknait, A.; Anand, P.; Gupta, S.; Sharma, M.; Mathur, D.; Joshi, A.; Singh, S.; Gautam, A.; Raghava, G. P. *Nucleic Acids Research* **2015**, *43*, D837–D843.

(117) Thundimadathil, J. *Journal of Amino Acids* **2012**, *2012*, 1–13.

(118) Karagiannis, E. D.; S. Popel, A. *Cmes-Computer Modeling in Engineering Sciences* **2007**, *1*, 119–131.

(119) Karagiannis, E. D.; Popel, A. S. *International Journal of Biochemistry and Cell Biology* **2007**, *39*, 2314–2323.

(120) Liang, P.; Ballou, B.; Lv, X.; Si, W.; Bruchez, M. P.; Huang, W.; Dong, X. *Advanced Materials* **2021**, *33*, 1–20.

(121) Folkman Judah Brookline, Javaherian Kashi Lexington, Becker, C. O. ANTIANGIOGENICPEPTIDES FOR TREATING OR PREVENTING ENDOMETRIOSIS, 2008.

(122) Yin, R.; Zheng, H.; Xi, T.; Xu, H. M. *Bioconjugate Chemistry* **2010**, *21*, 1142–1147.

(123) Tripodi, A. A. P.; Ranđelović, I.; Biri-Kovács, B.; Szeder, B.; Mező, G.; Tóvári, J. *Pathology and Oncology Research* **2020**, *26*, 1879–1892.

(124) Zhang, Z.; Ji, S.; Zhang, B.; Liu, J.; Qin, Y.; Xu, J.; Yu, X. *Biomedicine and Pharmacotherapy* **2018**, *108*, 1135–1140.

(125) Annese, T.; Tamma, R.; Ruggieri, S.; Ribatti, D. *Cancers* **2019**, *11*, DOI: 10.3390/cancers11030381.

(126) Vogt, M.; Stumpfe, D.; Maggiora, G. M.; Bajorath, J. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 191–208.

(127) Bailey, T. L. *Bioinformatics* **2021**, DOI: 10.1093/bioinformatics/btab203.

(128) Bailey, T. L. *Bioinformatics* **2011**, *27*, 1653–1659.

(129) Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S. *Nucleic Acids Research* **2009**, *37*, 202–208.

(130) Calvete, J.; Marcinkiewicz, C.; Sanz, L. *Current Pharmaceutical Design* **2007**, *13*, 2853–2859.

(131) Sanz, L.; Chen, R. Q.; Pérez, A.; Hilario, R.; Juárez, P.; Marcinkiewicz, C.; Monleón, D.; Celda, B.; Xiong, Y. L.; Pérez-Payá, E.; Calvete, J. J. *Journal of Biological Chemistry* **2005**, *280*, 40714–40722.

(132) Storni, T.; Ruedl, C.; Schwarz, K.; Schwendener, R. A.; Renner, W. A.; Bachmann, M. F. *The Journal of Immunology* **2004**, *172*, 1777–1785.

(133) *Infection and Immunity* **1993**, *61*, 3490–3495.

(134) Vermeulen, W.; Van Troys, M.; Bourry, D.; Dewitte, D.; Rossenu, S.; Goethals, M.; Borremans, F. A.; Vandekerckhove, J.; Martins, J. C.; Ampe, C. *Journal of Molecular Biology* **2006**, *359*, 1277–1292.

(135) Piao, S.; Lee, S. H.; Kim, H.; Yum, S.; Stamos, J. L.; Xu, Y.; Lee, S. J.; Lee, J.; Oh, S.; Han, J. K.; Park, B. J.; Weis, W. I.; Ha, N. C. *PLoS ONE* **2008**, *3*, DOI: 10.1371/journal.pone.0004046.

Attachments

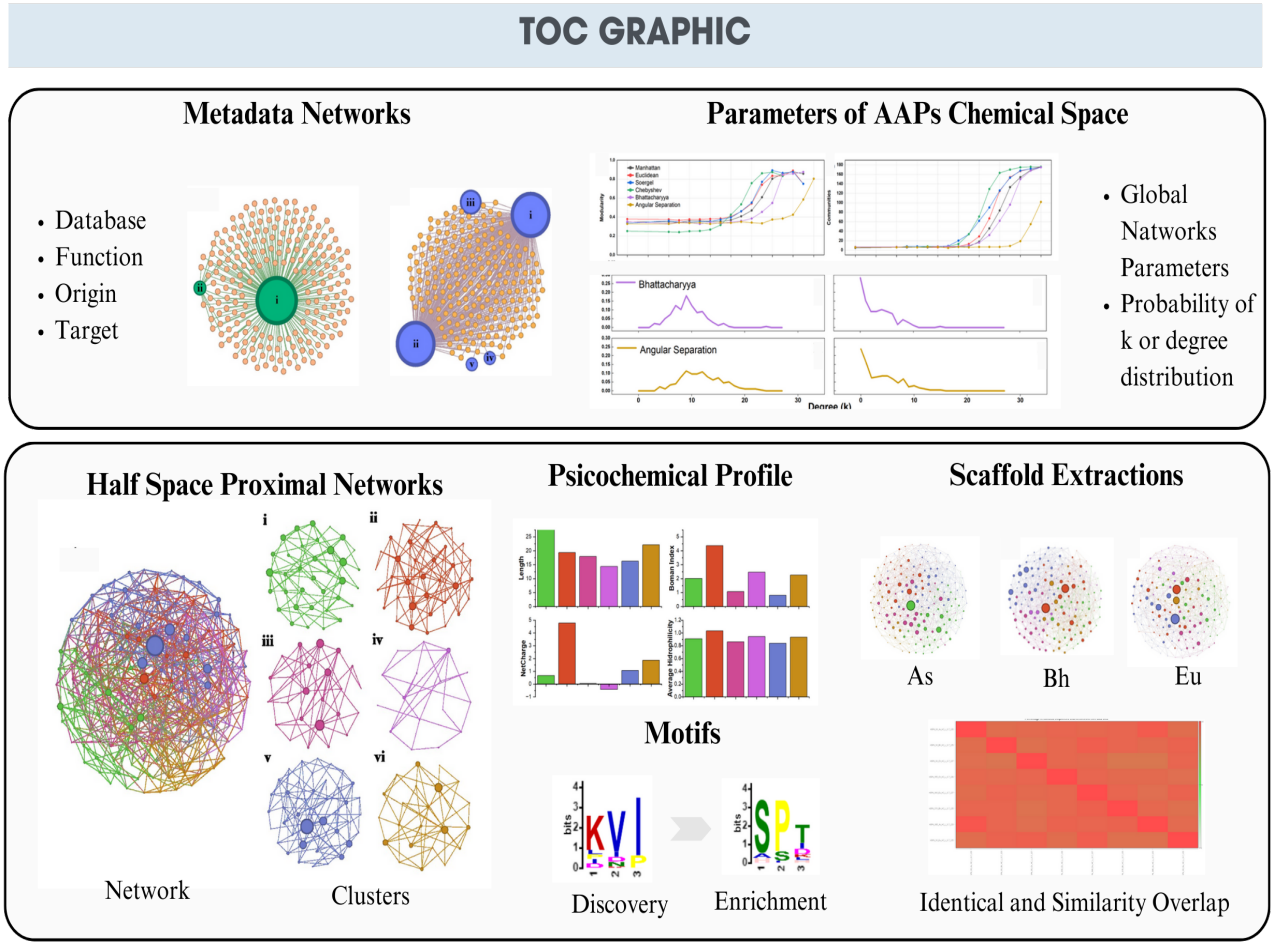

Figure 5.1: Research Overview: Exploring the Antiangiogenic Activity of Peptides for Cancer Treatment: An Analysis Using Visual Data Mining and a Similarity Search Based Approach.

Table 5.1: Motifs discovered by STREME for Chevyshev metric.

HSPN_00_Ch_								
No.	**Motif**	**Cluster**	**Cluster Size**	**Matches (+) Seq.**	**Matches (-) Seq.**	**Sites(%)**	**Score**	**Match Threshold**
1	LKKF	1	30	10	0	33.3	0.0004	7.84538
2	STRI			8	0	26.7	0.0023	10.7027
3	WSSCS			8	0	26.7	0.0023	5.06424
4	TCGDGV			6	0	20	0.012	1.71709
5	YCNINZ			3	0	10	0.12	0.70716
6	QKR	2	23	11	0	47.8	0.0001	5.96461
7	ING			7	0	30.4	0.0046	7.97888
8	PWSQCS			7	0	30.4	0.0046	1.13462
9	RCGRGT			5	0	21.7	0.025	2.54819
10	ADRAA			4	0	17.4	0.054	1.71699
11	KGRG	3	20	14	0	70	0.0000017	1.89673
12	PSDK			14	0	70	0.0000017	7.20803
13	EKQR			10	0	50	0.00022	8.23836
14	KRRR			10	1	50	0.0042	7.49419
15	RRP			7	1	35	0.046	6.93898
16	SPW	4	25	22	2	88	0.00000018	5.91394
17	VTCG			14	0	56	0.0000048	8.62409
18	GVQTR			13	0	52	0.000015	8.25228
19	QCS			12	1	48	0.00096	8.11015
20	APC			7	1	28	0.049	7.97233
21	APF	5	39	13	0	33.3	0.000037	7.18049
22	GRELCL			11	0	28.2	0.00022	10.1337
23	IIEK			11	0	28.2	0.00022	7.27161
24	VQK			10	0	25.6	0.00051	8.04087
25	NPASP			5	0	12.8	0.027	8.45514
26	FGN	6	10	3	0	30	0.1100	4.23788
27	ISNA			3	0	30	0.1100	12.1092
28	GCG			2	0	20	0.2400	9.01765
29	GYCS			2	0	20	0.2400	11.0779
30	HGKG			2	0	20	0.2400	13.7461
31	GAR	7	18	7	0	38.9	0.0038	3.30461
32	ATC			6	0	33.3	0.0095	8.37408
33	LHLV			5	0	27.8	0.023	10.3365
34	EGL			7	1	38.9	0.044	7.15219
35	SPSGGP			4	0	22.2	0.052	2.81781
36	SPLS	8	11	3	0	27.3	1.10E-01	8.03363
37	TGKI			3	1	27.3	1.10E-01	1.86612
38	HPH			5	1	45.5	1.60E-01	3.68463
39	PEN			5	0	45.5	1.60E-01	9.29149
40	DDDD			1	0	9.1	5.00E-01	14.1391

Table 5.2: Motifs discovered by STREME for Euclidean metric.

HSPN_00_Eu								
No.	**Motif**	**Cluster**	**Cluster Size**	**Matches (+) Seq.**	**Matches (-)Seq.**	**Sites(%)**	**Score**	**Match Threshold**
1	SPL	1	24	12	0	50	3.90E-05	7.52098
2	CGGGV			7	0	29.2	4.70E-03	3.73733
3	CSSR			6	0	25	1.10E-02	0.934526
4	GPWERC			5	0	20.8	2.50E-02	1.48061
5	ETWRTE			4	0	16.7	5.50E-02	3.25366
6	CGRG	2	35	19	0	54.3	6.40E-08	8.87225
7	WSPCSV			16	0	45.7	1.60E-06	11.5784
8	QRPS			14	0	40	1.20E-05	6.78501
9	DKPRR			8	0	22.9	2.50E-03	7.82051
10	KRRREK			8	0	22.9	2.50E-03	5.47961
11	CGG	3	43	14	0	32.6	1.70E-05	5.84324
12	WSPCS			14	0	32.6	1.70E-05	8.14282
13	RTR			11	0	25.6	2.40E-04	7.64453
14	PFL			14	1	32.6	3.60E-04	6.80863
15	RDI			5	0	11.6	2.80E-02	9.7223
16	TGAL	4	35	10	0	28.6	4.60E-04	1.2123
17	EEDP			6	0	17.1	1.20E-02	9.61524
18	HRI			6	0	17.1	1.20E-02	9.06642
19	ASP			5	0	14.3	2.70E-02	8.83988
20	KNWP			5	0	14.3	2.70E-02	10.1901
21	YCNI	5	16	9	0	56.3	4.10E-04	7.26261
22	LKK			8	0	50	1.20E-03	6.99532
23	TVT			6	0	37.5	8.80E-03	7.25726
24	FST			3	0	18.8	1.10E-01	11.2518
25	IKRY			3	0	18.8	1.10E-01	1.00024
26	VQKI	6	23	13	0	56.5	1.10E-05	7.71998
27	CLD			12	0	52.2	3.50E-05	9.3464
28	NGRE			9	0	39.1	7.40E-04	10.9047
29	PEAP			9	0	39.1	7.40E-04	8.66254
30	VRRA			7	0	30.4	4.60E-03	8.96213

Table 5.3: Motifs discovered by STREME for Angular Separation metric.

HSPN_00_As								
No.	**Motif**	**Cluster**	**Cluster Size**	**Matches (+) Seq.**	**Matches (-) Seq.**	**Sites(%)**	**Score**	**Match Threshold**
1	SPT	1	31	14	0	45.2	0.0000091	6.26986
2	KVI			14	1	45.2	0.00023	6.50308
3	KEICLD			9	0	29	0.00099	6.84349
4	SGG			9	0	29	0.00099	5.2302
5	PEAPF			5	0	16.1	0.026	5.42949
6	LAS	2	17	11	1	64.7	0.0012	6.23357
7	SLD			7	0	41.2	0.0036	6.85265
8	IVL			5	0	29.4	0.022	4.74948
9	TFE			5	0	29.4	0.022	11.0156
10	GRK			7	1	41.2	0.043	7.24671
11	WSDC	3	29	19	0	65.5	2.1E-08	6.96496
12	CGRG			18	0	62.1	7.7E-08	7.26064
13	REKQR			11	0	37.9	0.00015	7.03206
14	PRR			12	3	41.4	0.051	7.71382
15	QKR			4	0	13.8	0.056	9.55798
16	SPWE	4	51	19	0	37.3	0.22	5.02883
17	CSASCG			19	0	37.3	0.5	1.05052
18	GVQTR			15	0	29.4	0.5	8.32856
19	HGHK	5	16	6	0	37.5	0.0088	8.86924
20	QCSV			6	0	37.5	0.0088	9.73766
21	NPR			4	0	25	0.051	9.79705
22	KIKSST			2	0	12.5	0.24	21.0446
23	RSRQVR			2	0	12.5	0.24	18.7546
24	CNIN	6	32	8	0	25	0.0024	10.2809
25	EVCH			8	0	25	0.0024	11.1098
26	MPF			8	0	25	0.0024	10.1288
27	RIST			6	0	18.8	0.012	4.82639
28	NDYSYW			4	0	12.5	0.057	17.5945

Table 5.4: Motif enrichment by SEA, first stage (first part)

E-values ≤ 10								
No.	**Motif**	**P-Value**	**E-value**	**Q-Value**	**TP**	**FP**	**Enrichment Ratio**	**Score Threshold**
1	TC	6.52E-23	6.85E-21	1.15E-21	63 / 182 (34.6%)	31 / 729 (4.3%)	7.98	2.85
2	WSPCSV	1.04E-21	1.09E-19	9.15E-21	30 / 182 (16.5%)	0 / 729 (0.0%)	123.66	5.75
3	PWSQCSV	5.2E-21	5.46E-19	3.05E-20	29 / 182 (15.9%)	0 / 729 (0.0%)	119.67	3.14
4	WSSCS	2.6E-20	2.73E-18	1.15E-19	28 / 182 (15.4%)	0 / 729 (0.0%)	115.68	9.34
5	CSASCG	1.63E-17	1.72E-15	5.76E-17	24 / 182 (13.2%)	0 / 729 (0.0%)	99.73	12.18
6	RTR	4.23E-17	4.44E-15	1.24E-16	65 / 182 (35.7%)	53 / 729 (7.3%)	4.88	4.2
7	SC	2.92E-16	3.07E-14	7.27E-16	46 / 182 (25.3%)	25 / 729 (3.4%)	7.21	2.85
8	WSXC	3.3E-16	3.47E-14	7.27E-16	24 / 182 (13.2%)	1 / 729 (0.1%)	49.86	9.11
9	CSSR	8.64E-16	9.07E-14	1.69E-15	43 / 182 (23.6%)	22 / 729 (3.0%)	7.63	2.44
10	PXC	1.12E-15	1.18E-13	1.97E-15	100 / 182 (54.9%)	132 / 729 (18.1%)	3.03	1.67
11	WXSXC	1.31E-15	1.37E-13	2.09E-15	53 / 182 (29.1%)	38 / 729 (5.2%)	5.52	0.69
12	SXC	2.92E-15	3.07E-13	4.28E-15	97 / 182 (53.3%)	128 / 729 (17.6%)	3.03	1.67
13	CGXG	9.01E-15	9.46E-13	1.22E-14	27 / 182 (14.8%)	5 / 729 (0.7%)	18.62	7.61
14	WSDC	2.52E-14	2.65E-12	3.18E-14	24 / 182 (13.2%)	3 / 729 (0.4%)	24.93	6.06
15	CXN	3.07E-14	3.22E-12	3.6E-14	81 / 182 (44.5%)	99 / 729 (13.6%)	3.27	2.22
16	VTCGDGV	5.14E-14	5.39E-12	5.65E-14	19 / 182 (10.4%)	0 / 729 (0.0%)	79.78	6.59
17	SPSGGP	1.43E-13	1.5E-11	1.48E-13	24 / 182 (13.2%)	4 / 729 (0.5%)	19.95	2.34
18	REKQR	2.57E-13	2.7E-11	2.52E-13	18 / 182 (9.9%)	0 / 729 (0.0%)	75.79	6.55
19	RXGRGT	1.6E-12	1.68E-10	1.48E-12	41 / 182 (22.5%)	29 / 729 (4.0%)	5.58	0.47
20	GCG	2.15E-12	2.26E-10	1.84E-12	26 / 182 (14.3%)	8 / 729 (1.1%)	11.97	6.36
21	CGGGV	2.2E-12	2.31E-10	1.84E-12	21 / 182 (11.5%)	3 / 729 (0.4%)	21.94	6.36
22	HGKG	3.35E-12	3.52E-10	2.68E-12	56 / 182 (30.8%)	58 / 729 (8.0%)	3.85	1.52
23	XPWERC	3.96E-12	4.16E-10	3.03E-12	18 / 182 (9.9%)	1 / 729 (0.1%)	37.9	4.83
24	CXGXG	4.83E-12	5.07E-10	3.54E-12	35 / 182 (19.2%)	21 / 729 (2.9%)	6.53	5.28
25	SXTXG	9.65E-12	1.01E-09	6.8E-12	20 / 182 (11.0%)	3 / 729 (0.4%)	20.94	6.18
26	SP	1.38E-11	1.44E-09	9.32E-12	39 / 182 (21.4%)	29 / 729 (4.0%)	5.32	8.19
27	QRPS	4.69E-11	4.93E-09	3.06E-11	20 / 182 (11.0%)	4 / 729 (0.5%)	16.75	3.01
28	ATC	4.21E-10	4.42E-08	2.65E-10	61 / 182 (33.5%)	81 / 729 (11.1%)	3.02	0.09
29	SPXS	1.56E-09	1.64E-07	9.47E-10	49 / 182 (26.9%)	58 / 729 (8.0%)	3.38	4.02
30	CSXT	2.46E-09	2.59E-07	1.45E-09	53 / 182 (29.1%)	68 / 729 (9.3%)	3.12	0.52
31	GYCS	3.18E-09	3.34E-07	1.75E-09	22 / 182 (12.1%)	10 / 729 (1.4%)	8.34	3.38
32	SXTXC	3.18E-09	3.34E-07	1.75E-09	22 / 182 (12.1%)	10 / 729 (1.4%)	8.34	5.99
33	SPLS	3.59E-09	3.77E-07	1.92E-09	26 / 182 (14.3%)	16 / 729 (2.2%)	6.34	2.55
34	SXCXS	4.81E-09	5.05E-07	2.49E-09	27 / 182 (14.8%)	18 / 729 (2.5%)	5.88	5.58
35	CG	7.62E-09	0.0000008	3.83E-09	41 / 182 (22.5%)	45 / 729 (6.2%)	3.64	3.23
36	SPT	8.84E-09	9.28E-07	4.32E-09	46 / 182 (25.3%)	56 / 729 (7.7%)	3.29	4.48
37	APC	1.24E-08	0.0000013	5.88E-09	39 / 182 (21.4%)	42 / 729 (5.8%)	3.71	3.89
38	DKPRR	3.45E-08	0.00000362	1.6E-08	24 / 182 (13.2%)	16 / 729 (2.2%)	5.87	0
39	YCNINZ	5.91E-08	0.00000621	2.67E-08	14 / 182 (7.7%)	3 / 729 (0.4%)	14.96	2.28
40	PXW	0.00000012	0.0000126	5.29E-08	29 / 182 (15.9%)	27 / 729 (3.7%)	4.27	3.81
41	TXC	2.37E-07	0.0000249	1.02E-07	37 / 182 (20.3%)	45 / 729 (6.2%)	3.3	3.14
42	GVQTR	3.69E-07	0.0000387	1.55E-07	16 / 182 (8.8%)	7 / 729 (1.0%)	8.48	5.33
43	PSDK	0.00000104	0.000109	4.26E-07	22 / 182 (12.1%)	18 / 729 (2.5%)	4.83	2.81
44	QKR	0.00000125	0.000132	5.02E-07	41 / 182 (22.5%)	59 / 729 (8.1%)	2.79	4.43
45	RIST	0.00000139	0.000146	5.45E-07	49 / 182 (26.9%)	79 / 729 (10.8%)	2.49	0.93

Table 5.5: Motif enrichment by SEA, first stage (Second part)

E-values ≤ 10								
No.	**Motif**	**P-Value**	**E-value**	**Q-Value**	**TP**	**FP**	**Enrichment Ratio**	**Score Threshold**
46	CXSXT	0.00000157	0.000165	6.01E-07	44 / 182 (24.2%)	67 / 729 (9.2%)	2.64	1.2
47	RRP	0.00000326	0.000342	0.00000122	12 / 182 (6.6%)	4 / 729 (0.5%)	10.37	6.7
48	RT	0.00000341	0.000358	0.00000125	35 / 182 (19.2%)	48 / 729 (6.6%)	2.93	2.39
49	RTS	0.0000043	0.000451	0.00000155	16 / 182 (8.8%)	10 / 729 (1.4%)	6.16	5.43
50	SXPXS	0.00000448	0.00047	0.00000158	10 / 182 (5.5%)	2 / 729 (0.3%)	14.63	7.74
51	ETWRTE	0.0000159	0.00167	0.00000549	10 / 182 (5.5%)	3 / 729 (0.4%)	10.97	2.91
52	KNWP	0.0000204	0.00214	0.00000691	41 / 182 (22.5%)	69 / 729 (9.5%)	2.39	0.89
53	SLD	0.000064	0.00672	0.0000213	13 / 182 (7.1%)	9 / 729 (1.2%)	5.58	5.21
54	TR	0.000125	0.0131	0.0000408	26 / 182 (14.3%)	38 / 729 (5.2%)	2.76	8.09
55	HHQK	0.000163	0.0171	0.0000521	32 / 182 (17.6%)	54 / 729 (7.4%)	2.39	0.54
56	STRI	0.00026	0.0273	0.0000818	27 / 182 (14.8%)	43 / 729 (5.9%)	2.54	1.56
57	FST	0.000312	0.0327	0.000095	7 / 182 (3.8%)	2 / 729 (0.3%)	10.64	6.44
58	RGD	0.000313	0.0328	0.000095	38 / 182 (20.9%)	73 / 729 (10.0%)	2.1	3.75
59	GRELCL	0.000369	0.0387	0.00011	6 / 182 (3.3%)	1 / 729 (0.1%)	13.96	2.27
60	NPR	0.000533	0.056	0.000156	29 / 182 (15.9%)	51 / 729 (7.0%)	2.3	4.65
61	EGL	0.00146	0.154	0.000423	9 / 182 (4.9%)	7 / 729 (1.0%)	4.99	5.27
62	KRRREK	0.00236	0.247	0.000669	21 / 182 (11.5%)	36 / 729 (4.9%)	2.37	0
63	ING	0.00305	0.32	0.000852	6 / 182 (3.3%)	3 / 729 (0.4%)	6.98	12.08
64	APF	0.00361	0.379	0.000994	31 / 182 (17.0%)	66 / 729 (9.1%)	1.91	4.01
65	KEICLD	0.00465	0.488	0.00126	5 / 182 (2.7%)	2 / 729 (0.3%)	7.98	2.12
66	IQ	0.00506	0.531	0.00135	87 / 182 (47.8%)	249 / 729 (34.2%)	1.4	1.08
67	EVCH	0.0057	0.598	0.0015	26 / 182 (14.3%)	54 / 729 (7.4%)	1.96	1.49
68	GRK	0.00582	0.612	0.00151	36 / 182 (19.8%)	84 / 729 (11.5%)	1.74	3.79
69	NGRE	0.00609	0.64	0.00155	10 / 182 (5.5%)	12 / 729 (1.6%)	3.38	2.42
70	KIKSST	0.00784	0.824	0.00197	11 / 182 (6.0%)	15 / 729 (2.1%)	2.99	1.39
71	NPASP	0.00797	0.837	0.00198	3 / 182 (1.6%)	0 / 729 (0.0%)	15.96	7.47
72	TVT	0.0157	1.65	0.00385	23 / 182 (12.6%)	51 / 729 (7.0%)	1.84	4.36
73	HGR	0.018	1.89	0.00433	7 / 182 (3.8%)	8 / 729 (1.1%)	3.55	6.01
74	MPF	0.0255	2.67	0.00606	22 / 182 (12.1%)	51 / 729 (7.0%)	1.76	4.61
75	PFL	0.0271	2.85	0.00628	3 / 182 (1.6%)	1 / 729 (0.1%)	7.98	12.02
76	HWGF	0.0271	2.85	0.00628	3 / 182 (1.6%)	1 / 729 (0.1%)	7.98	6.2
77	VQKI	0.0299	3.14	0.00683	6 / 182 (3.3%)	7 / 729 (1.0%)	3.49	8.62
78	YH	0.0351	3.69	0.00793	50 / 182 (27.5%)	146 / 729 (20.0%)	1.38	2.05
79	LVFF	0.0428	4.5	0.00954	17 / 182 (9.3%)	39 / 729 (5.3%)	1.8	1.55
80	PEAPF	0.0558	5.86	0.0122	8 / 182 (4.4%)	14 / 729 (1.9%)	2.39	2.01
81	ADRAA	0.0561	5.89	0.0122	4 / 182 (2.2%)	4 / 729 (0.5%)	3.99	5.23
82	HPH	0.0766	8.04	0.0165	17 / 182 (9.3%)	43 / 729 (5.9%)	1.63	0.37

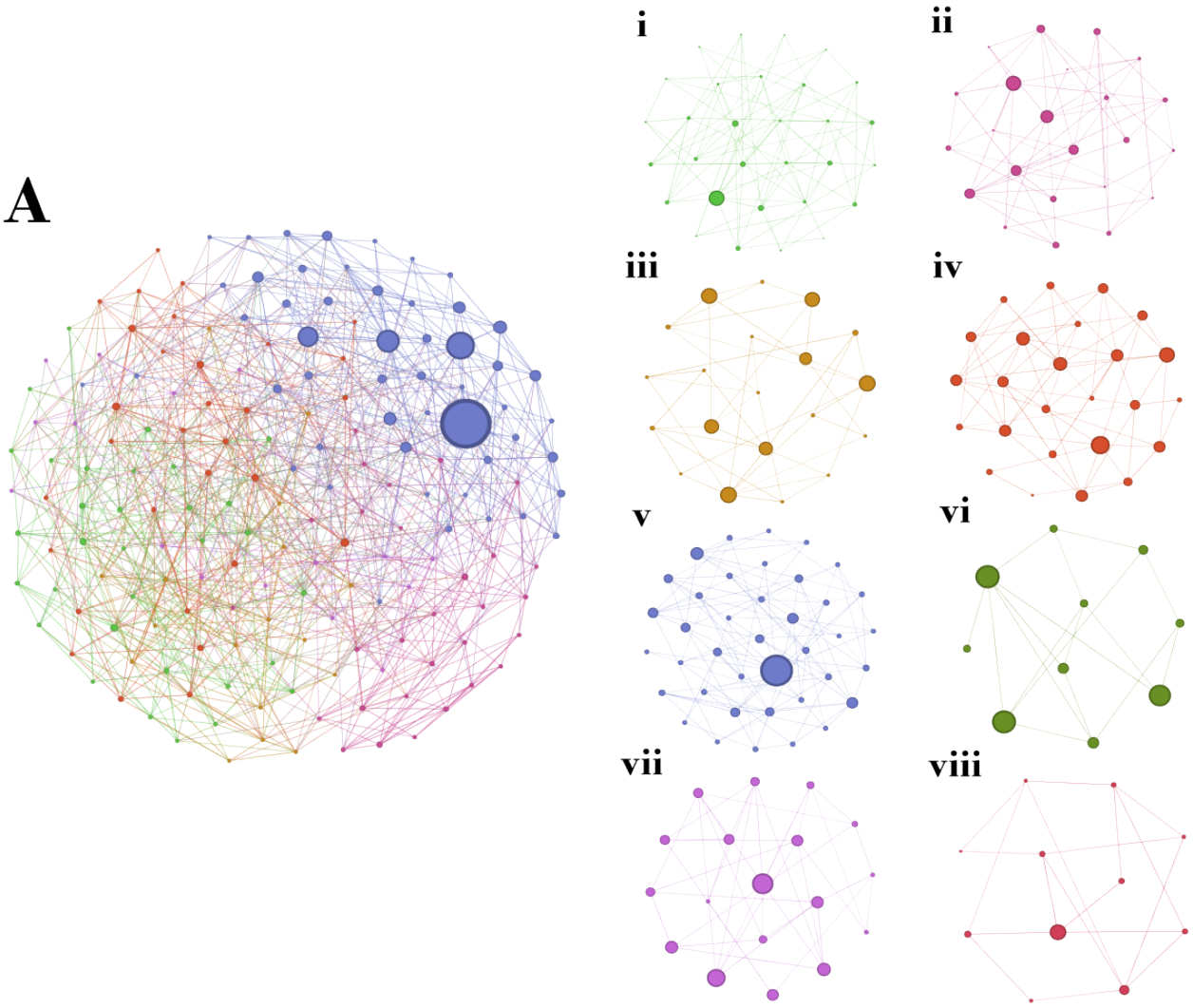

Figure 5.2: Graphical representation of HSPN of Chevyshev Metric with t = 0.00 showcasing its respective clusters: i) Cluster 1, ii) Cluster 2, iii) Cluster 3, iv) Cluster 4, v) Cluster 5, vi) Cluster 6, vii) Cluster 7, viii) Cluster 8 . Node colors signify distinct peptide communities, and the size of the node was calculated by the HB centrality value. Layout: Fruchterman-Reingold [101]. Networks were created with StarPep toolbox [36], visualized in Gephi [87] and edited with Inkscape

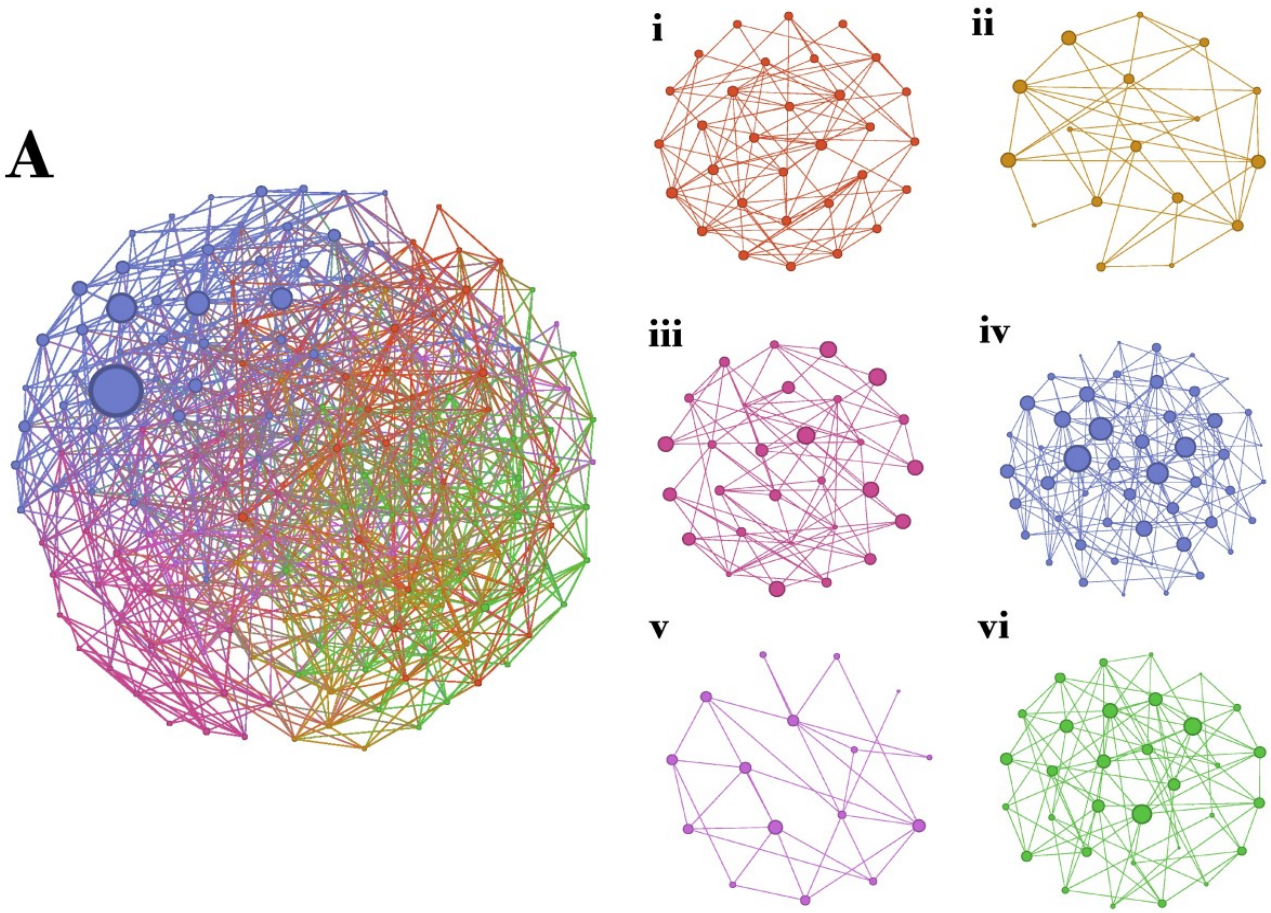

Figure 5.3: Graphical representation of HSPN of Angular Separation Metric with t = 0.00 showcasing its respective clusters: i) Cluster 1, ii) Cluster 2, iii) Cluster 3, iv) Cluster 4, v) Cluster 5, vi) Cluster 6. Node colors signify distinct peptide communities, and the size of the node was calculated by the HB centrality value. Layout: Fruchterman-Reingold [101]. Networks were created with StarPep toolbox [36], visualized in Gephi [87] and edited with Inkscape

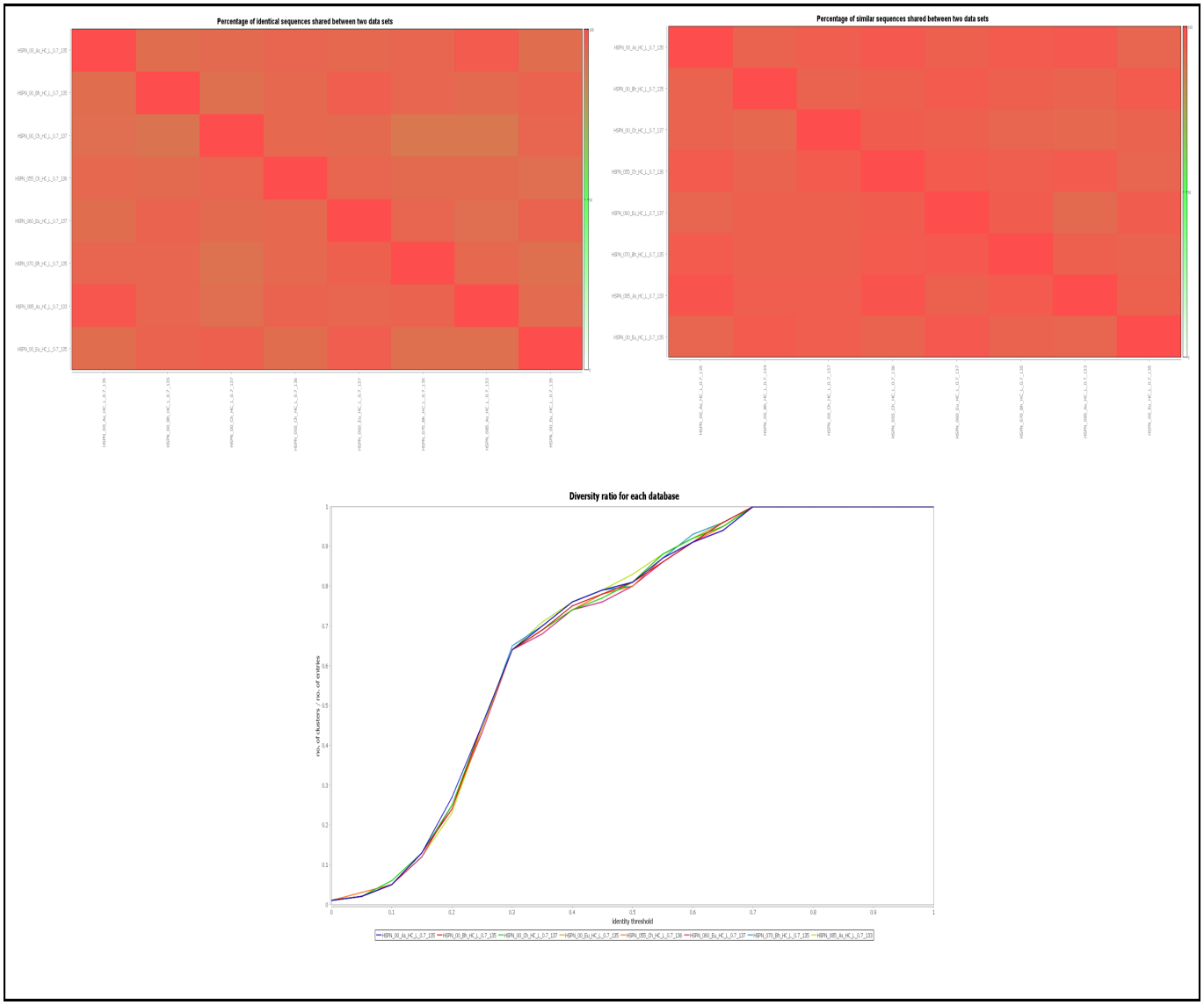

Figure 5.4: Representation of parameters considered for the Analysis 1 of scaffolds extraction in Dover Analyzer. Identical Overlap, Similarity overlap, Diversity Ratio.

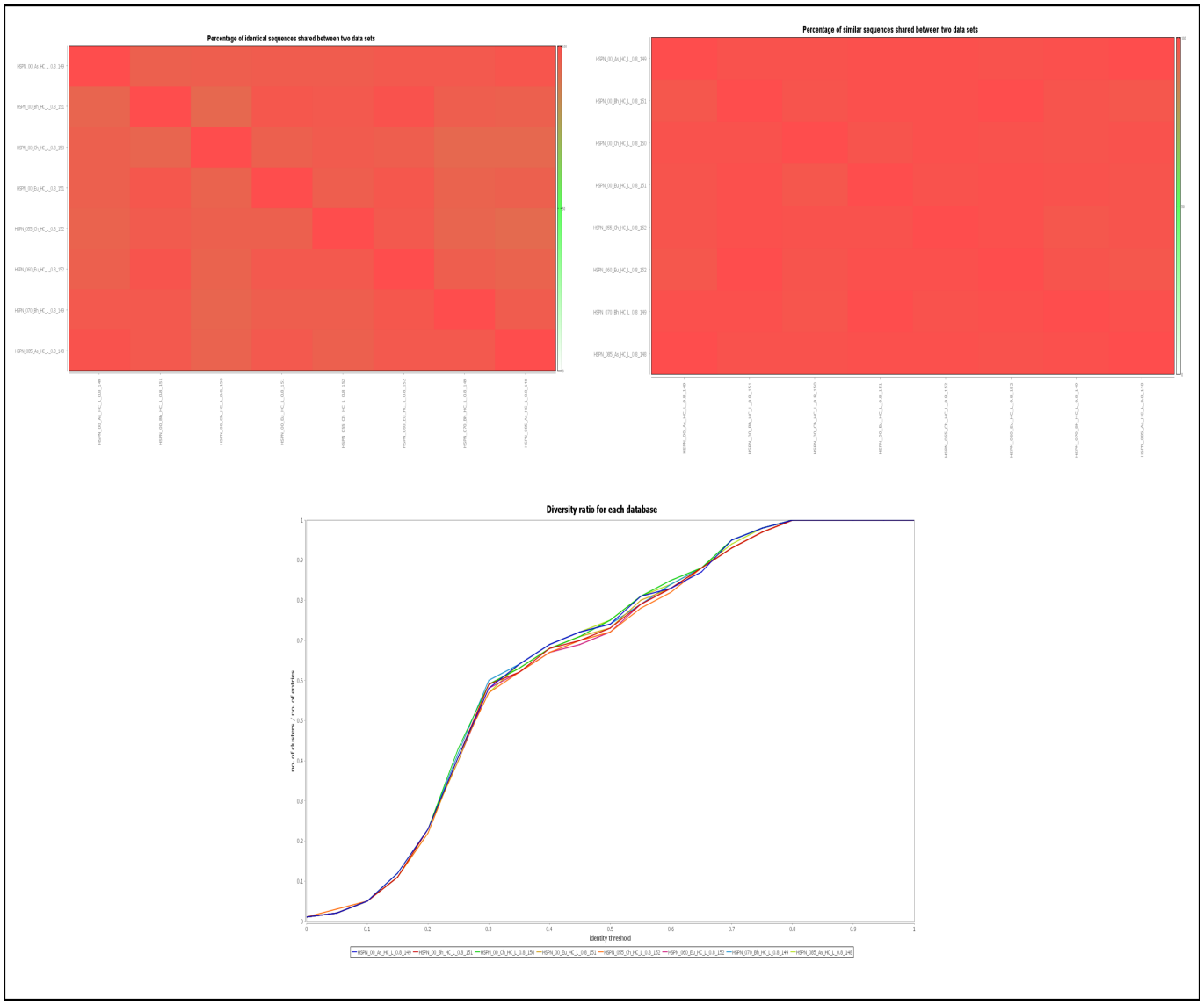

Figure 5.5: Representation of parameters considered for the Analysis 2 of scaffolds extraction in Dover Analyzer. Identical Overlap, Similarity overlap, Diversity Ratio.

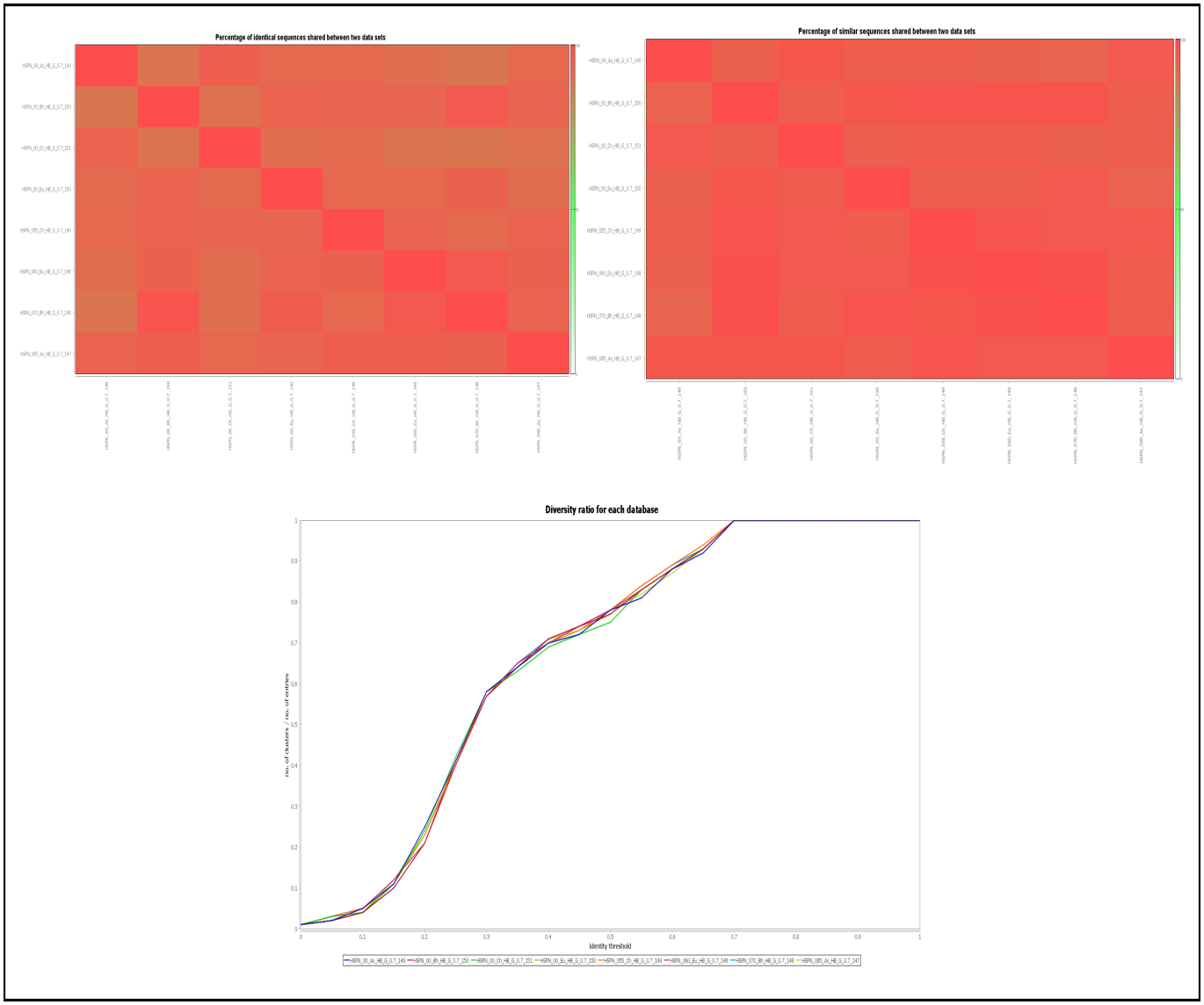

Figure 5.6: Representation of parameters considered for the Analysis 3 of scaffolds extraction in Dover Analyzer. Identical Overlap, Similarity overlap, Diversity Ratio.

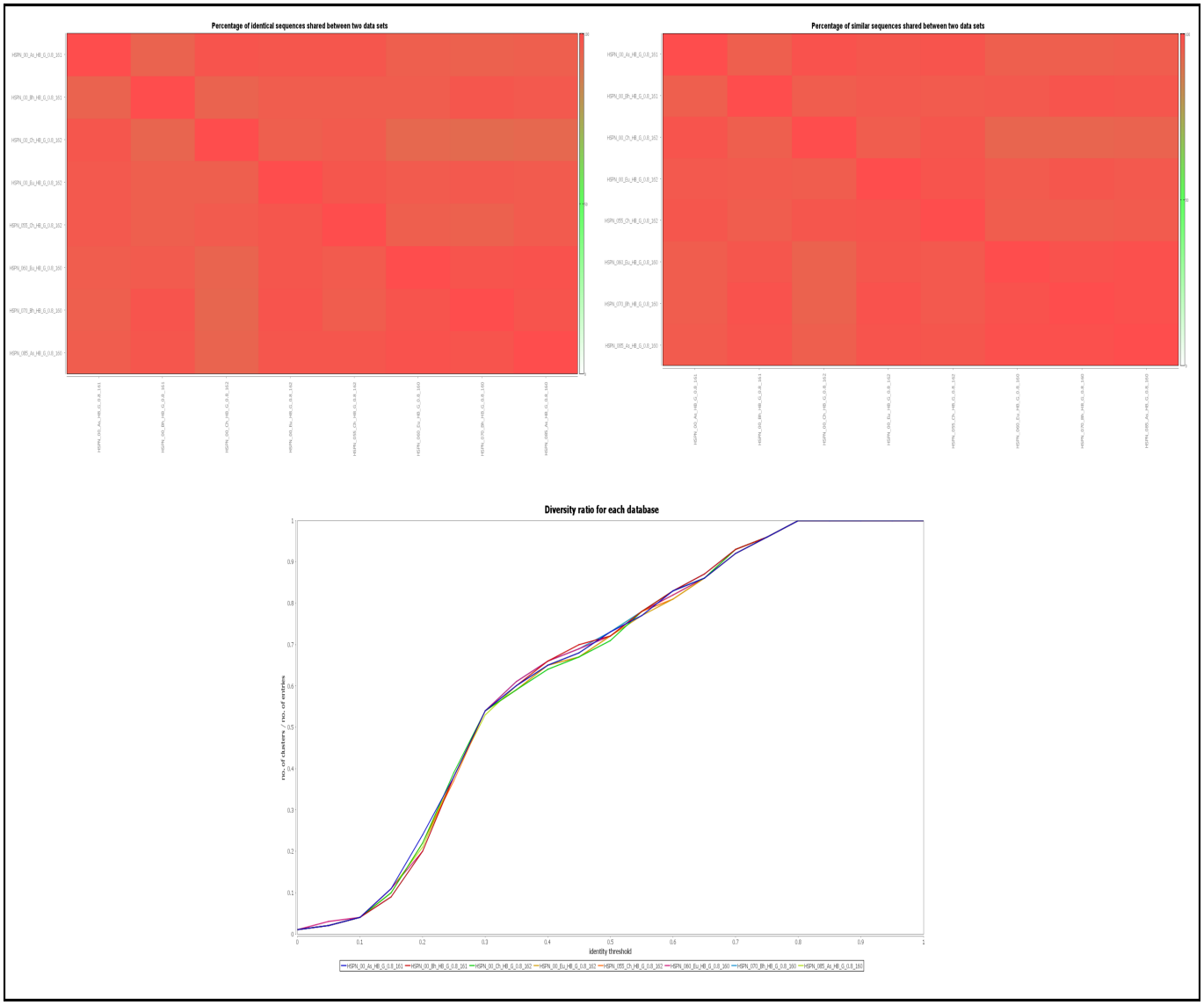

Figure 5.7: Representation of parameters considered for the Analysis 4 of scaffolds extraction in Dover Analyzer. Identical Overlap, Similarity overlap, Diversity Ratio.

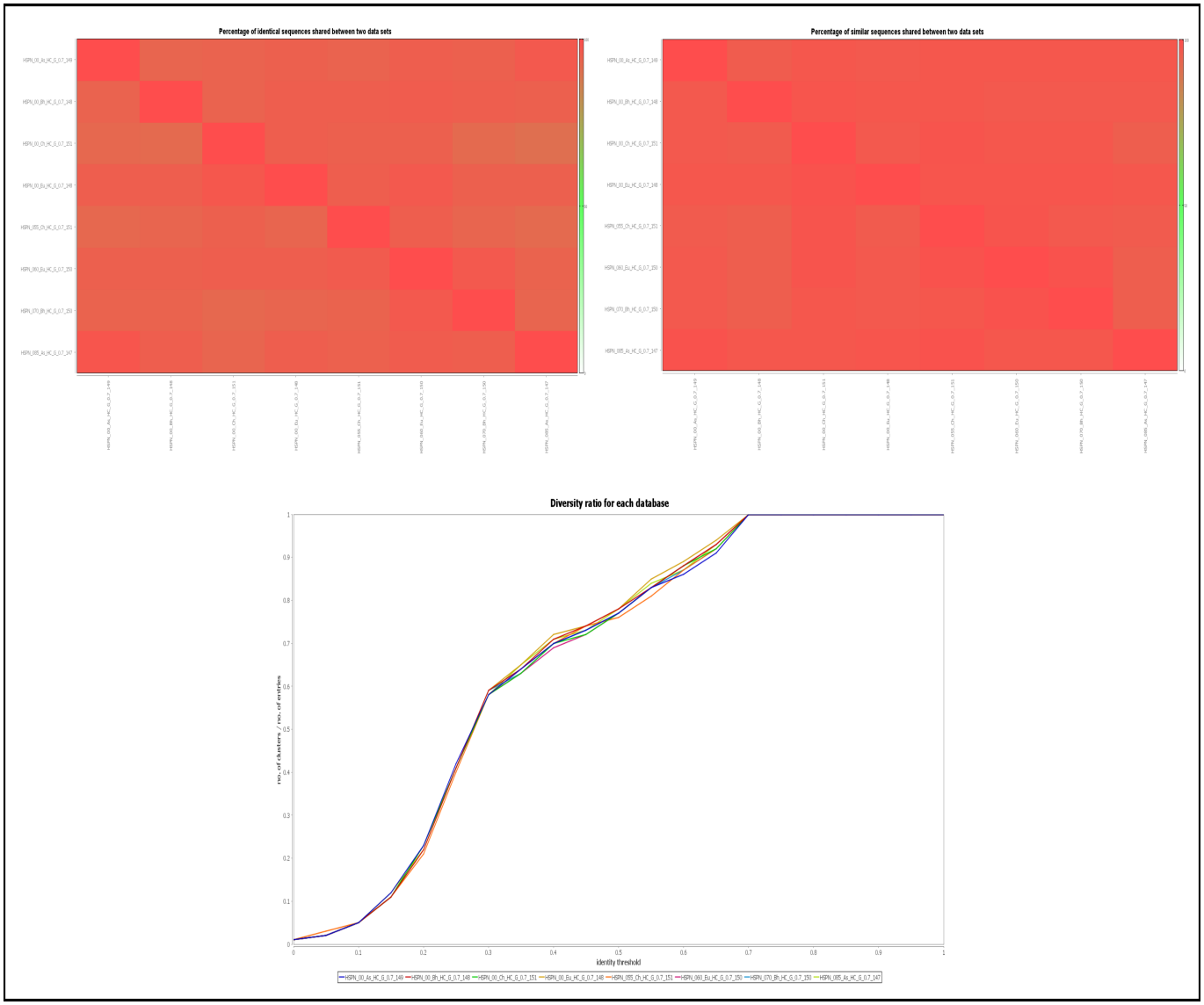

Figure 5.8: Representation of parameters considered for the Analysis 5 of scaffolds extraction in Dover Analyzer. Identical Overlap, Similarity overlap, Diversity Ratio.

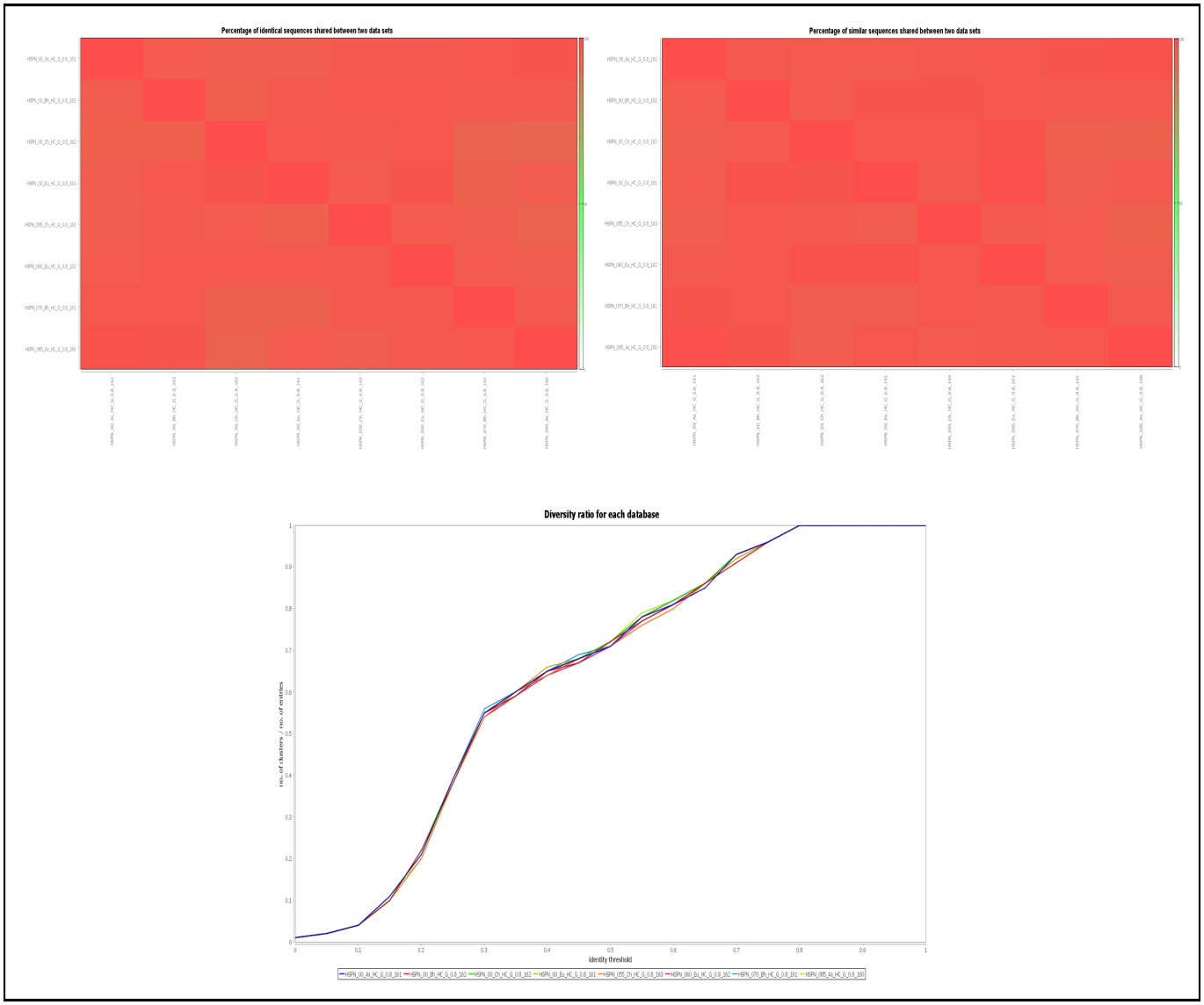

Figure 5.9: Representation of parameters considered for the Analysis 6 of scaffolds extraction in Dover Analyzer. Identical Overlap, Similarity overlap, Diversity Ratio.

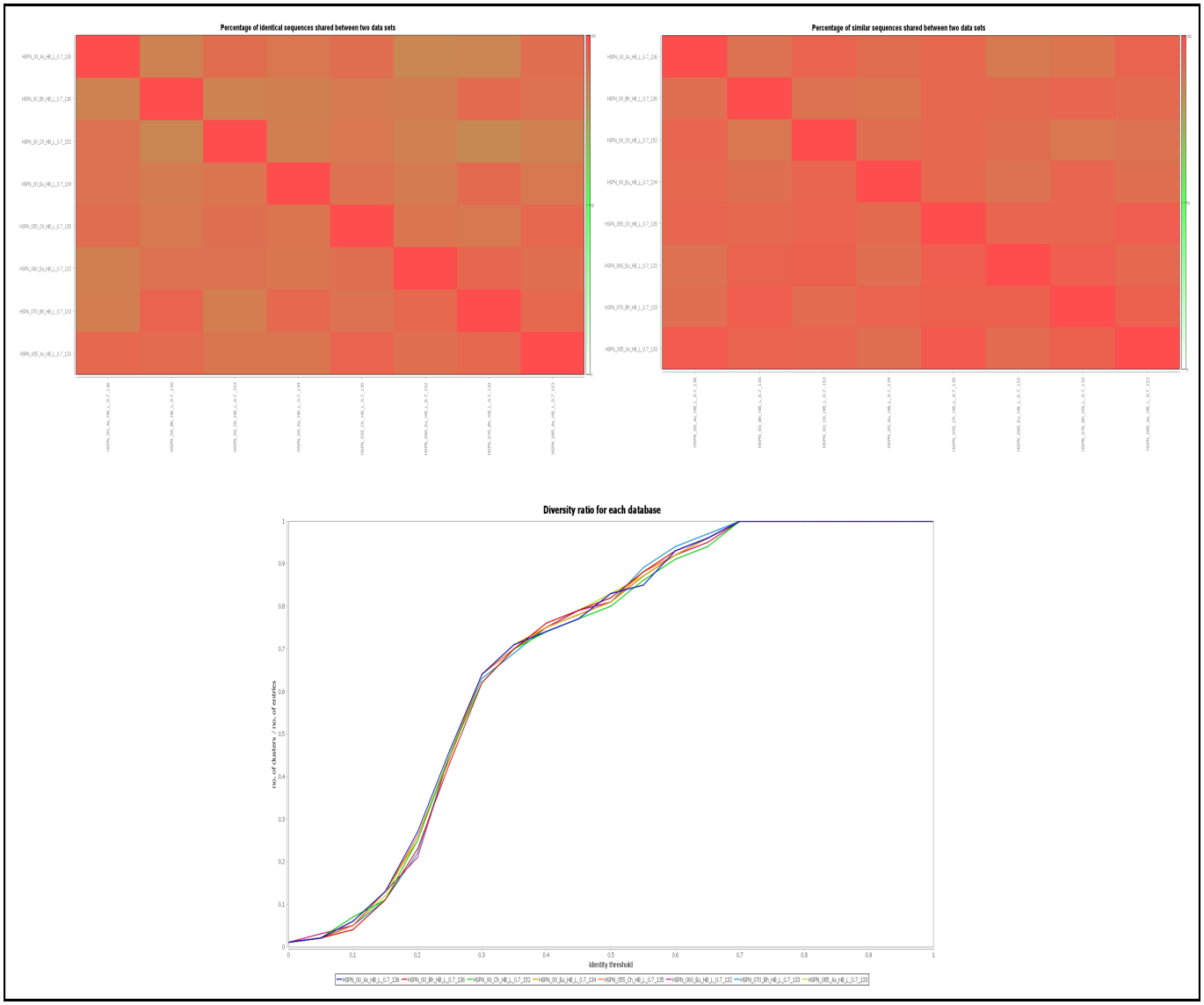

Figure 5.10: Representation of parameters considered for the Analysis 7 of scaffolds extraction in Dover Analyzer. Identical Overlap, Similarity overlap, Diversity Ratio.

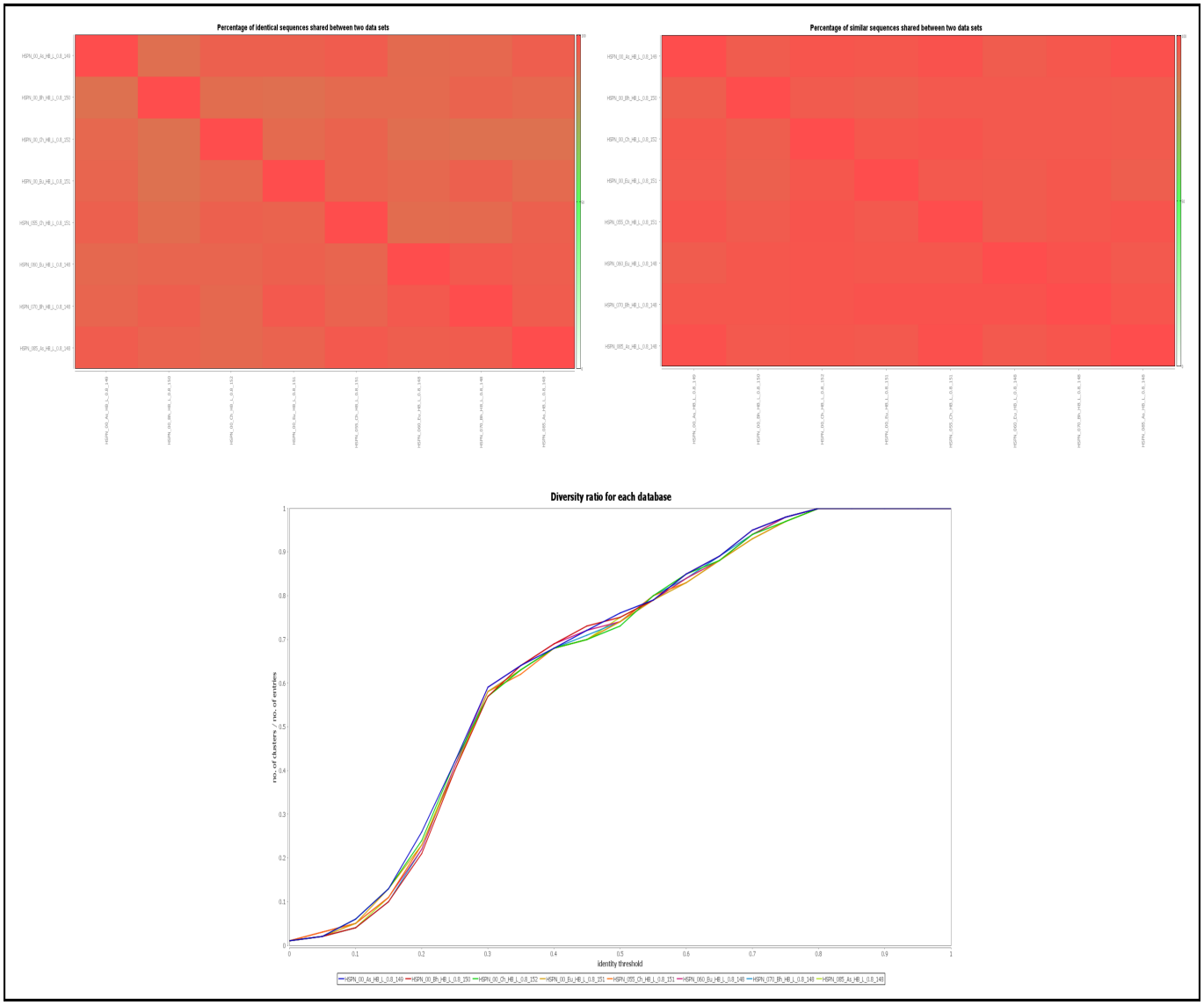

Figure 5.11: Representation of parameters considered for the Analysis 8 of scaffolds extraction in Dover Analyzer. Identical Overlap, Similarity overlap, Diversity Ratio.