Exploring the Chemical Space of Antiparasitic Peptides and Discovery of New Promising Leads through a Novel Approach based on Network Science and Similarity Searching

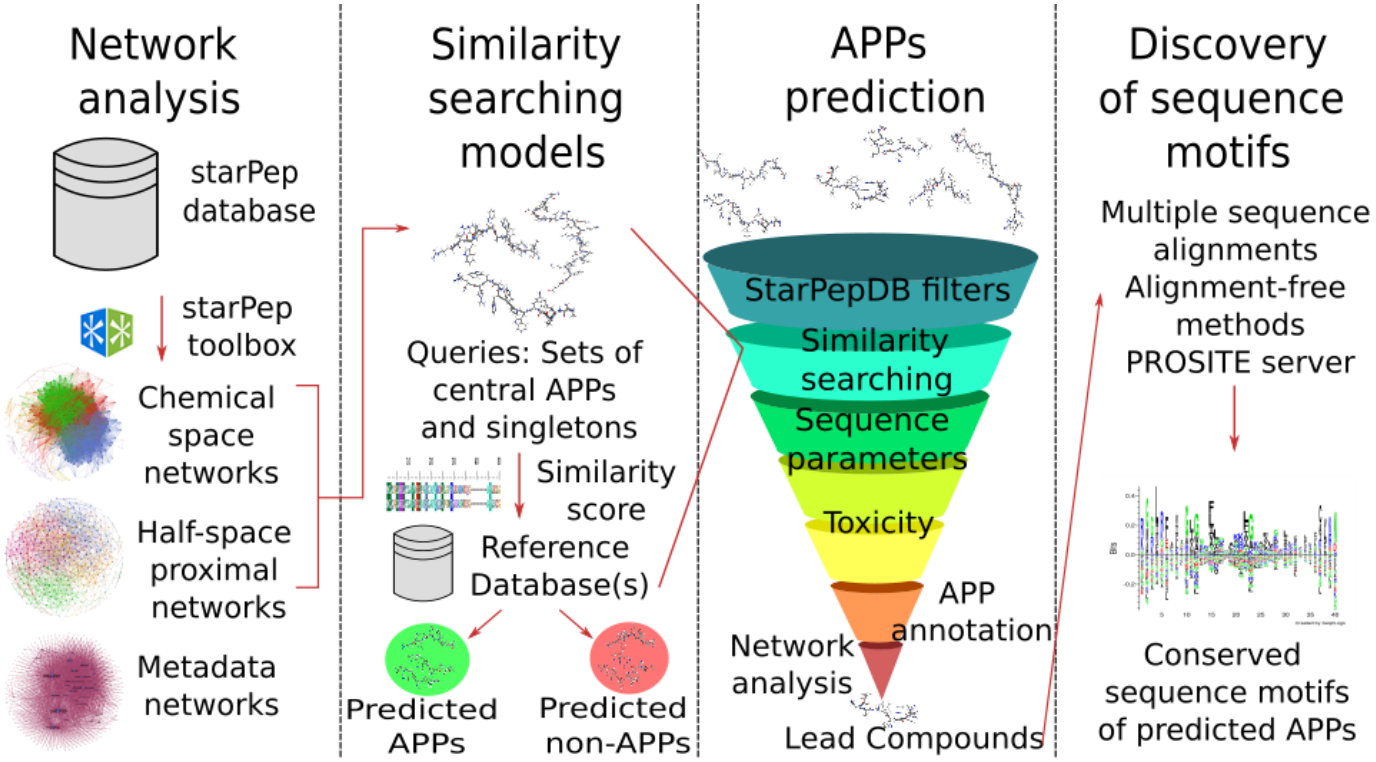
Sebastián Ayala-Ruano,[1,2] Yovani Marrero-Ponce,[1,3,4*] Longendri Aguilera-Mendoza,[5] Noel Pérez,[2] Guillermin Agüero-Chapin,[6,7] Agostinho Antunes,[6,7] & Ana Cristina Aguilar.[1]

[1]*Universidad San Francisco de Quito, Grupo de Medicina Molecular y Traslacional (MeM&T), Escuela de Medicina, Colegio de Ciencias de la Salud (COCSA), Av. Interoceánica Km 12 1/2 y Av. Florencia, 17-1200-841 Quito, Ecuador.*
[2]*Colegio de Ciencias e Ingenierías "El Politécnico", Universidad San Francisco de Quito (USFQ), Quito, Ecuador.*
[3]*Computer-Aided Molecular "Biosilico" Discovery and Bioinformatics Research International Network (CAMD-BIR IN), Cumbayá, Quito, Ecuador.*
[4]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Valencia, Spain.*
[5]*Departamento de Ciencias de la Computación, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Baja California 22860, Mexico.*
[6]*CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n 4450-208 Porto, Portugal.*
[7]*Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal.*

Corresponding authors (*):

Y. Marrero-Ponce, ymarrero@usfq.edu.ec or ymarrero77@yahoo.es ; Tel.: +593-2-297-1700 (ext. 4021). http://www.uv.es/yoma/ or http://ymponce.googlepages.com/home; **ORCID ID: http://www.orcid.org/0000-0003-2721-1142**

TOC/ABSTRACT GRAPHIC



ABSTRACT

Antimicrobial peptides (AMPs) have appeared as promising compounds to treat a wide range of diseases. Their clinical potentialities reside in the wide range of mechanisms they can use for both killing microbes and modulating immune responses. However, the hugeness of the AMPs' chemical space (AMPCS), represented by more than 10^{65} unique sequences of 50 residues long or less, has represented a big challenge for the discovery of new promising sequences and for the identification of common structural motifs and even relevant biological functions. Here, we present a new approach based on network science and similarity searching to discover new promising AMPs, specifically antiparasitic peptides (APPs). We exploited the network-based representation of APPs' chemical space (APPCS) to retrieve valuable information by using three types of networks: chemical space (CSN), half-space proximal (HSPN), and metadata (METN). In the network analysis, some centrality measures were applied to identify in each network the most important and non-redundant nodes/peptides. Then, these central peptides were considered as queries (Qs) in group fusion similarity-based searches against an existing collection of known bioactive compounds, stored in the graph database starPepDB, to discover new potential APPs. The performance of the resulting multi-query similarity-based search models (mQSSMs) was evaluated in five benchmarking data sets of APP/non-APPs. The predictions performed by the best mQSSM showed a strong-to-very strong performance since their external Matthews correlation coefficient (MCC) values ranged from 0.834 to 0.965. Outstanding MCC values (> 0.85) were attained by the mQSSM with 219 Qs from both networks CSN and HSPN with 0.5 as similarity threshold in external datasets. Then, the performance of our best mQSSM was compared with the APPs prediction servers AMPDiscover and AMPFun. The proposed model showed its relevance by outperforming state-of-the-art machine learning models to predict APPs with statistically significant differences. After applying the best mQSSM and additional filters on the non-APP space, 95 AMPs were repurposed as potential antiparasitic leads. Due to the high sequence diversity of these peptides, different computational approaches were applied to identify relevant motifs for searching and designing new promising APPs. These results support that network-based similarity searches identify APPs with high effectivity and reliability. The proposed models and pipeline are freely available through the starPep toolbox software at http://mobiosd-hub.com/starpep.

Keywords: Antimicrobial peptide, Antiparasitic peptide, Network science, Chemical space network, Half-space proximal network, Motif identification, Scaffold extraction, Centrality, Similarity searching, Group fusion, StarPepDB, StarPep Toolbox.

Reading title: *Exploring the Chemical Space of APPs and Discovery of New Promising Leads…*

1. INTRODUCTION

In the last decades, antimicrobials have contributed to preventing and treating infectious diseases caused by bacteria, viruses, fungi, and parasites.[1] Nonetheless, inappropriate use of these medicines, poor sanitary conditions, and other causes have created conditions for the emergence of multidrug-resistant (MDR) pathogens that are not treatable with the available drugs.[2] According to the World Health Organization, antimicrobial resistance (AMR) is one of the top ten global public health threats facing humanity in this century, so this is a worrying issue with potential consequences for human, animal, and environmental health.[3]

In this scenario, it is mandatory to search for new antimicrobials less susceptible to evolutionary resistance mechanisms and that decrease damaging inflammation. Antimicrobial peptides (AMPs) or cationic host defense peptides (CHDPs) have appeared as promising compounds to control infectious diseases avoiding AMR, due to their exceptional microbicidal properties and/or by immunomodulating host responses.[4] AMPs are small bioactive compounds, commonly with fewer than 50 amino acids, amphipathic properties, and a net positive charge between 2-9 at physiological pH.[5] These molecules are part of the primary immune responses, the first defense barrier against microbial pathogens of different living organisms, including bacteria, plants, fungi, invertebrates, amphibians, and mammals.[6]

Some advantages of AMPs over traditional antimicrobials are slower emergence of resistance, antibiofilm activity, modes of action that do not rely on specific targets, and capacity to modulate host immune responses.[7] Therefore, the effectiveness of CHDPs resides on the wide range of mechanisms they can use for both killing microbes and modulating immune responses, which depend on their concentration and dose, external stimuli, target cell or tissue, administration mechanism, host microbiota, and so forth.[5] AMPs can kill microbes affecting

their extracellular dynamics, mainly by the membrane perturbation of pathogens, using different mechanisms. Moreover, these compounds can interrupt transcription, replication, cell wall synthesis, and other important processes by binding to intracellular molecular targets.[4] Some of the CHDPs' immunomodulatory actions are the recruitment of leukocytes to the site of infection, modulating neutrophil responses, influencing on adaptive immune responses, altering inflammatory cytokine patterns, enhancing phagocytosis, microbiome modulation, and wound healing.[8]

It has also been proven that AMPs have potential uses to treat a wide range of diseases, including infections caused by MDR bacteria;[4,9] chronic inflammatory diseases like asthma,[10] arthritis,[11] and colitis;[12] and some types of cancers.[13] Some CHDPs' antimicrobial activities such as antiparasitic, antiviral, and antifungal have been less explored but have a great potential to combat infectious diseases caused by non-bacterial pathogens.[14] Considering this prospect, in this report, we have focused on the AMPs' antiparasitic activity, which could help to treat malaria and neglected tropical diseases such as Leishmaniases, Chagas disease, among others.[15]

There are multiple sources to retrieve AMPs, such as natural peptides produced as part of the immune system of different organisms,[6] synthetic peptides derived from natural CHDPs,[16] cryptic peptides obtained from proteomes or microbiomes using bioinformatics,[17,18] mass spectrometry-based proteomics experiments with fragmentation techniques,[19] and so forth. Therefore, AMPs chemical space (AMPCS) is huge, it is estimated that there are more than 10^{65} unique sequences of peptides with 50 residues or fewer,[14] which represent a big challenge for the discovery of new promising bioactive peptides and the identification of common features (e.g. sequence and structural motifs determining their relevant biological functions).[7] In this context,

computational-aided pipelines have been proposed as efficient alternatives to do high-throughput screening (HTS) of CHDPs.[20]

Traditionally, strategies applied for the discovery of AMPs have relied on bioinformatics methods such as sequence and structure-based alignment searches; pattern-matching approaches like profile Hidden Markov Models and regular expressions; evolutionary algorithms; molecular fingerprint comparisons; and quantitative structure-activity relationship (QSAR) models.[21–23] More recently, machine learning (ML) algorithms, sometimes in combination with the aforementioned methods, have been extensively used to predict and discover new potential AMPs.[24] Most of the ML methods to predict AMPs have focused on supervised strategies, requiring labeled datasets to train these models . These supervised algorithms have shown some issues regarding the size, quality, diversity, application domain, and representativeness of datasets required to train the models, which can produce inappropriate predictions and wrong results.[25]

Considering the limitations and drawbacks of the available methods to discover AMPs, we present a novel approach based on network science tools and multi-query similarity searching models (*mQSSMs*) to discover new potential AMPs, specifically antiparasitic peptides (APPs). Network science is a discipline that studies complex systems, large collections of components that are characterized by having a lot of interactions, emergence, dynamics, self-organization, adaptation, among other properties.[26,27] Similarity searching is a virtual screening strategy that compares a molecular query, characterized by descriptors, against the set of descriptors of other molecules from a database, obtaining a ranked list that possesses the most similar molecules to the query at the top of the list. That is, chemical similarity searching involves the use of a similarity measure (coefficient) to score the degree of similarity between a query structure (or

several queries) and each target compound in a database, and the similar property principle means that the highest-ranked query structures by a similarity measure are likely to have similar properties to that of the template(s).[28,29]

We have taken advantage of network-based representation of APPs' chemical space (APPCS) to retrieve valuable information by using chemical space networks (CSNs), half-space proximal networks (HSPNs), and metadata networks (METNs). Some centrality measures were applied to identify the most important nodes, and these peptides were taken as queries to perform similarity-based searches by group fusion (MAX-SIM rule) models against the graph database *starPepDB* [30] (http://mobiosd-hub.com/starpep). These *mQSSMs* allowed us to repurpose new potential APPs. It is worth mentioning this is the first time we are exploring the chemical space from *starPepDB* to retrieve valuable information of certain known AMPs. To validate the worth of this strategy, we evaluated the *mQSSMs* performance in five benchmarking data sets of APP/non-APPs, and classification results were contrasted with the performance metrics of ML APPs prediction servers *AMPDiscover* (https://biocom-ampdiscover.cicese.mx)[31] and *AMPFun* (http://fdblab.csie.ncu.edu.tw/AMPfun/index.html).[32]

2. DATASETS AND METHODS

Our workflow was divided into four stages: i) networks analysis, ii) multi-query similarity searching models, iii) APPs prediction, and iv) discovery of sequence motifs. The first stage consisted of data extraction, networks building, similarity cutoff analysis, the study of global networks properties, calculation of centrality measures, and retrieval of the most central APPs sets by each metric. The second stage included the design of network-based similarity searching models, selection of the best ones, and comparison of our models with ML approaches to predict

APPs. The third stage was the prediction of new potential APPs by applying the best *mQSSMs* to screen the entire *starPepDB* and subsequently some filters of toxicity and hemolytic activities, the resulting APPs were confirmed with the annotation of external webservers. The last stage was the discovery of sequence motifs shared by the potential antiparasitic leads, using multiple sequence alignments, alignment-free methods, and the *PROSITE* server. TOC/Abstract graphic summarizes all these stages.

2.1. Networks Analysis

2.1.1 Data collection

The APPs were obtained from *starPepDB*, a graph database that contains 45,120 nodes representing AMPs and additional nodes for metadata, obtained from about 40 data sources.[33] As far as we know, this is one of the largest integrated AMPs database until now. The *starPepDB* is embedded in the *starPep toolbox*, a software designed to perform network analysis of data contained in this resource.[30] Thus, we filtered the database by metadata function using the "*Antiparasitic*" search term and retrieved 550 APPs (see SI1-A, a FASTA file).

2.1.2 Creation of networks

The starPep toolbox allowed us to create three types of networks: CSNs, HSPNs, and METNs. CSNs and HSPNs are similarity or correlation networks,[30] defined as $G = (V, E)$ where V is a set of nodes and E is a set of edges. In these networks, nodes in V represent AMPs, characterized by multi-dimensional molecular descriptors vectors, and edges linking nodes in E are pairwise similarity relationships between sequence-based descriptors of the peptides. Thus, nodes of CSNs and HSPNs are connected because they are similar to each other, instead of the existence of physical interactions among these compounds.[30] In addition, METNs are multilayer networks, defined as $G = (V, E, L)$, where V and E are the sets of nodes and edges, same as in

CSNs and HSPNs, and L is the set of layers representing different edge types or labels.[34] METNs have two layers: metadata and AMPs. Metadata consist in additional information of AMPs such as origin, database, function, target pathogen, crossref, N-terminus, C-terminus, and unusual amino acids. Note that CSNs and HSPNs are networks where there is only one kind of node and relationship. However, a more interconnected system has been considered for further analysis, by connecting the nodes representing AMPs with their different types of metadata such as origin, database, function, target pathogen, and so on.[35] Among these nodes, peptides and metadata, the edges depict multi-type links and hierarchical connections for a better organization and network navigation. Hierarchy relationships between nodes are established by the edges in both layers of METNs.

In CSNs and HSPNs, the set of molecular descriptors that codifies an AMP can be derived from sequence molecular descriptors by applying statistical and aggregation operators on amino acid property vectors (see SI1-2 in ref.[30]). The cross-out sequence's descriptors were calculated using StarPep by selecting all the available amino acid properties (e.g., the heat of formation, side-chain mass, among others), all groups of amino acid types (e.g., aliphatic, aromatic, unfolding, and so forth), and traditional aggregation operators but those based on GOWAWA and Choquet integral. The neighborhood (k neighbors up to 6) was included as one of the aggregation operators.[30]

The selection of suitable sequence descriptors to map the chemical space of AMPs is a key parameter to create CSNs and HSPNs, which was widely explored in ref.[30] From the defined chemical space, the similarity relationships between AMPs form a symmetric similarity matrix M of size $|V|x|V|$, being $|V|$ the number of AMPs. The symmetric property of M means that $\forall_{u,v}$ $M_{u,v} = M_{v,u}$, where u and v are any two nodes from V. Each entry $M_{u,v}$ corresponds to the

similarity score between nodes *u* and *v* in *M*. Then, a similarity threshold *t* is applied on *M* to filter the most prominent similarity relations, and if the similarity scores are greater than or equal to *t* they remain on *M*, otherwise, they are assigned to zero.[36,37] The new matrix is known as threshold matrix *T*, and both CSN and HSPN were constructed from *T*.[30] Edges are filtered according to the criteria summarized in the following expression:

$$T_{u,v} = \begin{cases} T_{u,v} \; if \; u \neq v, M_{u,v} \geq t \\ 0 \; otherwise \end{cases} \qquad \textbf{(1)}$$

Therefore, CSN and HSPN are weighted and undirected networks, with similarity values between AMPs as weights, and there exists an edge between two nodes if this value is greater than or equal to a given cutoff *t*.[30] The criteria applied to choose this similarity threshold are explained in the next section.

The main difference between CSNs and HSPNs is the way they are constructed. CSNs create a similarity matrix of all pairwise relationships between nodes and establish an edge only if the pairwise similarity value is equal or greater than a given threshold.[38] On the other hand, HSPNs do not consider all the possible pairwise relationships between nodes, instead, these networks apply the half-space proximal test over the set of nodes,[39] obtaining a connected network with a small fraction of the maximum number of edges.[30] HSPNs have been applied to create a vector representation of residues contacts in protein 3D structures,[40] but this is the first time we are using them to represent the AMPCS.

In this report, we created CSN, HSPN, and METNs of the 550 APPs available in starPepDB. METNs of origin, database, function, and target pathogen were constructed using the *Metadata Network option* from the *starPep toolbox*. For CSN and HSPN, we firstly chose the default sequence identity value to remove redundant APPs. Thus, 415 APPs sharing a maximum of 98% of identity were used to generate the networks (see SI1-A_I), applying the local alignment

algorithm Smith-Waterman[41] and Blosum 62 substitution matrix. Secondly, an optimized set of alignment-free sequence descriptors and methods were used to represent them as recommended in ref.[30] Next, the Euclidean distance metric with min-max normalization to establish the pairwise similarity relationships among nodes was applied. The similarity threshold of both networks was set up considering different parameters, as is explained in the next section. Then, we retrieved CSN and HSPN giant components (405 APPs comprised the giant component of both networks, see SI1-A_II), defined as the largest connected component of a network, using the *Central Informative Nodes in Network Analysis (CINNA)* R package.[42] The giant components of both networks were used onwards for all calculations. To visualize these networks in a meaningful way, we examined a family of force-directed layout algorithms that can be used to spatialize the network and rearrange nodes. These algorithms change the position of nodes by considering that they repulse each other, whereas similarity relationships may attract their attached nodes like springs.[43] Particularly, the *Fruchterman-Reingold* algorithm[44] was the most suitable for drawing CSN and HSPN of APPs.

In addition, we created two null network models with the same number of nodes and edges of CSN and HSPN giant components by applying the *Gilbert* method, a variation of the well-known *Erdős-Rényi* model. In this random network model, the edges are chosen uniformly and randomly from the set of all possible edges of the network.[45] We created these random networks using the *sample_gnm* function from the *igraph* R package,[46] applying a random seed of 100 with the *seed* function. All network visualizations were customized with *Gephi* [47] and Inkscape.[48]

2.1.3 Networks similarity threshold analysis

We constructed CSNs and HSPNs of APPs according to specifications explained in previous section but changing similarity threshold in the range of 0.05 and 0.90 with a step of 0.05 (36

networks in total, 18 for each network type). As long the cutoff increases, some edges were removed from networks, becoming increasingly sparser graphs. Then, we retrieved some metrics of these networks at different similarity cutoffs using the *starPep toolbox*. The first metric was the network density, which is the actual number of edges over the maximum number of edges in a network[49] (equation 2). This metric was obtained with *starPep toolbox*, according to the variations of the similarity cutoffs.

$$D = \frac{2|E|}{|V|(|V|-1)} \qquad \textbf{(2)}$$

where, $|V|$ is the number of vertices and $|E|$ the number of edges.

We also removed singletons, nodes without edges connected to it or the ones with zero-degree,[27] filtering the 550 APPs by *Network measure* with attribute *Weighted Degree* to be greater than zero.

The modularity of the networks was also analyzed at each similarity cutoff. This is a network measure that compares the density in a community with the expected density for the same group of nodes on a random network.[34] We calculated modularity and the number of communities using the *modularity optimization* clustering algorithm (based on the Louvain method).[50] The modularity of undirected and weighted networks is defined as:

$$M = \frac{1}{2|E|}\sum_{u,v\epsilon V}\left[A_{uv} - \frac{k_u k_v}{2|E|}\right]\sigma(c_u, c_v) \qquad \textbf{(3)}$$

where, $|E|$ is the number of edges, A_{uv} is an entry of an adjacency matrix, k is the degree of a particular node, and σ is the Kronecker delta, a function that returns one if u and v are in the same community ($c_u = c_v$) or zero otherwise.[51]

According to a previous report, the average clustering coefficient (ACC), a global measure of nodes neighborhood connectivity,[27] is a good indicator to set up the proper similarity threshold to create similarity networks,[36] so we calculated the ACC for all networks using the *transitivity*

function from the *igraph* R package,[46] which applies the definition of ACC for weighted networks proposed in ref.[52] Therefore, we decided the best similarity threshold for CSNs and HSPNs evaluating all the aforementioned network measures. Plots of this sections were created with *ggplot2*[53] R package.

2.1.4 Study of global networks properties

The global characterization of networks is useful for identifying general topological and structural patterns, and thus, understanding the phenomena we are modeling, in this case the representation of the APPCS. These calculations were applied only to CSN and HSPN with the best similarity thresholds. In the last section, we obtained some of these features such as density, number of communities, modularity, singletons, and ACC for CSNs and HSPNs and their respective null network models. These properties are related to the number of edges, connectivity, and community structure of networks.

Moreover, we measured some properties associated with component structure and reachability of networks, including the number of connected components, diameter, and average shortest path (ASP). A connected component is a subnetwork whose nodes can be reached from one another by traversing edges.[34] Diameter is defined as the largest shortest path of a network, and ASP corresponds to the expected length of the shortest path between two nodes chosen at random.[49] The formal definition of ASP is presented below:

$$ASP = \frac{\sum_{u,v \epsilon V} |P_{uv}|}{|V|(|V|-1)} \qquad \textbf{(4)}$$

where, $|V|$ is the number of vertices and $|P_{uv}|$ is the length of a path to go from u to v.

We also plotted the degree distributions of networks. All these metrics were calculated with the *starPep toolbox* and *igraph* R package.

2.1.5 Centrality analysis

Centrality is a key concept in network science, it provides an intuition of the importance of nodes in networks, which may play critical roles in the system that is being modeled.[54] In this study, the most influential nodes can provide useful information from the APPCS, and also these peptides can be used to retrieve new potential APPs by similarity searches. Therefore, we calculated the four centrality measures available in the *starPep toolbox* (weighted degree (WD), betweenness (BE), harmonic (HC), and hub-bridge (HB)) for both CSN and HSPN, and all of these values were normalized with the min-max method. Also, we explored possible correlations between these measures with the Spearman coefficient, using the *corrmorant* R package.[55] To corroborate the correlation analysis, common APPs were identified within the top-50 most central nodes retrieved by different centrality measures, using *in-house* R scripts.

To retrieve the most central and unique APPs sets by each metric, first, we decreased the redundancy among them by applying the *Scaffold extraction* plugin from the *starPep toolbox*. These peptides were ranked in decreasing order by each centrality measure, and redundant sequences were removed at a given percentage of sequence identity. We chose a sequence identity value of 50% to consider that a particular peptide is related to an already selected central peptide and, as a consequence, removed from the network. For these sequence comparisons, we applied the Smith-Waterman local alignment algorithm[41] and BLOSUM62substitution matrix. Then, we retrieved APPs whose centrality scores were at least 10% of the most central APP value by each metric. We applied this process for both CSN and HSPN.

2.2. Similarity Searching Models

2.2.1 Description of models

Our models consisted of multi-query searches against some databases, and the combination of similarity scores by group fusion applying various similarity thresholds. All the components of our models are explained below:

- **Query datasets:** The queries of our model were the most central and non-redundant APPs sets by the four centrality measures considered in this study for CSN, HSPN, and the consensus sets of both networks. In addition, we considered the set of 13 singletons (see SI1-B, a FASTA file), which was the same for both networks, and some combinations of the most promising sets. Finally, we had twenty-one query datasets, seven for each network, six for the combination of both networks, and the set of singletons.
- **Target or calibration databases:** We considered five databases of APPs and non-APPs reported in ref.[31] There were different balanced and unbalanced datasets stored in five FASTA files with thousands of labeled APPs and non-APPs (see SI1C-G).
- **Similarity coefficient:** Smith-Waterman local alignment algorithm,[41] implemented in BioJava,[56] with Blosum62 substitution matrix allowed us to calculate the similarity scores, which were numbers between 0 and 1.
- **Group fusion:** In this fusion model, the reference peptide can be presented by anyone of the extracted peptidic scaffolds (reduced chemical space), however the similarity measure (defined below) was kept constant Some studies have demonstrated that fusion by similarity scores and the maximum fusion rule are the best parameters for these models,[29,57] so we implemented these standards in our pipeline. Therefore, given a reference peptide *Q* and a peptide *D* from the target database, the algorithm of group fusion measures similarity scores *S(Q, D)* between *Q* and all the molecules of the

database, and retrieves the single fused score by the maximum fusion rule. Thus, the fused score is the largest of all similarity scores.

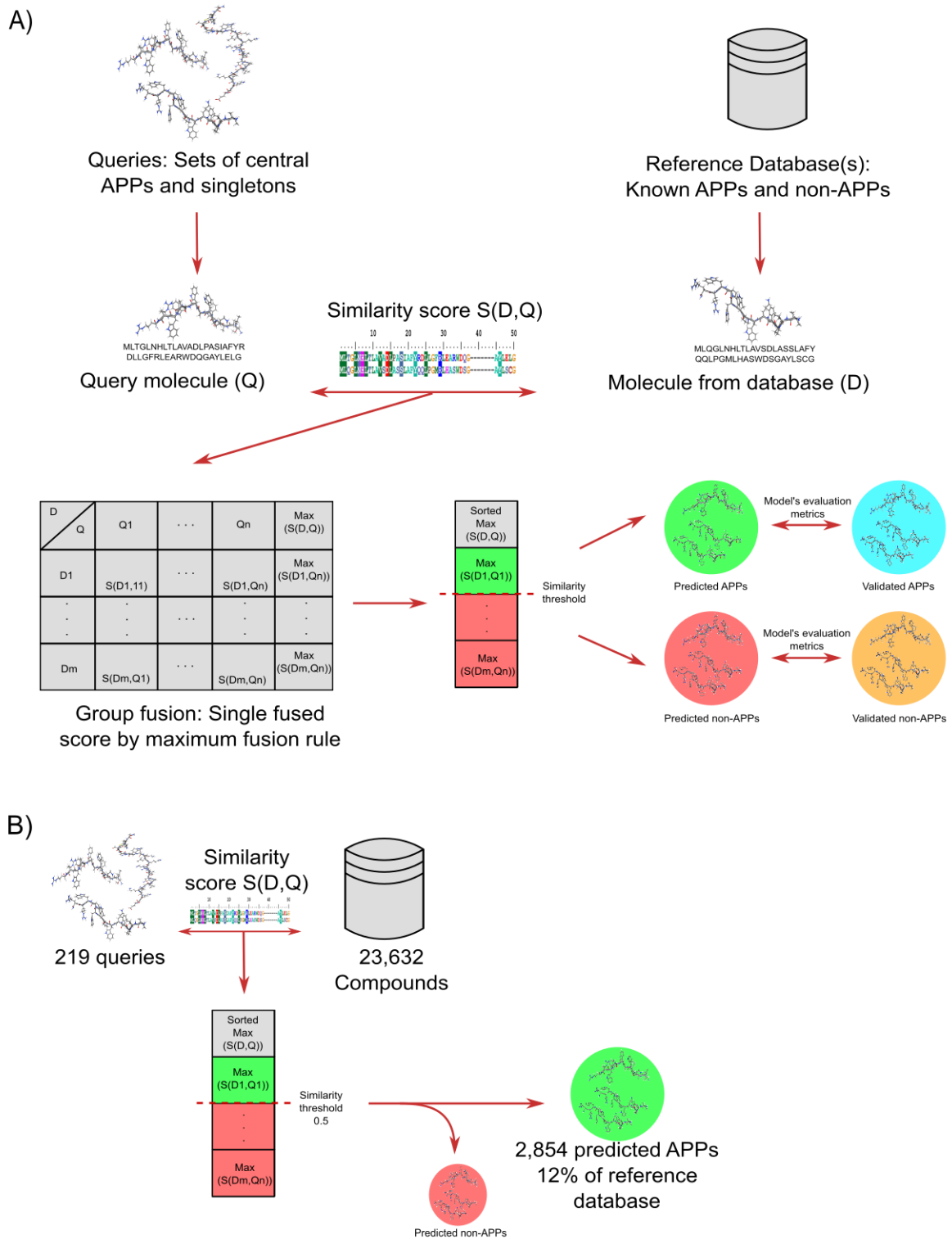
- **Similarity threshold:** After applying the group fusion model for all queries of a dataset, we ranked the results in decreasing order of the fused scores. Then, we tested seven similarity thresholds in the range of 0.3 and 0.9 with a step of 0.1. Therefore, all of the peptides with fused scores greater than or equal to the specific cutoff were predicted as APPs.

We performed these models with the *starPep toolbox*, using the *Multiple query sequences* option of the *Peptides search by* menu. In this software, group fusion by similarity scores and the maximum fusion rule are implemented by default, and users can change the query set, target/calibration dataset, similarity coefficient, and similarity threshold. Thus, we imported each of the five target databases to the starPep toolbox in different workspaces, and we applied multi-query sequence searches with each of the twenty-one queries sets against each of the target databases. The query datasets were composed of central and singleton peptides, previously selected by *scaffold extraction* protocol with the starPep toolbox. As we had twenty-one query datasets and seven similarity thresholds, we evaluated 147 different mQSSMs. The best models were identified using several classification performance measures, which are explained in the next section. We presented a graphical summary of the pipeline used in our mQSSMs in Scheme 1A.

2.2.2 Selection of the best of models and comparisons with ML APPs prediction servers

To assess the relative performance of the *mQSSMs*, we used the five data sets of APPs and non-APPs recently provided in ref.[31] These datasets were obtained from starPepDB, whose description can be found in https://biocom-ampdiscover.cicese.mx/dataset. Each set of queries

Scheme 1. A) Workflow corresponding to the similarity searching modeling process (retrospective study) and B) APPs selection process (prospective virtual screening study). This scheme was created with *Inkscape*.[48]



and similarity thresholds were wrapped into a calibration algorithm, comprising a modified virtual screening simulation technique.[31] In these models, we used just the queries' subset of APPs as the multi-query calibration group, while the active and inactive subsets were the target datasets. The prediction ensemble, composed of similarity scores of each peptide *D* in the target dataset with each query *Q*, was ordered with the MAX-SIM multi-classifier.[58,59] The ordered list was scanned for every active and inactive APP-labeled peptide of the target database, and these results were used to calculate performance metrics, obtained from the confusion matrix.[60] (Scheme 1A). The performance metrics were used to evaluate the quality of the early retrieval.

In ref.[61] the authors reported a unified overview of methods that are currently used for evaluating classification tasks, as well as the advantages and downsides of each approach. Here, we used the following performance metrics derived from the confusion matrix of the actual *versus* predicted class: i) sensitivity (*SN*, also called true positive rate, hit rate, and recall), ii) precision (*PR*), iii) specificity (*SP*, also called true negative rate), iv) accuracy (Q% - global good classification), v) kappa, and vi) Matthews correlation coefficient (MCC). These performance metrics were calculated with the formulas presented below:

$$\begin{cases} SN = \frac{TP}{TP+FN} \\ PR = \frac{TP}{TP+FP} \\ SP = \frac{TN}{TN+FP} \\ Q\% = \frac{TP+TN}{TP+TN+FP+FN} \\ k = \frac{p_o - p_e}{1-p_e} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \quad \textbf{(5)}$$

where, TP are true positives, TN true negatives, FP false positives, FN false negatives, p_o the relative observed agreement among raters, and p_e the hypothetical probability of chance agreement.

Finally, the performance metrics for the five calibration datasets were used to carry out comparisons among models by using the statistic of Iman and Davenport.[62] These statistical tests showed that Friedman's value was undesirably conservative. Whenever significant differences were detected, the post hoc tests[58,63,64] were used to compare the Friedman best-ranked models or reference measure with the remaining ones. This step-up procedure works in the opposite direction to Holm's test and allows the control of the so-called familywise type I error arising from multiple pair-wise comparisons.[63] After applying these statistical tests, we obtained the best *mQSSM*. A second comparison was carried out to compare our best similarity searching models with ML-based models reported in the literature for APP prediction[31,32] by using the same five calibration datasets.

2.3. New APPs Predictions

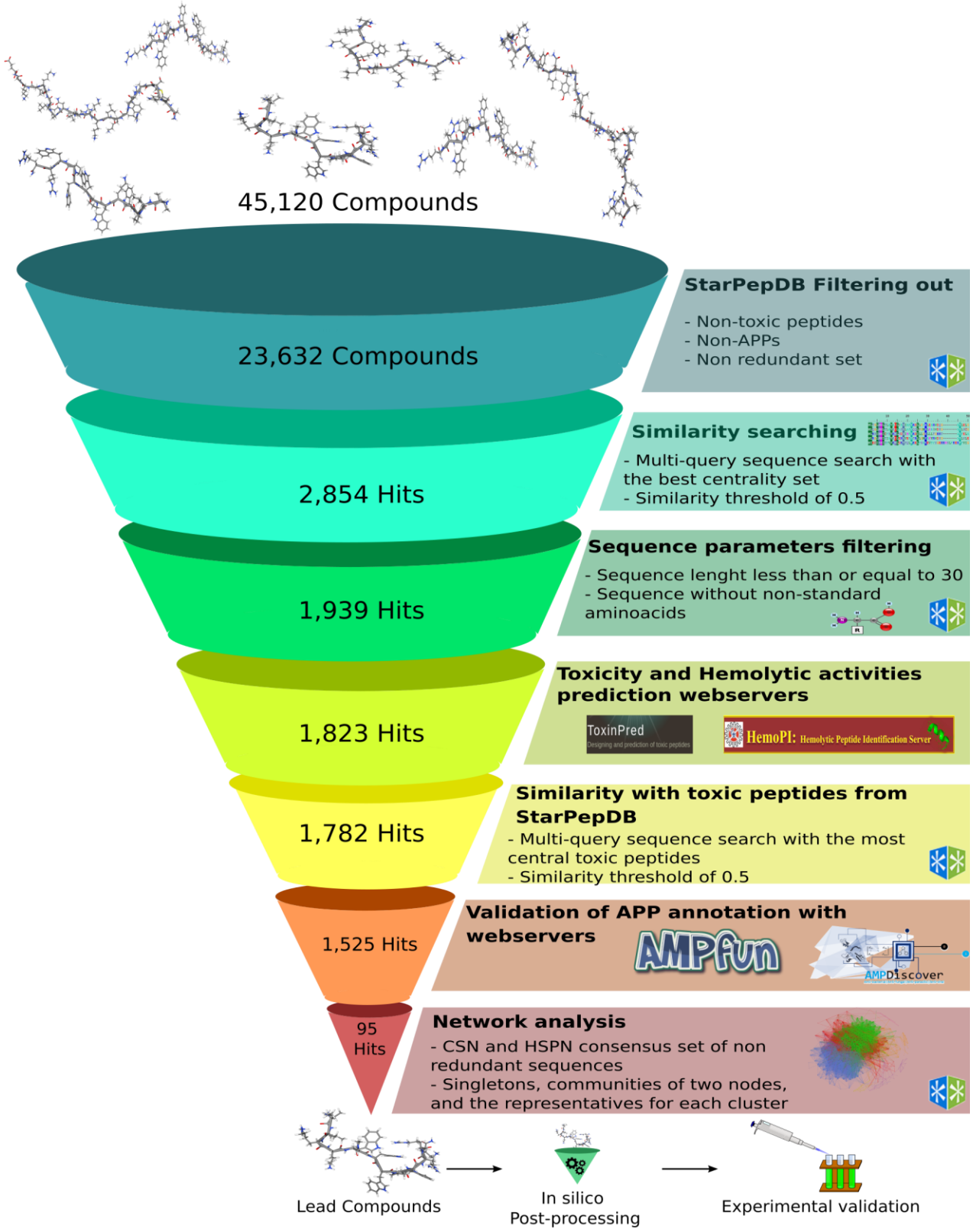
We used the *starPepDB* as a space of search to discover new potential APPs and the *starPep toolbox* for exploring the APPCS. First, we removed the toxic peptides and known APPs from starPepDB, applying the not operator and filtering the database by metadata Function with *Antiparasitic*, *Toxic*, and *Toxic/Venom* queries. Then, we reduced the redundancy of these sequences with the *nonredundant* plugin, applying a similarity identity value of 0.95 with the local alignment algorithm Smith-Waterman,[41] and BLOSUM62 substitution matrix. These non-toxic, non-APP, and nonredundant peptides were the chemical space to search for new potential APPs. Hence, we used the best *mQSSM*, obtained in the previous section, as a prediction tool to detect new APPs in the before-mentioned chemical space.

We removed from the space of unknown and non-toxic APPs, those virtual hits with sequence length greater than or equal to 30, sharing similarity score of one, and that contain non-standard amino acids. To avoid toxic peptides in our list of APPs candidates, we uploaded the

FASTA file of these sequences to the *ToxinPred* server (https://webs.iiitd.edu.in/raghava/toxinpred/)[65], applying the *SVM (Swiss-Prot) + Motif based* model with an SVM threshold of zero. The APPs predicted as toxic by ToxinPred were discarded. We also used the *HemoPI* server (https://webs.iiitd.edu.in/raghava/hemopi/)[66] to remove potential hemolytic peptides, applying *SVM + Motif (HemoPI-1)* and *SVM + Motif (HemoPI-2)* models, and removing peptides with a PROB score greater than or equal to 0.7 in both models. Then, the remaining virtual hits were further reduced by developing mQSSM aimed at detecting similarities with the most central toxic peptides available in *starPepDB.* The centrality analysis was performed as explained above.

In addition, we used *AMPDiscover*[31] and *AMPFun*[32] servers to confirm our APP predictions. Therefore, Random Forest and Deep Learning models of *AMPDiscover*, and *AMPFun* were evaluated on the virtual hits. Then, we created a CSN and an HSPN with the remaining virtual hits and applied non-redundant scaffold reduction based on Harmonic centrality with a 0.7 identity threshold on each network. Thus, we obtained the consensus set of sequences between both networks. Ultimately, a CSN was constructed with the remaining set of virtual hits. We selected the singletons and communities with two nodes, applied the *Modularity optimization* clustering algorithm, and extracted the non-redundant set for each community applying a similarity threshold of 0.5, the Harmonic centrality, and the other parameters established by default. Therefore, singletons, communities of two nodes, and representatives for each cluster were the lead peptides proposed as potential APPs in this study. A graphical summary of this section is depicted in Scheme 2.

Scheme 2. Filtering workflow to obtain the new potential APPs. This scheme was created with *Inkscape*.[48]



2.4. Discovery of Sequence Motifs

2.4.1. Multiple sequence alignments

We created a CSN with lead compounds and obtained its communities using the *Modularity optimization* clustering algorithm. Then, these resulting clusters were aligned independently by using multiple sequence alignment (MSA), publicly available at https://www.ebi.ac.uk/Tools/msa/. To determine consensus motifs within each cluster, three different MSA algorithms were applied with their default parameters: Multiple Alignment using Fast Fourier Transform (*MAFFT*) v7 with the iterative refinement FFT-NS-i option,[67] Multiple Sequence Comparison by Log-Expectation (*MUSCLE*),[68] and Tree-based Consistency Objective Function for Alignment Evaluation (*T-Coffee*).[69]

The resulting MSAs were employed to extract the consensus sequences by considering the frequency of each residue at every column of the alignment. The residues with a higher score than a certain threshold estimated for each column will conform to the positions (putative motifs) in the consensus. Both the *Jalview* software v2.11.1.4[70] and the *EMBOSS Cons* web server[71] (https://www.ebi.ac.uk/Tools/msa/emboss_cons/) were used for this aim.

2.4.2. Alignment-free method.

Lead compounds were analyzed with the Sensitive, Thorough, Rapid, Enriched Motif Elicitation (*STREME*) software[72] to discover fixed-length patterns (ungapped motifs) that were enriched in each cluster. The predictions were performed via its web server (https://meme-suite.org/meme/tools/streme), fully integrated within the widely-used *MEME* Suite of sequence analysis tools (https://meme-suite.org/meme/).[73] Control sets were generated by shuffling input peptides. The motif width was set between 3-5 amino acids length. *STREME* evaluated motifs

using a statistical test of the enrichment of matches for the target motif in the query set of sequences compared to a set of control sequences.[72]

2.4.3. Motif Search in PROSITE

Potential APPs were queried by the Motif Search tool (https://www.genome.jp/tools/motif/), integrated into the *GenomeNet Suite* (https://www.genome.jp/).[74] *PROSITE* Pattern and *PROSITE* Profile libraries[75] were only considered for the motif search within each cluster.

3. RESULTS AND DISCUSSION

3.1. Navigating and Mining the APPCS

3.1.1 Networks of APPs

Before creating CSN and HSPN, we conducted some analyses to decide the proper similarity threshold for both networks. The similarity cutoff to define edges is a mandatory parameter to create CSNs and it is optional for HSPNs. The selection of this threshold is not trivial because it modifies network topology and some properties like density, modularity, among others.[36] There is a lack of predefined standard values for this task because it depends on the input data, similarity relationships between nodes, and other aspects. Therefore, it is recommended to define this threshold case by case,[30] so we studied what would be the best cutoff values for both networks taking into account some network metrics.

Some previous articles have found similarity networks have an inversely proportional relationship between their similarity threshold and density values,[30,36] which means that networks with high cutoffs have fewer edges and are sparser. Both CSNs and HSPNs showed the mentioned behavior between similarity thresholds and density (Figure 1A, Tables SI2-1 and SI2-2). HSPN density values were much lower than the corresponding values of CSN, as we

expected because of the differences between methods building these networks, as explained in the Datasets and Methods section. In addition, density values were the same on HSPNs with a cutoff between 0.05 and 0.45 because the number of edges almost did not change (Table SI2-2). If the density is too high it would be complicated to interpret network topological features, while at low values it is likely to lose information, so an equilibrium between both extremes is a must.[34]

The ACC had a particular behavior, it increased at low and high similarity thresholds in the CSNs, and the HSPNs with high cutoffs even had larger ACC than networks with lower values (Figure 1B, Tables SI2-1 and SI2-2). These results were counterintuitive because the logic output would be that dense networks increase their connectedness, while the sparser ones decrease this parameter. Nonetheless, adding edges to some nodes does not guarantee that their neighbors are connected, which is measured by ACC, instead, the opposite could occur,[36] as is shown in the HSPN results (Figure 1B and Table SI2-2).

In ref [36] the authors studied the relationship between ACC and similarity threshold in networks of small molecules, obtained from the World of Molecular Bioactivity database and the PubChem Molecular Libraries Small Molecule Repository. They found that the ACC versus similarity threshold function of networks reconstructed from different datasets had a local maximum in a cutoff value associated with the best clustering outcome, and it would be the best option to choose.[36] Our results showed that the local maximum similarity thresholds for CSN and HSPN were 0.90 and 0.65, respectively (Figure 1B, Tables SI2-1 and SI2-2). Moreover, additional parameters were analyzed to confirm if these values were the best cutoffs, as explained below.

The modularity of both types of networks did not change too much at initial similarity cutoffs, but then these values increased until a global maximum of 0.94 and 0.96 for CSNs and HSPNs respectively (Figure 1C, Tables SI2-1 and SI2-2). Higher values of this network measure indicate if a community structure exists,[51] and it is associated with the number of communities (Figure 1D). An excessive number of communities is not desirable because it is likely that some of these clusters would be artifacts.[34] In both CSN and HSPN, the number of communities at high similarity thresholds increased too much compared to their low counterparts (Figure 1D, Tables SI2-1 and SI2-2), so these aspects were considered to choose the similarity cutoff for both types of networks.

The last parameter was the number of singletons (also well known as *outliers* or *atypical* sequences), unique APPs not similar to other nodes from our networks. These peptides are worth exploring because they could have new properties that enhance their antiparasitic activity. This network measure had a behavior as modularity, with no change at initial similarity thresholds, but increasing their values at higher cutoffs (Figure 1E, Tables SI2-1 and SI2-2), and it is not desirable to have an excessive number of singletons nor very few.

Considering all metrics from the 36 networks with different similarity cutoffs (18 cutoff values for both CSNs and HSPNs), the best similarity threshold for both types of networks was 0.65. The HSPN with this similarity cutoff was the local maximum point of ACC, and it presented intermediate values of density, modularity, communities, and singletons. Although CSN with a similarity threshold of 0.65 was not the local maximum of ACC, the other parameters were the most appropriate (Figure 1, Tables SI2-1 and SI2-2).

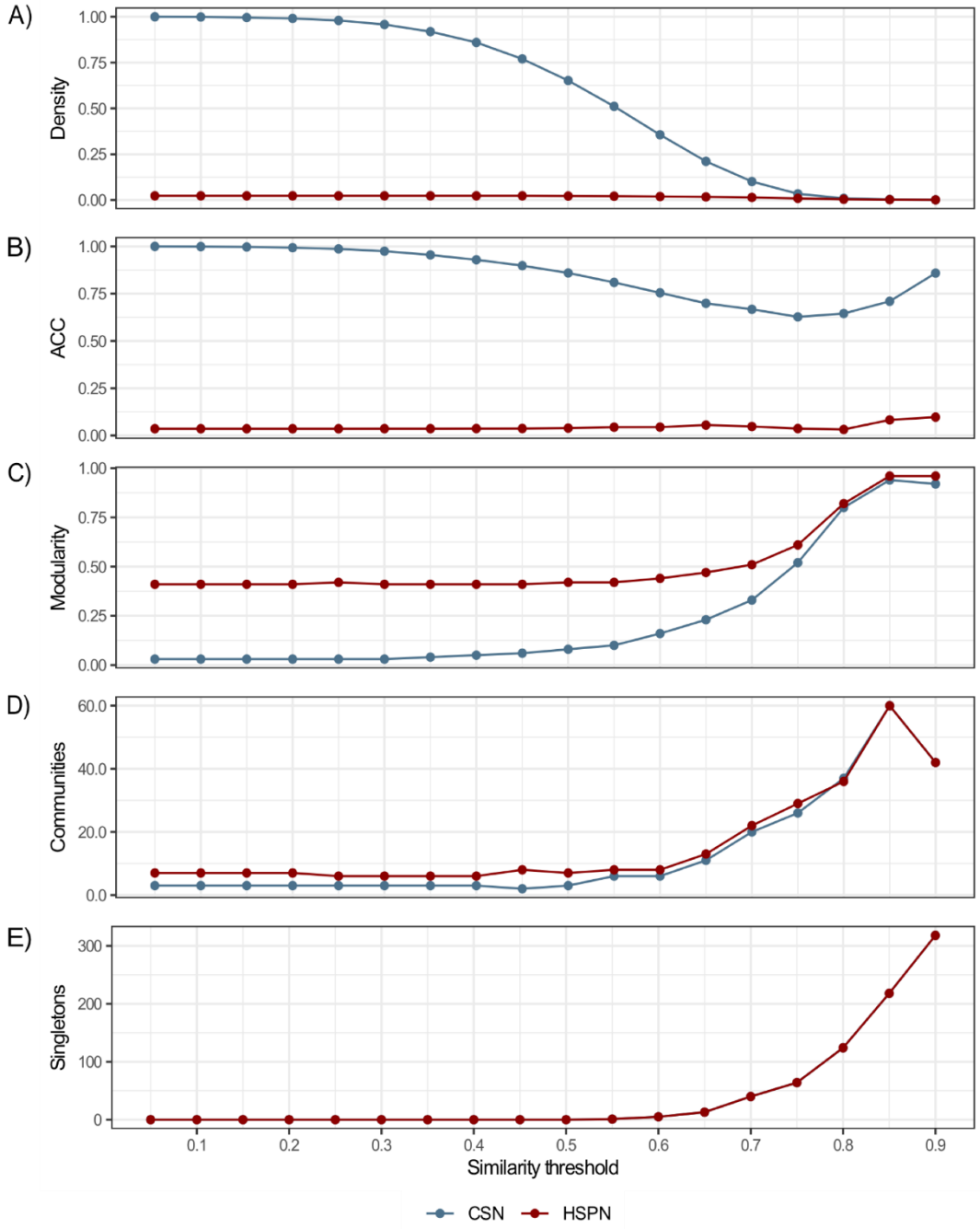


Figure 1. Network measures to determine the proper similarity threshold for CSN and HSPN. A) Density, B) Average clustering coefficient (ACC), C) Modularity, D) Communities, and E) Singletons. This Figure was created with ggplot2 R package[53] and edited with Inkscape.[48]

Therefore, we created CSN and HSPN applying the 0.65 similarity threshold. We obtained the giant components of both networks, and we constructed null models with the same number of

nodes and edges of the giant components. Figure 2 shows visualizations of both CSN and HSPN giant components and their null models, with nodes colored by their community and sized by their weighted degree or strength, calculated by summing up edge weights of the adjacent edges for each node.[49] The graphml files of all networks are available as SI3.

The size of all nodes from null networks was the same (Figures 2C and 2D) because we created these graphs using a random model that did not have APPs as nodes, so there were no weights to calculate strength. Another noticeable feature from Figure 2 was that the real networks had an apparent community structure, absent in the random models.

We calculated some networks metrics to measure in a formal way the differences between CSN and HSPN, as well as between real and random networks, in terms of their community structures and other aspects. The number of vertices, singletons, and connected components of CSN and HSPN were the same, which was shown in complete networks, giant components, and random models (Table 1). Indeed, the singletons of both networks were the same 13 APPs (see SI1-B), so we had only one set of these unique nodes for further analysis.

Density and ACC of HSPN tended to have lower values than CSN, while the opposite occurred with modularity, communities, diameter, and ASP. This behavior is related to the way each network model is constructed and the number of edges obtained by each process, as we explained in Materials and Methods. If a network has many links, as is the case of CSNs, the network's fraction of the possible number of edges and its connectivity would increase, which are the aspects measured by density and ACC, respectively.

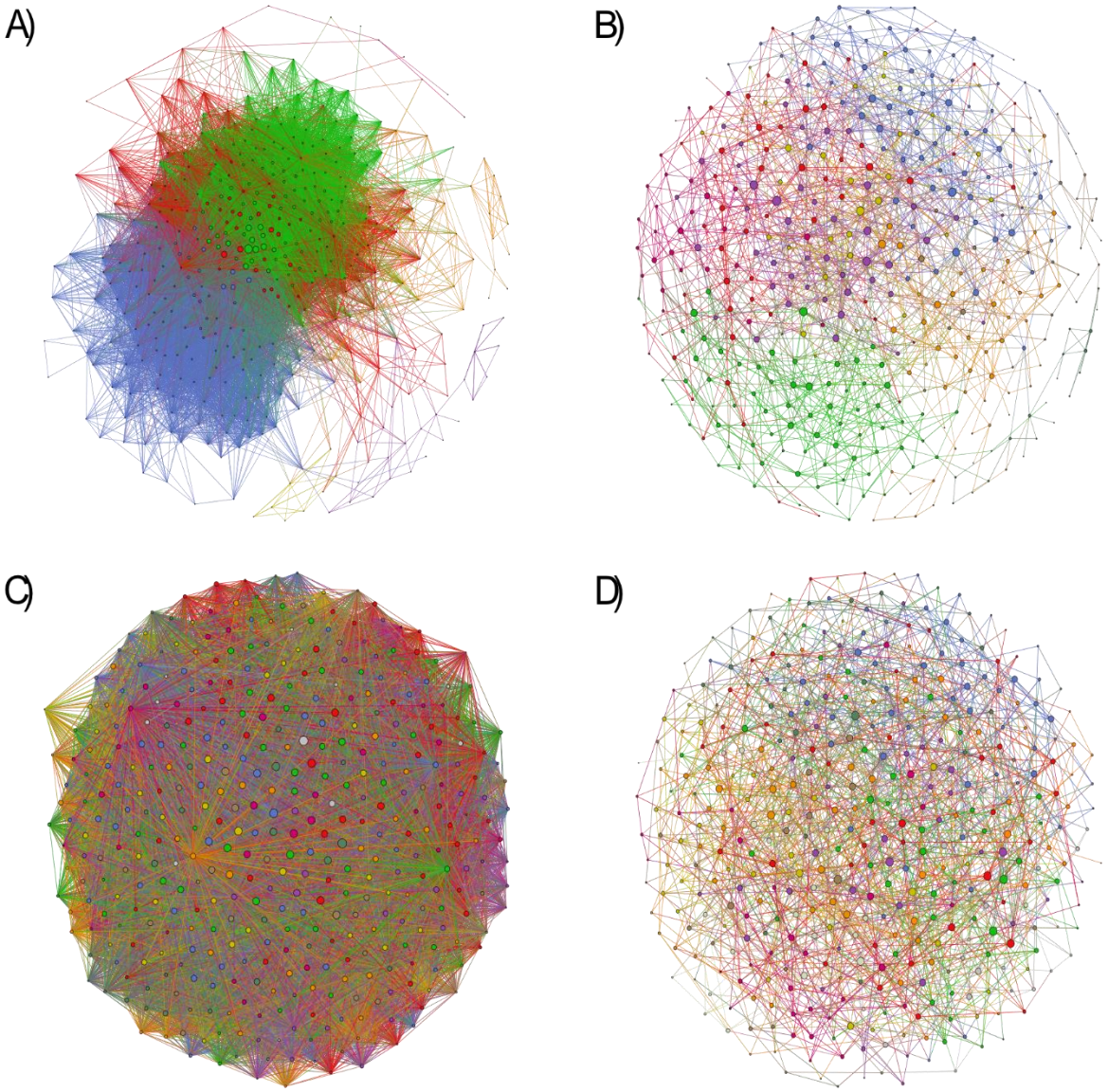


Figure 2. Network of APPs. A) CSN and B) HSPN giant components. C) CSN and D) HSPN random models with the same number of nodes and edges as A) and B). In all networks, nodes are colored by their community and sized by their weighted degree. All the visualizations were created with *Gephi*,[47] applying the *Fruchterman-Reingold* layout algorithm,[44] and edited with *Inkscape*.[48]

On the other hand, community detection in dense networks would be difficult due to their interconnectedness between nodes, so assigning a node to a specific cluster would be fuzzy.[49] This fact was observed in the lower modularity and number of communities values in CSNs compared to HSPNs. The same pattern appeared in the two measures related to reachability,

Table 1. Global Networks Properties of the Complete Graphs, Giant Components, and Random Models.

Network[a]	CSN	HSPN	CSN GC	HSPN GC	CSN RM	HSPN RM
Vertices	415	415	405	405	405	405
Edges	19,302	1,564	19,294	1,557	19,294	1,557
Connected components	5	5	1	1	1	1
Density	0.2247	0.0182	0.2358	0.0190	0.2358	0.0190
ACC	0.6988	0.0551	0.6943	0.0480	0.2361	0.0226
Modularity	0.234	0.455	0.233	0.452	0.071	0.335
Communities	11	15	7	9	10	12
Singletons	13	13	0	0	0	0
Diameter	∞	∞	9	12	2	6
ASP	∞	∞	2.254	3.732	1.764	3.166

[a]All the measures were calculated with igraph.[46] GC: giant component, RM: random model, ACC: average cluster coefficient, ASP: average shortest path.

diameter and ASP (Table 1), a logical result because dense networks like CSN have more possible paths to reach a node from another one, so the diameter and ASP from CSN were lower than the corresponding values in HSPN. The diameter and ASP for both complete networks were assigned to be infinite because these models had more than one component, so it is not possible to link some of their nodes, and for convenience these values are infinite.

Comparing the giant components with their random model counterparts, the number of vertices, edges, connected components, density, and singletons were the same. However, modularity, number of communities, and ACC values of giant components were greater than the random models (Table 1). Hence, the real networks showed a better community structure and neighbor connectivity, as is shown visually in Figure 2. In CSN and HSPN, communities could be APP families that share certain chemical and structural properties. The diameter and ASP values were also greater in real networks, so the reachability of these graphs was lower compared to the null models.

In addition, we plotted the degree distributions of giant components and random models from CSN and HSPN to explore some properties and if these networks behave as general models.[49]

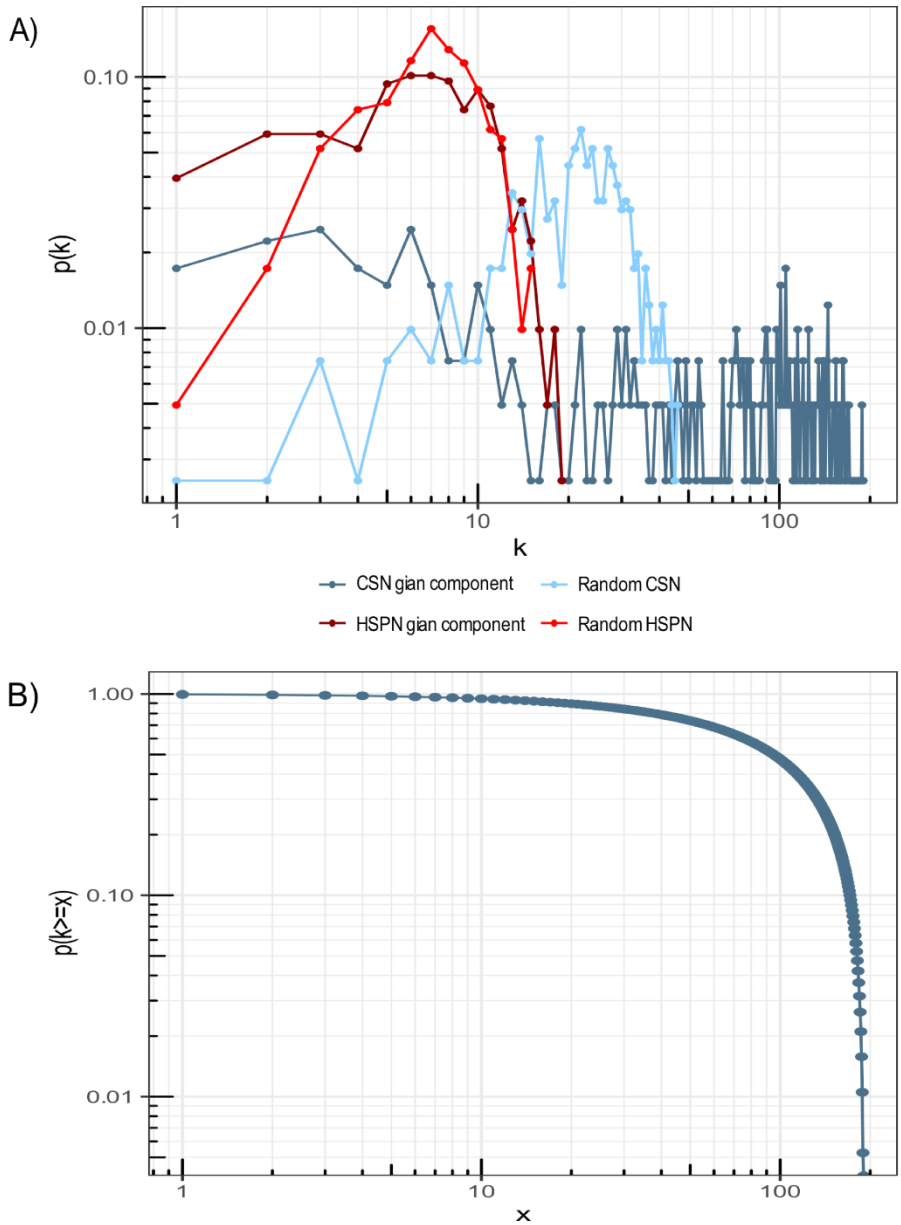


Figure 3. A) Degree distributions in a log-log scale of the giant components and random models from both CSN and HSPN. The horizontal axis is vertex degree *k* and the vertical axis is the probability of a node to have a degree of *k*. B) Complementary cumulative distribution function of the CSN giant component degree. The horizontal axis is vertex degree *x* and the vertical axis is the probability for finding a node of degree *k* greater than or equal to *x*. This Figure was created with *ggplot2* R package[53] and edited with *Inkscape*.[48]

The degree of HSPN giant component and their random models are distributed normally, as is shown with their bell-shaped distribution (Figure 3A), revealing the random behavior of HSPN. The randomness in HSPN was expected due to the way this network is constructed, following the half-space proximal test, as was explained in Materials and Methods. By contrast, the CSN giant component pattern distribution was not as evident as the other networks in Figure 3A, so we visualized its degree distribution with the complementary cumulative distribution function (CCDF) because it reflects these patterns in a better way.[34] The cumulative degree distribution showed that in CSN the probabilities for finding a node with degree x or higher were similar across different degree values (Figure 3B), so there was no power-law behavior due to the lack of scale invariance between degree and CCDF.[76]Therefore, the CSN was more related to a random model, as well as HSPN.

Moreover, the METNs showed valuable information about the APPCS. We observed that most of the APPs come from synthetic constructs, which are observed as the largest node or hub in the network, and a few of them are derived from parasites, bacteria, and animals (Figure 4A). As we expected, the most prevalent functions were antiparasitic and antimicrobial, but some of the APPs have been associated with antibacterial (Gram-positive and Gram-negative), antifungal, anticancer, among other activities (Figure 4B). The most predominant pathogen targets were some bacteria such as *Escherichia coli* and *Staphylococcus aureus*, although we also found parasites such as *Plasmodium*, *Leishmania*, and *Trypanosoma* (Figure 4C). The results of pathogen targets can be biased because there are much more antimicrobial essays made in bacteria than in parasites.[13] Regarding databases, most of the APPs come from DBAASP,[77] SATPdb,[78] ParaPep,[79], and APD[80](Figure 4D). ParaPep is one of the biggest databases of

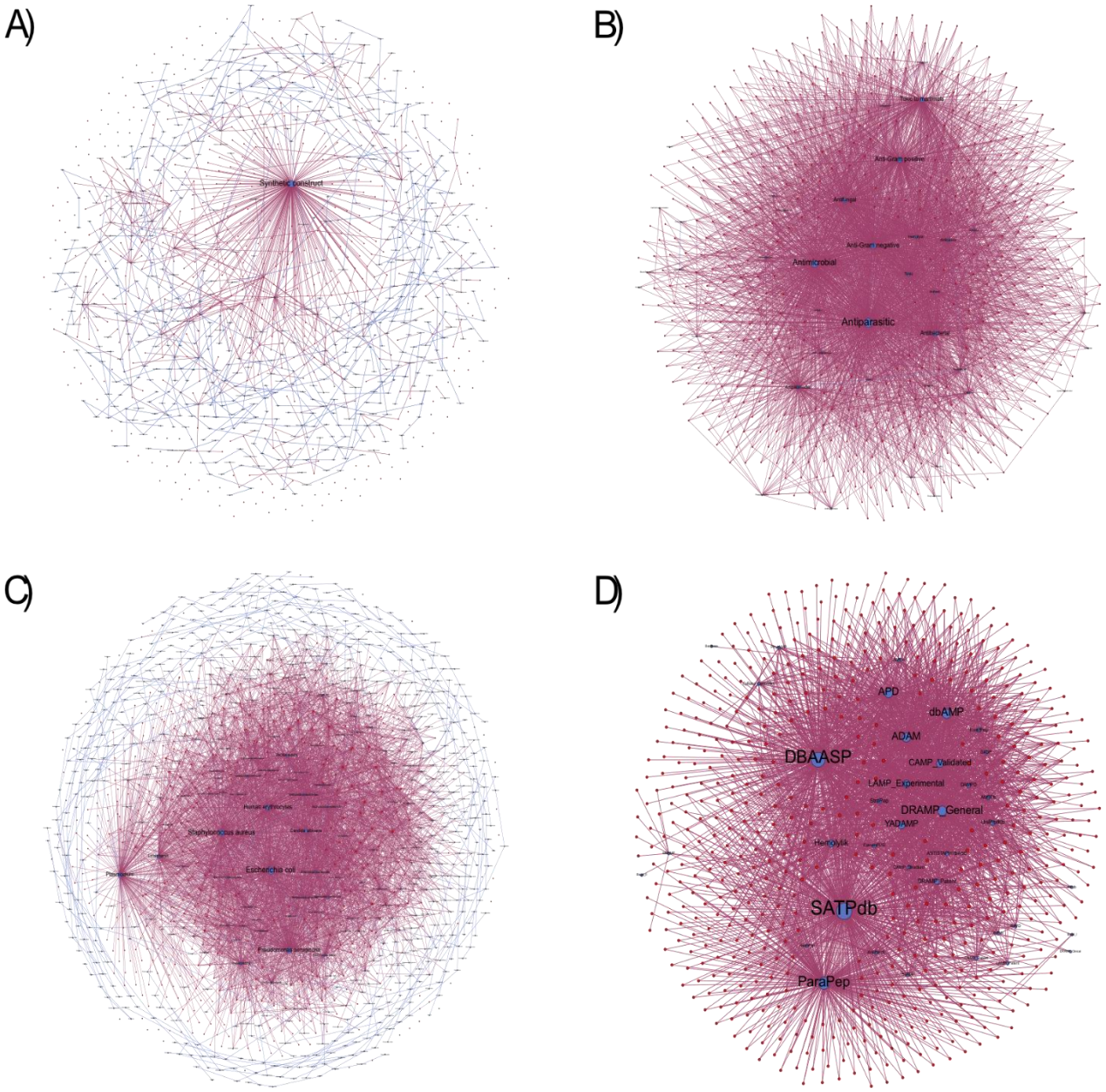


Figure 4. METNs with metadata of A) origin, B) function, C) target pathogen and D) database. In all networks, red nodes are APPs and the blue ones are metadata, and all of them are sized by their degree. All the visualizations were created with *Gephi*,[47] applying the *Fruchterman-Reingold* layout algorithm,[44] and edited with *Inkscape*.[48]

validated APPs,[79] so including its information in starPepDB helped us to map the known APPCS.

3.1.2 Centrality analysis and influential but non-redundant APPs

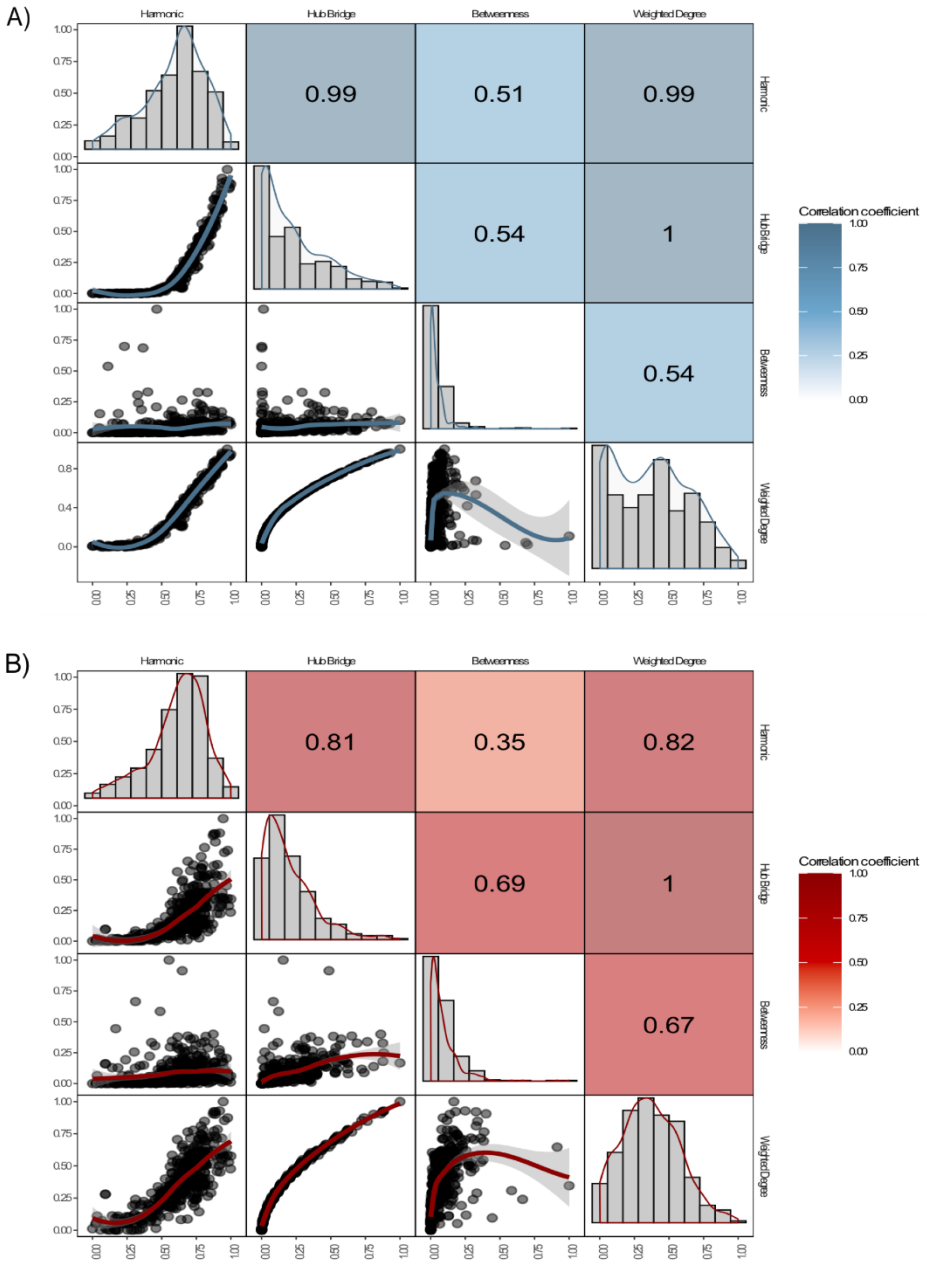


Figure 5. Spearman correlation analysis of centrality measures from A) CSN and B) HSPN. In A) and B) the distribution of each variable, the pairwise scatter plots with a fitted line, and the values of the spearman correlation are shown in the main diagonal, bottom diagonal, and upper diagonal respectively. On the right side of each correlogram, the legend color shows the scale of the correlation coefficients. This figure was created with *corrmorant* R package[55] and edited with *Inkscape*.[48]

The results of normalized centrality measures for all nodes from both CSN and HSPN can be found as SI4-A and SI4-B, respectively. These centrality measures consider different network

properties to identify influential nodes, but some metrics might have similar results. Hence, we studied possible correlations between these variables with the Spearman coefficient. We applied this correlation measure because the distributions of some of these metrics were skewed (see main diagonal at Figures 5A and 5B), so the assumption of normality was not satisfied, and other correlation coefficients like Pearson would not be a good choice.[81,82]

In both kinds of networks, the centrality measures of harmonic, hub-bridge, and weighted degree had a high positive correlation between them, greater than 0.80 in all cases, which is also shown graphically in the pairwise scatter plots between these variables (Figure 5). These results showed that the notion of the importance of these three centrality measures is highly related. Betweenness centrality had intermediate correlations values with the rest of the metrics, so it was also associated with the other measures, but at a lower level compared to the relationships among the others. Correlation analysis was supported by the common APPs in the top 50 most central nodes retrieved by different centrality measures, showing similar associations between the centrality measures (Tables SI2-3 and SI2-4). Considering these outcomes, the centrality sets from correlated measures were merged and tested together as queries against the validated datasets, as is explained in the next section.

An important aspect to consider with the obtention of central nodes was their representativeness by different communities, which was achieved with the chosen criteria, as is shown in Table SI2-5. Thus, our central nodes for each network and centrality measure belonged to different communities, which in these networks could be APPs families. In this way, the central nodes represented the APPCS and most of its potential APPs families. However, these influential nodes could be redundant in their communities because they would be highly similar to one another.

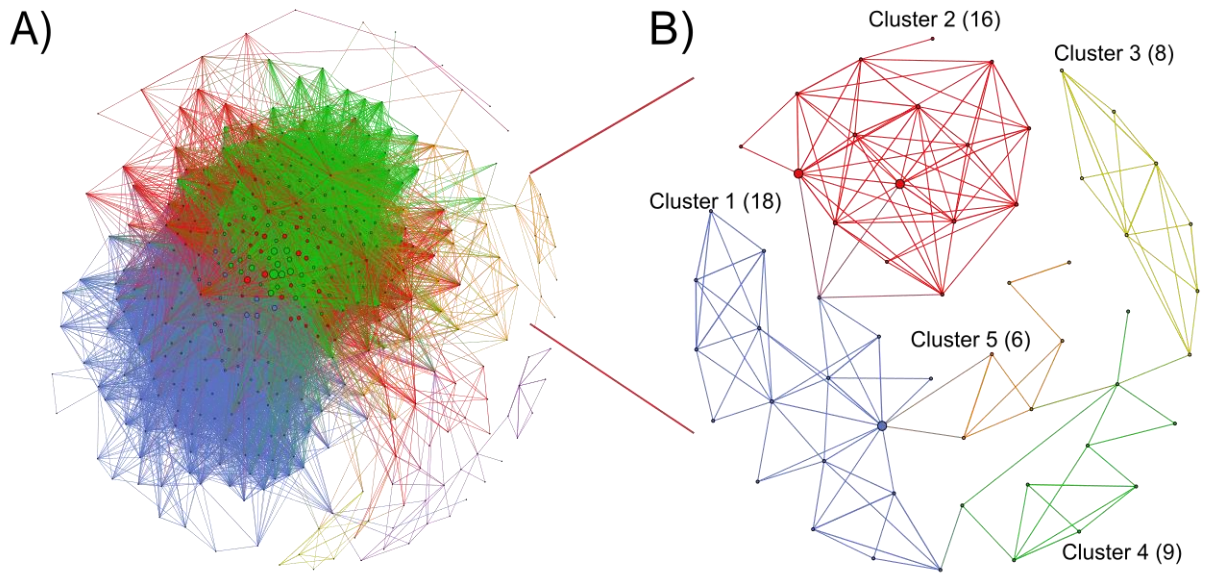


Figure 6. A) APPs CSN and B) subnetwork of 57 APPs from CSN's community 3 and its sub-communities. In both networks, nodes are colored by their community and sized by their weighted degree. All the visualizations were created with Gephi,[47] applying the *Fruchterman-Reingold* layout algorithm,[44] and edited with *Inkscape*.[48]

As a proof of concept for the redundancy of peptides inside communities of our networks, we extracted the APPs from CSN's community 3 and obtained sub-communities of this cluster using the *modularity optimization* algorithm, as is shown in Figure 6.

Then, we obtained the sequences and other physicochemical properties of some representative APPs from each sub-community (Table 2). We observed that several from CSN's community 3 had the same sequence length of 29, 10, 5, and 16 amino acids in 1, 2, 3, and 5 sub-communities, respectively. In addition, the same (or rather similar) amino acid residues composed of those peptides, and their physicochemical properties had similar values (Table 2). Hence, in general, it is expected that inside the most central nodes we could obtain highly similar APPs, so it may be better to extract some non-redundant sequences from the networks instead of just selecting the highest-ranked ones by each centrality measure. To remove this potential

Table 2. Sequences and some physicochemical properties of representative APPs from sub-communities of CSN's community 3.

Community	Name[a]	Sequence	Length	Charge[b]	Mol wt[b]	Hydrophobicity[b]
1	starPep_09852	GKGLXXGKXXGLXXG KXXGLXXGKXXGKR	29	6	1611.25	-0.26
	starPep_09855	GKGLXXGRXXGFXXG RXXGFXXGRXXGKR	29	6	1763.30	-0.37
2	starPep_20193	FPFFNQYVKL	10	1	1302.68	0.04
	starPep_20234	FPWFNQYVKL	10	1	1341.72	0.02
3	starPep_09474	FHPHE	5	0	665.77	-0.18
	starPep_11159	LHPHE	5	0	631.76	-0.19
4	starPep_04155	IASASCTTCICTCSCSS	17	0	1641.10	0.02
	starPep_02009	SCTTCVCTCSCCTT	14	0	1415.83	-0.05
5	starPep_13642	WIQXITXLXXQXXXPF	16	0	1145.51	0.15
	starPep_13916	YIQXITXLXXQXXXPF	16	0	1122.47	0.11

[a]ID of the peptides in starPepDB. [b]The physicochemical properties of APPs were calculated with ToxinPred server.[65] Mol wt: molecular weight.

redundancy in APPs, and obtain central but non-redundant peptides, we applied the *Scaffold extraction* plugin from the *starPep toolbox*, as was explained in the centrality analysis section of Materials and Methods. Thus, we obtained the most central and non-redundant APPs by each centrality measure and network, and we exported them as FASTA files, which are available as SI5. These sets of influential and unique APPs were used in the next section to retrieve new potential APPs by similarity searching.

The numbers of central nodes obtained from CSN with hub-bridge, weighted degree, and betweenness centrality measures were lower compared to the values derived from HSPN, while for harmonic centrality these values were almost the same for CSN and HSPN (Table SI2-5). Moreover, the numbers of central APPs from both CSN and HSPN with harmonic, hub-bridge, and weighted degree centrality measures were greater than one hundred APPs, which was not the case for betweenness centrality (Table SI2-5).

3.2. Multi-Query Similarity Searching Models for APPs

3.2.1. Performance of the best *mQSSMs*

Table 3. Antiparasitic datasets to calibrate/evaluate and compare the *mQSSMs* proposed in this report with several methods reported in the literature.

ID/Fasta file	Name	Number of sequences	Positive (APPs)	Negative (non-APPs)
D1/SI1-C	TR_starPep_AP	198	99	99
D2/SI1-D	TS_starPep_AP	62	31	31
D3/SI1-F	EX_starPep_AP	11,182	411	10,771
D4/SI1-F	B-TS_starPep_AP	57	26	31
D5/SI1-G	B-EX_starPep_AP	11,080	309	10,771

This table was adapted from Table1 of ref.[31] SI: supporting information, TR: training, TS: test, EX: external, B_TS: benchmarking test, B_EX: benchmarking external.

In our *mQSSMs*, the constant parameters were the similarity coefficient and group fusion model, while we varied the query set, target database, and similarity threshold (Scheme 1A). Five target databases (SI1C-G, five FASTA files) provided in the ref,[31] namely *D1-D5***,** were used to calibrate and evaluate the novel *mQSSMs*. These databases included balanced datasets (*D1*, *D2*, *D4*) with a similar proportion of positive and negative classes, and unbalanced datasets (*D3*, *D5*), which have much more negative instances than positive ones (Table 3). *D1-D3* databases contain sequences of lengths between 5 and 100 amino acids, while sequences from *D4-D5* have lengths between 5 and 30 amino acids.[31] The *D1-D3* datasets were recently used as training, test, and external validation data sets, respectively, to generate ML models by using genetic algorithm metaheuristics and random forest (RF); where the default configurations in the Weka tool v3.8 were applied.[31] On the other hand, *D4-D5* databases were previously used as external validation datasets to carry out a comparative study of the best RF-based classification models obtained for APPs discrimination. Here, we used these five benchmarking datasets of APP/non-APPs (Table 3) to compare the performance between our *mQSSMs* and with the algorithms reported in the literature for predicting APPs.

As we had twenty-one query sets, and seven similarity thresholds (0.3 to 0.9 with 0.1 step), we generated 147 different *mQSSMs*, which were evaluated with the five target databases (*D1-D5*), and their results were summarized in SI6 as four excel files containing output predictions

Table 4. Performance of the best nine *mQSSMs* to identify APPs, evaluated on the *D1-D5* databases.

Performance metrics (target database)[a]	**186Q_0.5 (HB-HC-Singletons)**	**178Q_0.5 (HC-Singletons)**	**165Q_0.5 HC**
	Best 3* mQSSMs *from CSN		
Accuraccy (**D1**)	0.934	0.914	0.894
KappaStatistic (**D1**)	0.869	0.828	0.788
AverageRecall (**D1**)	0.934	0.914	0.894
AveragePrecision (**D1**)	0.94	0.924	0.913
Accuraccy (**D2**)	0.952	0.952	0.935
KappaStatistic (**D2**)	0.903	0.903	0.871
AverageRecall (**D2**)	0.952	0.952	0.935
AveragePrecision (**D2**)	0.952	0.952	0.937
Accuraccy (**D3**)	0.991	0.991	0.991
KappaStatistic (**D3**)	0.86	0.86	0.863
AverageRecall (**D3**)	0.904	0.904	0.898
AveragePrecision (**D3**)	0.96	0.96	0.972
Accuraccy (**D4**)	0.947	0.965	0.965
KappaStatistic (**D4**)	0.894	0.929	0.929
AverageRecall (**D4**)	0.945	0.965	0.965
AveragePrecision (**D4**)	0.949	0.965	0.965
Accuraccy (**D5**)	0.991	0.991	0.992
KappaStatistic (**D5**)	0.832	0.832	0.842
AverageRecall (**D5**)	0.89	0.89	0.886
AveragePrecision (**D5**)	0.945	0.945	0.964
Performance metrics (target database)[a]	**200Q_0.5 HB-HC-Singletons**	**187Q_0.5 HB-HC**	**173Q_0.5 HC-Singletons**
	Best 3* mQSSMs *from HSPN		
Accuraccy (**D1**)	0.939	0.904	0.929
KappaStatistic (**D1**)	0.879	0.808	0.859
AverageRecall (**D1**)	0.939	0.904	0.929
AveragePrecision (**D1**)	0.946	0.919	0.936
Accuraccy (**D2**)	0.968	0.935	0.935
KappaStatistic (**D2**)	0.935	0.871	0.871
AverageRecall (**D2**)	0.968	0.935	0.935
AveragePrecision (**D2**)	0.968	0.937	0.937
Accuraccy (**D3**)	0.991	0.992	0.99
KappaStatistic (**D3**)	0.874	0.881	0.85
AverageRecall (**D3**)	0.921	0.915	0.898
AveragePrecision (**D3**)	0.955	0.969	0.957
Accuraccy (**D4**)	0.982	0.965	0.965
KappaStatistic (**D4**)	0.965	0.929	0.929
AverageRecall (**D4**)	0.984	0.965	0.965
AveragePrecision (**D4**)	0.981	0.965	0.965
Accuraccy (**D5**)	0.992	0.993	0.991

KappaStatistic (**D5**)	0.838	0.854	0.817
AverageRecall (**D5**)	0.904	0.9	0.882
AveragePrecision (**D5**)	0.935	0.958	0.939
Performance metrics (target database)[a]	**178Q_0.5 HC**	**206Q_0.5 HB-HC**	**219Q_0.5 HB-HC-Singletons**
	Best 3* mQSSMs *from both HSPN-CSN		
Accuraccy (**D1**)	0.919	0.909	0.944
KappaStatistic (**D1**)	0.838	0.818	0.889
AverageRecall (**D1**)	0.919	0.909	0.944
AveragePrecision (**D1**)	0.93	0.923	0.95
Accuraccy (**D2**)	0.935	0.952	0.984
KappaStatistic (**D2**)	0.871	0.903	0.968
AverageRecall (**D2**)	0.935	0.952	0.984
AveragePrecision (**D2**)	0.937	0.952	0.984
Accuraccy (**D3**)	0.991	0.993	0.992
KappaStatistic (**D3**)	0.868	0.89	0.885
AverageRecall (**D3**)	0.904	0.922	0.928
AveragePrecision (**D3**)	0.969	0.97	0.958
Accuraccy (**D4**)	0.965	0.965	0.982
KappaStatistic (**D4**)	0.929	0.929	0.965
AverageRecall (**D4**)	0.965	0.965	0.984
AveragePrecision (**D4**)	0.965	0.965	0.981
Accuraccy (**D5**)	0.992	0.993	0.992
KappaStatistic (**D5**)	0.845	0.866	0.856
AverageRecall (**D5**)	0.891	0.91	0.914
AveragePrecision (**D5**)	0.961	0.959	0.942

[a]Models performance on five **D1-D5** benchmarking datasets of APP/non-APPs, which are described in Table 3.
Q: query, CSN: Chemical space network, HSPN: half-space proximal network, HB: hub-bridge centrality, HC: harmonic centrality, mQSSMs: multi-query similarity searching models.

(active or inactive for APPs and non-APPs, respectively) for all models. We had ninety-eight *mQSSMs* for CSN (SI6-A) and HSPN (SI6-B), forty-nine for each network, forty-two models for the combination of both networks (SI6-C), and seven *mQSSMs* for the set of singletons (SI6-D). The query sets and number of queries for all the models are presented in Table SI2-6.

Table 4 shows the performance metrics of the best nine models to predict APPs, evaluated with the *D1-D5* validation datasets, whereas SI7-A contains the corresponding statistical parameters for all 147 *mQSSMs*. As it can be noted, the number of *Qs* included in the best 3 models for each network ranged from 165 to 219 sequences. It can also be observed that all these

best *mQSSMs* had good results according to their performance metrics, showing values of average recall, average precision, kappa statistic, and accuracy greater than 0.8.

We observed that the best similarity threshold was 0.5 in all *mQSSMs* (SI7-A). The best reference query sets were HC > WD > HB >> BE > singletons in both networks (see SI7-A and SI7-C for more details). However, the combination of query sets obtained with different centrality measures (in the same network and from both networks) was always better than any query set derived from a single centrality measure. Combinations of query datasets such as HB-HC-singletons in CSN, HSPN, and mixing both networks at the same time (CSN-HSPN) were the best *mQSSMs* (Tables 4 and Table 5). Similarly, the fusion of the thirteen singletons to any central query datasets enhanced the recovery of models, which is a logical result due to atypical nodes and central query sets represent the complete space of known APPs.

All the best 21 *mQSSMs* had successful predictive ability according to the average recall, average precision, kappa statistic, and accuracy performance metrics (see SI7-A and SI7-C for more details). Outstanding outcomes were attained by the *mQSSM* with 219 Qs from both networks (HB-HC-Singletons CSN-HSPN) and by using 0.5 as similarity threshold, with the previous performance metrics being greater than or approximately equal to 0.83 in *D4-D5* external validation datasets (Table 5, SI7-B, and SI7-C). Moreover, it can be stated that the predictions performed by the best *mQSSMs* were not random (MCC $\gg$ 0). It is important to remark that these antiparasitic *mQSSMs* had a strong-to-very strong predictive agreement since their test/external MCC values ranged from 0.834 to 0.965.

The best *mQSSMs* developed in this report for each network were used to perform a comparative study with *state-of-the-art* ML-based methods reported in the literature for

Table 5. Comparison between the best *mQSSMs* to predict APPs proposed in this study and those reported in the literature on the antiparasitic benchmarking test and external data sets.

Parameters	ProtDCal-AP_RF	ProtDCal-AP_RF_Hierarchical	AMPfun	178Q_0.5 (HC-Singletons) CSN	200Q_0.5 HB-HC-Singletons HSPN	219Q_0.5 HB-HC-Singletons HSPN-CSN
			D4 *Dataset*			
SN_{B-TS}	0.885	0.769	0.538	0.962	0.963	**0.963**
SP_{B-TS}	0.903	0.936	0.71	0.962	1	**1**
$Q\%_{B-TS}$	0.895	0.86	0.632	0.965	0.983	**0.983**
MCC_{B-TS}	0.788	0.721	0.252	0.929	0.965	**0.965**
			D5 *Dataset*			
SN_{B-EX}	0.799	0.783	0.45	0.896	0.875	**0.8893**
SP_{B-EX}	0.867	0.944	0.883	0.783	0.8123	**0.832**
$Q\%_{B-EX}$	0.865	0.939	0.871	0.9914	0.992	**0.993**
MCC_{B-EX}	0.306	0.45	0.165	0.834	0.839	**0.856**

AP: antiparasitic, RF: random forest, Q: query, CSN: Chemical space network, HSPN: half-space proximal network, HB: hub-bridge centrality, HC: harmonic centrality, *mQSSMs*: multi-query similarity searching models, SN: sensitivity, SP: specificity, Q%: accuracy, MCC: Matthew's correlation coefficient, B-TS: benchmarking test, B-EX: benchmarking external.

predicting APPs; the AMPfun sever,[32] and alignment-free quantitative sequence-activity models (AF-QSAMs) implemented in AMP-Discover.[31]

Regarding the outcomes achieved in the antiparasitic classification, the superiority of the three proposed models was remarkable, since the AMPfun model[32] and AMP-Discover AF-QSAMs[31] presented a weak predictive ability y (MCC < 0.26 and MCC < 0.45, respectively) on both benchmarking data sets (Table 5).

3.2.2. Statistical comparison

A sole accepted and established test doesn't exist for multiple comparison tests (MCT, for more detail, see http://sci2s.ugr.es/sicidm). In fact, models comparison and the selection of the best one is a staple among scientific investigations.[83] We selected the best *mQSSMs* by evaluating our models with various criteria (Q%, SE, SP, and MCC) on the five target databases, and applying a paired-parametric post hoc test (see SI7-C and SI7-D for more details). We determined the differences between our models by using several non-parametric statistical

tests.[58,63,64] In the first place, we applied an Iman–Davenport test[62] to check whether all the results obtained by the algorithms present any inequality, and in the case of finding some, then we can know, by using a Holm test[58,63,64] what algorithms partners' average results are dissimilar. That is to say, a Friedman's test, which rejected the null hypothesis that all predictors performed comparably on average. The same can be concluded from the results of an Iman and Davenport's test.

The MCTs showed according to the rankings' method that **219Q_0.5 HB-HC-Singletons HSPN-CSN** was the best algorithm, while **200Q_0.5 HB-HC-Singletons HSPN** and **178Q_0.5 HC-Singletons CSN** had the second and third best average value of ranking in the five validation datasets, in concordance with the results depicted by Tables 4 and Table 5 (see SI7-C and SI7-D for more details).

Besides, a second Iman–Davenport test[62] was carried out to detect if significant differences existed between our models and *state-of-the-art* algorithms to predict APPs (Table 5). In this sense, the null hypothesis (no-differences) was rejected for the case where the test values were higher than the critical value. Then, we performed the Holm test[58,63,64] and for the case of the benchmark datasets, we found statistically significant differences between our *mQSSMs* and literature methods, but no significant differences were found between our best *mQSSM* and the second and third best *mQSSMs* (SI7-D). Namely, in these five validation databases significant differences were observed by our best *mQSSMs*, at $\alpha = 0.05$. Figure 7 is a graphical representation of the average ranks (ranking scores) obtained by the best *mQSSMs* and literature methods in the *Friedman Test*, showing the relative position in the ranking of each of the six models and their differences with the best-ranked one (**219Q_0.5 HB-HC-Singletons HSPN-CSN**, see also SI7-D for more details). For example, **200Q_0.5 HB-HC-Singletons HSPN**

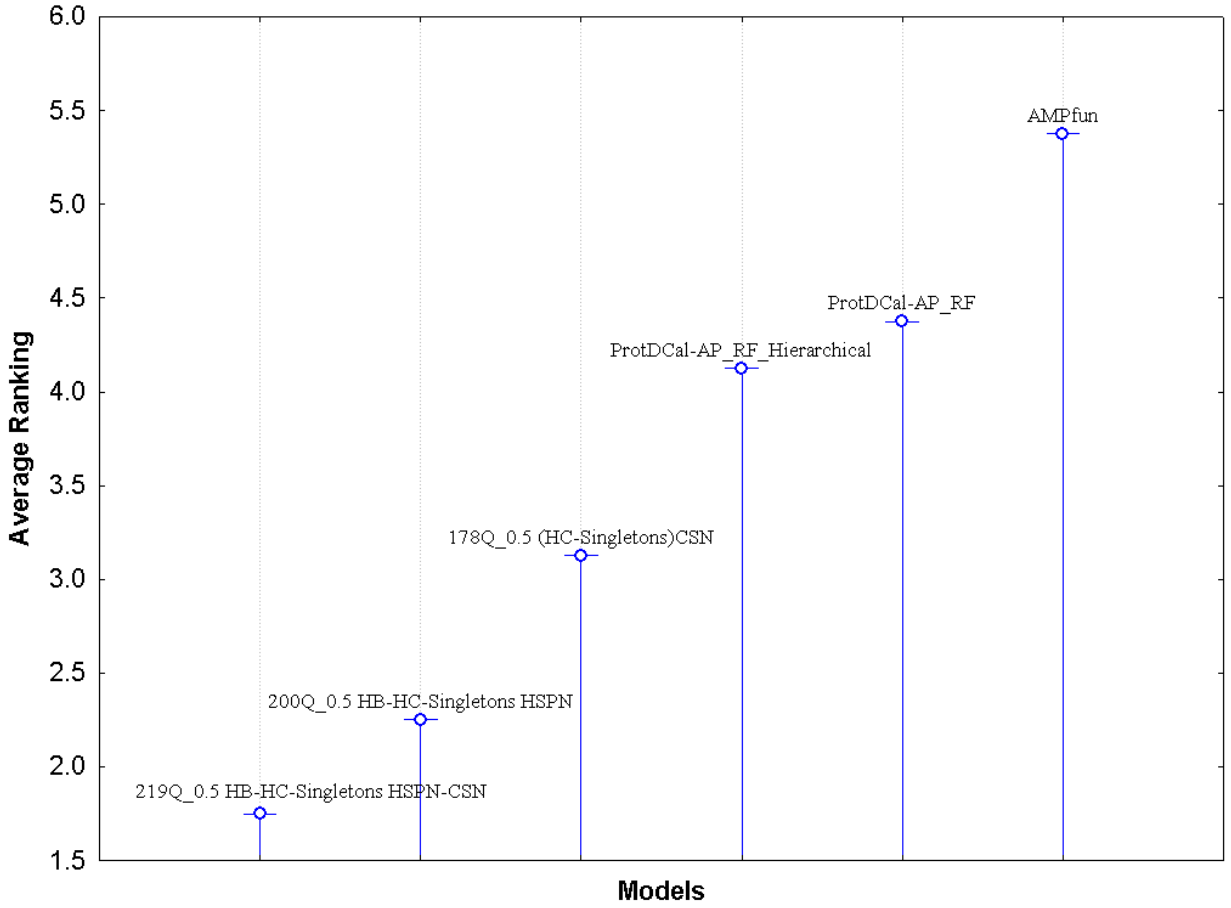


Figure 7. Average ranks obtained by each method in the Friedman Test. Friedman statistic (distributed according to chi-square with 5 degrees of freedom): 21.571. P-value computed by Friedman Test: 0.000631. Iman and Davenport statistic (distributed according to F-distribution with 5 and 35 degrees of freedom): 8.194. P-value computed by Iman and Davenport Test: 0.00003336.

performed similarly to the first ranked method, and slightly better than **178Q_0.5 HC-Singletons CSN**. Somewhat larger differences were detected between our *mQSSMs* and the literature models, including *ProtDCal-AP_RF_Hierarchical*, *ProtDCal-AP_RF*,[31] and AMPfun.[32]

3.3. Virtual Screening for Discovery of Putative APPs

Our starting search space was the entire *StarPepDB*, which contains about 45.000 peptides. After applying a series of filters with the *starPep toolbox* and some external servers, as well as the best *mQSSM*, we retrieved ninety-five leads that have never been associated with the

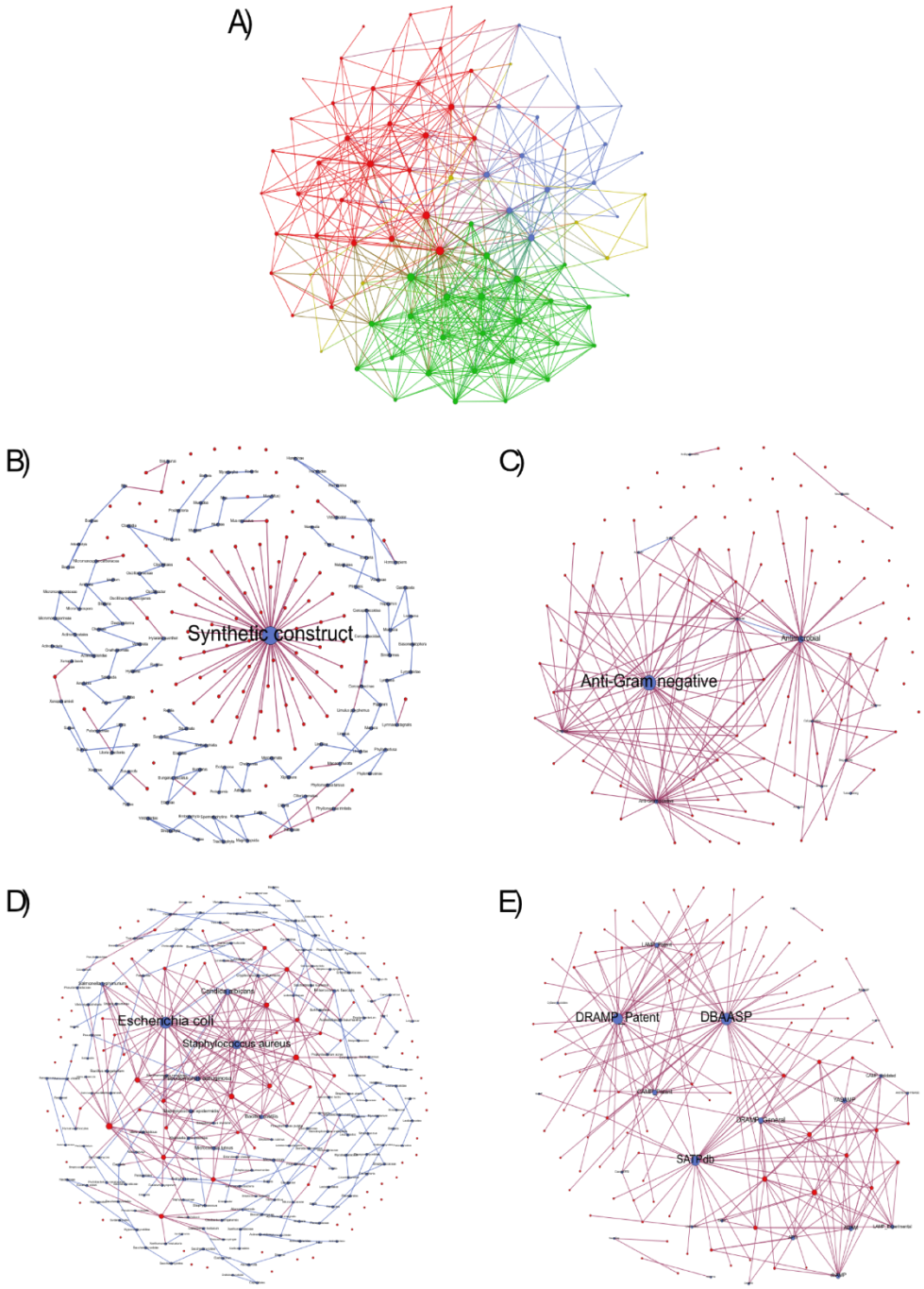


Figure 8. A) CSN of 95 potential AMPs, in which nodes are colored by their community and sized by harmonic centrality. METNs with metadata of B) origin, C) function, D) target pathogen, and E) database. In all METNs, red nodes are APPs, and the blue ones are metadata, and all of them are sized by their degree. All the visualizations were created with *Gephi*[47], applying the *Fruchterman-Reingold* layout algorithm,[44] and edited with *Inkscape*.[48]

antiparasitic activity, available in SI8-1 as a FASTA file. Scheme 2 summarizes the filtering process applied to retrieve the set of new potential APPs. In addition, Scheme 1B depicts the prospective virtual screening process to reduce most of the peptides from the initial search space, applying our best *mQSSM*. Figure 8A shows the CSN of the 95 potential APPs with its communities, which exhibit the diversity that these peptides still have (SI3-11 has the graphml file of this network).

To measure such similarity among the 95 leads, we calculated pairwise sequence identity among all of them. We found that most of the sequences share pairwise identity values below 30%, represented by the blue points in the heatmap depicted in Figure 9A. Figure 9B also shows the structural singularity of lead peptides because most pairwise identity values belong to 0-0.1, 0.1-0.2, and 0.2-0.3 bins of the histogram.

In addition, METNs of these peptides revealed some common characteristics shared by them. The origin of the leads was mainly from synthetic constructs (Figure 8B), as the METN showed for the APPs (Figure 3A). Regarding their annotated function, most of these peptides have antimicrobial and antibacterial (Gram-positive and Gram-negative) activities (Figure 8C), so their main pathogen targets are also bacteria such as *Escherichia coli* and *Staphylococcus aureus* (Figure 8D). Moreover, these potential APPs were mainly obtained from DRAMP,[84] DBAASP,[77] SATPdb,[78] among other databases (Figure 8E).

As far as we know, none of the lead 95 compounds reported in this study have been associated with the antiparasitic activity. However, some of them have been reported to have other activities. For instance, starPep_00322 or caerin 1.19 is a peptide derived from the skin secretion of the frog *Litoria gracilenta*, identified as a wide-spectrum antibiotic,[85] and as antiviral agent against the HIV.[86] There are other peptides associated with general antimicrobial

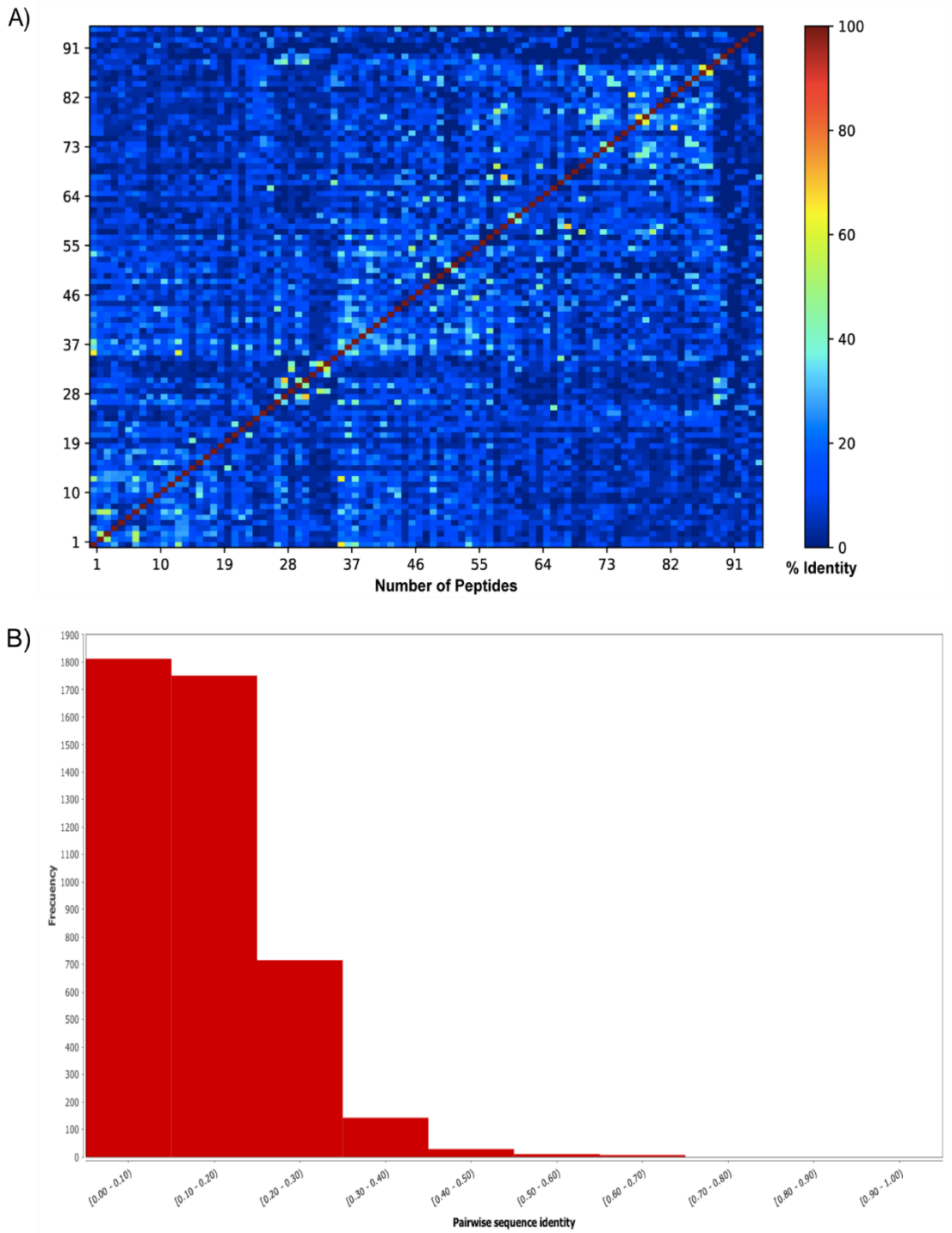


Figure 9. A) Heat map of pairwise sequence identity (all *vs* all), and B) Histogram of pairwise sequence identity of the 95 lead peptides.

activity such as starPep_15171 or N-Mag-C,[87] while others have antiviral activity like starPep_09816 or MG2d, antifungal activity such as starPep_17290 or Cap-LFampH-K,[88] and antibacterial activity like starPep_01732 or Phylloseptin-2.1TR.[89] Consequently, our method can

be considered as a drug repurposing strategy addressed to detect the antiparasitic activity on peptides with other previously reported activities.

3.4. Discovery of APPs Sequence Motifs.

To perform a wide exploration of motifs that could be determining a repurposed antiparasitic activity in peptides not labeled as APPs, the resulting ninety-five lead peptides identified after applying the above-mentioned filtering steps (Scheme 2), were clustered by mapping them onto the CSN space (Figure 8A). Thus, we identified five clusters; four out of the five contained members sharing some network regularities/properties, but the fifth cluster was selected to store singletons (peptides identified as atypical in the CSN). The five clusters were made up of twenty-six, nine, thirty-five, eighteen, and seven members, respectively (SI8-2-6 are FASTA files with sequences of the five clusters). The sequence diversity within each cluster was evaluated by all against all global alignments, reaching an overall identity lower than 30% in all clusters, which means that is very unlikely to find homologous sequences within each cluster.[90] This analysis confirmed the structural singularity of the ninety-five APPs considered as new sequence scaffolds.

In this sense, the five clusters were screened for discovering sequence patterns/motifs among these peptides that have been identified as potential APPs. As they represent new structural and singular scaffolds, new motifs accounting for the antiparasitic activity should be found. The motif search was performed by using different motif identification algorithms, including MSA, STREME,[72] and PROSITE.[75] We applied MSA algorithms developed after the classical ClustalW,[91] so that they can deal with the sequence diversity shown in each cluster and thus, detect more accurately any conserved signature or motif.

In this sense, MAFFT,[67] MUSCLE,[68] and T-Coffee[69] were applied to carry out MSAs in each cluster. The philosophy behind each MSA algorithm is different to improve alignment quality. MAFFT uses Fourier transform (FFT) to optimize protein alignments based on the amino acid sequence properties and have included iterative steps to refine the alignments,[67] while MUSCLE combines alignment-free (k-*mers* counting) and alignment-based (Kimura) distances to perform the progressive alignment, which is controlled by a log-expectation score function, and also includes iterative refinement alignment steps.[68] On the other hand, T-Coffee constructs progressive MSA by combining information derived from global and local alignment.[69]

Each MSA algorithm provided a consensus sequence that was estimated by the Jalview[70] and the EMBOSS Cons.[71] As EMBOSS Cons gives a more legible output, only displaying high scored amino acids/positions (capital letters), less scored but positive residues (lower-case letters), and non-consensus positions (X) that are under the threshold score, we identified the motifs using this software. Non-consensus positions were complemented by the visual inspection of the corresponding positions in the Jalview software and the Seq2Logo (available at http://www.cbs.dtu.dk/biotools/Seq2Logo)[92] by using default parameters. Table 6 depicts the consensus motifs, unraveled by each MSA algorithm, as well as the frequency of these motifs in the 550 APPs from *starPepDB* and the 95 lead compounds reported in this study. In general, most of the motifs had a low frequency of occurrence on both 550 APPs and 95 lead compounds, being the most frequent motif *KxxG* (x being any amino acid) with 120 occurrences in the 550 APPs and 20 in the 95 lead chemicals (Table 6). The low frequency from most of the motifs obtained by MSA could suggest they are novel signatures for characterizing APPs, so these motifs can be considered as scaffolds to search for new APPs. Moreover, alignments and sequence logos by each of the clusters are available at SI9.

Table 6. Discovered Motifs by Multiple Sequence Alignment.

No	Motif	Frequency of occurrence 550 APPs/95 potential APPs	EMBOSS Consensus	Frequency of occurrence 550 APPs/95 potential APPs	Cluster	Cluster size	MSA Method
1	K[fl]GK	22/4	KxGk	31/6			MAFFT
2	kK[fl][ga]K	15/3	kKxxK	57/14	1	26	MUSCLE
3	K[fy][fl]G	0/2	KxxG	117/20			T-Coffee
4	RK[vi]AL	0/0	RKxAL	0/0			MAFFT
5	aLLAL	0/0	axLAL	8/3	2	9*	MUSCLE
6	K[l]K[pa]RPa	0/0	KKxRPa	0/0			T-Coffee/ MUSCLE
7	L[kl]I[la]RK	0/0	LxIxRK	0/0			MAFFT
8	IL[kr]K	2/1	ILxK	6/2	3	35*	MUSCLE
9	r[ilv]I[il]K	0/0	rxIxK	6/2			T-Coffee
10	RWR[rw]r[mrs]RR	0/0	RWRxrxRR	0/0	4	18	MAFFT MUSCLE T-Coffee
11	L[ap]L[lp]L	0/0	LxLxL	14/5	5 (singletons)	7	MUSCLE

*The MSA quality of clusters 2 and 3 was improved by removing noised peptides, so we removed starPep_36552 from cluster 2, and starPep_16010-starPep_16459 from cluster 3.

To perform a wide motif search, unaligned patterns in the peptides should be also discovered. In this sense, the alignment-free approach STREME was used to find enriched patterns ranging from three to five amino acids length within the peptide clusters. STREME has been reported as the most accurate and sensitive algorithm among its competing state-of-art partners[72] (*e.g.*, DREME,[93] HOMER,[94] MEME[73]). Unlike previously algorithms, STREME counts efficiently position matches by using a position weight matrix (PWM) representing the motif candidate and also creates a Markov Model of a user-specified order from the control sequences. Both elements are considered when counting motif matches, keeping away the search on those that are mere artifacts of lower-order statistics of the input sequences.[72] Table 7 displays enriched motifs found within each cluster with respect to the control sequences. Motifs appearing in more than 30% of the query sequences were listed according to their statistical significance or score. We observed that motifs obtained with STREME also had a low frequency of occurrence in the 550 APPs and the 95 lead compounds, so these motifs can be new patterns to search novel APPs (Table 7).

Table 7. Discovered Motifs by STREME.

No	Motif	Cluster	Cluster size	Matches in positive sequences.	Matches in control sequences	Score	Frequency of occurrence 550 APPs/95 potential APPs
1	GAI			15	1	2.0e-005	6/2
2	LHS	1	26	11	0	1.3e-004	7/5
3	GKF			12	2	1.9e-003	7/5
4	PRPY			4	0	4.1e-002	0/1
5	ALKKA	2	9	3	0	1.0e-001	2/2
6	KKALL			3	0	1.0e-001	4/2
7	RLGI			8	0	2.5e-003	0/1
8	L[IA]KKF	3	35	7	0	5.6e-003	0/0
9	GLL			9	1	6.7e-003	15/1
10	WQWR			8	0	1.4e-003	8/2
11	MRR	4	18	7	1	2.0e-002	3/4
12	RRF			5	0	2.3e-002	2/2
13	LLLRL	5	7	2	0	2.3e-001	0/0

APPs: antiparasitic peptides.

Table 8. Discovered Motifs found in PROSITE.

No	Motif	Cluster	Hit Peptide	PROSITE Database	Match with	Signature	Frequency of occurrence 550 APPs/95 potential APPs
1	AGLQFPV	1	starPep_36218	Pattern	[AC]GLxFPV	Histone H2A	0/1
2	CGETCVLGTC		starPep_10020		C[GA]E[ST]C[FTV][GLTI]G[TSK]C	Cyclotides Moebius	0/1
3	CYCRIPACLAGERRYGTCFYRRRVWAFCC		starPep_01640		CxCx(3,5)Cx(7)GxCx(9)CC	Mammalian defensins	0/1
4	DAIWNLLRQAQEKFG		starPep_17290	Profile	ECIWHLLQRMQQLFGHGGKDP	Transferrin-like domain	0/1
5	GSAFCGETCVLGTCYTPDCSCTALVCLKN		starPep_10020		GLPVCGETCVWGPCNTPGCTCKWPVCYRN	Cyclotides	0/1
6	KMDSRWRWKSCKK	4	starPep_27296	Profile	KMDCRWRWKCCKK	Myotoxins_2	0/1

APPs: AntiParasitic Peptides.

Lastly, we also queried the peptide clusters against PROSITE Pattern and PROSITE Profile databases[75] by using the search engine Motif Search of the GenomeNet suite.[74] Significant hits were only found among a few members of clusters one and four (Table 8). Matching patterns and profiles can be straightforward associated with AMP-related signatures such as the histone 2A,

cyclotides, mammalian defensins, and myotoxins. Although transferrin-like domains are found in many proteins with diverse functions, some of them like the mammalian blood serotransferrin may have an antibacterial effect by removing toxic free iron from the blood, as well as the lactoferrin, found in the mammalian milk, which showed antimicrobial activity.[88]

As we mentioned before, motifs listed in Tables 6 and 7 were searched against the APPs registered in StarPepDB and the 95 lead compounds to discriminate the possible new signatures from the existing ones. We need to consider that new motifs should not appear in any of the registered APPs or should be at a very low frequency, which was the case for most of the motifs obtained by MSA and STREME methods.

4. CONCLUSIONS

A novel approach based on network science methods and similarity searches was introduced to explore the APPCS. We explored the chemical space of starPepDB with three types of networks (CSNs, HSPNs, METNs) and *mQSSMs* to retrieve valuable information from this database. We demonstrated that the pipeline developed in this research outperformed *state-of-the-art* ML models available for APP prediction with statistically significant differences. The novel *mQSSMs* were comparatively tested with the largest experimentally validated non-redundant peptide set reported to date and they largely outperformed several methods from the literature. Thus, we have arrived at a novel computational strategy regardless machine learning algorithms that recognizes APPs at high effectivity and reliability. This strategy promises to support research aimed at repurposing peptides associated with different activities as promising leads for a specific bioactivity. In fact, as a result of our method and other filters, we proposed 95 repurposed lead compounds as potential APPs that have not been associated with this activity

until now. Moreover, we explored sequence similarities and motifs shared by these leads and discovered some promising common motifs that can serve as templates for searching novel APPs.

ASSOCIATED CONTENT

The Supporting Information is available free of charge at Zenodo: https://doi.org/10.5281/zenodo.5650160. **Supporting information 1 (SI1)** contains Fasta files of 550/415/405 APPs used to generate networks as well as all Fasta files of the five benchmarking datasets of APP/non-APPs used to compare the performance of our similarity searching method from the literature algorithms. **Supporting information 2 (SI2)** contains an MS word file with Tables of parameters of similarity threshold analysis from CSN and HSPN of APPs; common APPs in the top 50 of most central nodes from CSN and HSPN retrieved by weighted degree, hub-bridge, harmonic, and betweenness centrality measures; the number and community membership percentage of the most central APPs by weighted degree, hub-bridge, harmonic, and betweenness centrality measures, and query sets and number of queries from the multi-query similarity searching models to find novel APPs for CSN, HSPN, and CSN-HSPN. **Supporting information 3 (SI3)** has graphml files of all the networks created in this study. **Supporting information 4 (SI4)** is an excel file with normalized centrality measures of CSN and HSPN. **Supporting information 5 (SI5)** has FASTA files of the most central and non-redundant APPs by each type of network and centrality measures, that is the query sets. **Supporting information 6 (SI6)** contains excel files with results of multi-query similarity searching models (Output Predictions). **Supporting information 7 (SI7)** has 3 kinds of files, namely SI7-A contains 3 folders with original results for each mQSSMs generated, as well excel file with statistical parameters. SI7-B is an excel file with the performance parameter of the best 21 mQSSMs proposed as well as the ranking of these models. Finally, SI7-C and SI7-D are pdf files with results of multiple comparisons of our mQSSMs and with literature algorithms, respectively. **Supporting information 8 (SI8)** contains FASTA files of 95 lead compounds and communities obtained from the CSN of these compounds. **Supporting information 9 (SI9)** contains a PowerPoint file with alignments and sequence logos of 95 lead compounds and communities obtained from the CSN of these peptides.

AUTHOR INFORMATION

Corresponding Author:

Prof. Dr. Yovani Marrero-Ponce. Email: ymarrero@usfq.edu.ec or ymarrero77@yahoo.es; Tel.: +593-2-297-1700 (ext. 4021). http://www.uv.es/yoma/ or http://ymponce.googlepages.com/home; **ORCID ID: http://www.orcid.org/0000-0003-2721-1142**

Author's ORCID:

Sebastián Ayala-Ruano (https://orcid.org/0000-0001-9756-6745)

Longendri Aguilera-Mendoza (https://orcid.org/0000-0002-2421-0382)
Noel Pérez (https://orcid.org/0000-0003-3166-745X)
Guillermin Agüero-Chapin (https://orcid.org/0000-0002-9908-2418)
Agostinho Antunes (http://orcid.org/0000-0002-1328-1732)
Ana Cristina Aguilar (https://orcid.org/0000-0001-7519-1583)

Notes

Availability of software and material: The starPep toolbox software and the respective user manual, as well as *mQSSMs*, are freely available online at http://mobiosd-hub.com/starpep.

Conflict of Interest: The authors declare no conflict of interest.

Funding*:* Declared none.

ACKNOWLEDGMENT

Yovani Marrero-Ponce (**M-P, Y**) thanks to the program *Profesor coinvitado* for a post-doctoral fellowship to work at Valencia University in 2020. **Y.M.-P.** and **Noel Pérez** acknowledge the support from Collaboration Grant 2019–2020 (Project **ID**16897) and Med Grant 2019-2020 (Project **ID**16905). A-C, G. and A, A were supported by the Strategic Funding UIDB/04423/2020 and UIDP/04423/2020 through national funds provided by FCT and European Regional Development Fund (ERDF), in the framework of the program PT2020.

ABBREVIATIONS

AMPs, antimicrobial peptides; AMPCS, AMPs' chemical space; APPs, antiparasitic peptides; APPCS, APPs' chemical space; CSN, chemical space network; HSPN, half-space proximal network, METN, metadata network; *mQSSM*, the multi-query similarity searching model; MCC, Matthews correlation coefficient; MDR, multidrug-resistant; AMR, antimicrobial resistance; CHDPs, host defense peptides; ML, machine learning; ACC, average clustering coefficient; ASP, average shortest path; WD, weighted degree; BE, betweenness; HC, harmonic; HB, hub-bridge; Qs, queries; ref, reference; SN, sensitivity; PR, precision; SP, specificity; Q%, accuracy; MSA, multiple sequence alignment; MAFFT, Multiple Alignment using Fast Fourier Transform; MUSCLE, Multiple Sequence Comparison by Log-Expectation; T-Coffee, Tree-based Consistency Objective Function for Alignment Evaluation, STREME, Sensitive, Thorough, Rapid, Enriched Motif Elicitation; CCDF, complementary cumulative distribution function; RF, random forest; AF-QSAMs, alignment-free quantitative sequence-activity models; MCT, multiple comparison tests.

5. REFERENCES

(1) Jones, K. E.; Patel, N. G.; Levy, M. A.; Storeygard, A.; Balk, D.; Gittleman, J. L.; Daszak, P. Global Trends in Emerging Infectious Diseases. *Nature* **2008**, *451* (7181), 990–993. https://doi.org/10.1038/nature06536.

(2) Andersson, D. I.; Balaban, N. Q.; Baquero, F.; Courvalin, P.; Glaser, P.; Gophna, U.; Kishony, R.; Molin, S.; Tønjum, T. Antibiotic Resistance: Turning Evolutionary Principles into Clinical Reality. *FEMS Microbiol. Rev.* **2020**, *44* (2), 171–188. https://doi.org/10.1093/femsre/fuaa001.

(3) WHO. Antimicrobial resistance https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance (accessed 2021 -05 -07).
(4) Mookherjee, N.; Anderson, M. A.; Haagsman, H. P.; Davidson, D. J. Antimicrobial Host Defence Peptides: Functions and Clinical Potential. *Nat. Rev. Drug Discov.* **2020**, *19* (5), 311–332. https://doi.org/10.1038/s41573-019-0058-8.
(5) Mahlapuu, M.; Björn, C.; Ekblom, J. Antimicrobial Peptides as Therapeutic Agents: Opportunities and Challenges. *Crit. Rev. Biotechnol.* **2020**, *40* (7), 978–992. https://doi.org/10.1080/07388551.2020.1796576.
(6) Lazzaro, B. P.; Zasloff, M.; Rolff, J. Antimicrobial Peptides: Application Informed by Evolution. *Science* **2020**, *368* (6490). https://doi.org/10.1126/science.aau5480.
(7) Magana, M.; Pushpanathan, M.; Santos, A. L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M. A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A. L.; Cherkasov, A.; Seleem, M. N.; Pinilla, C.; de la Fuente-Nunez, C.; Lazaridis, T.; Dai, T.; Houghten, R. A.; Hancock, R. E. W.; Tegos, G. P. The Value of Antimicrobial Peptides in the Age of Resistance. *Lancet Infect. Dis.* **2020**, *20* (9), e216–e230. https://doi.org/10.1016/S1473-3099(20)30327-3.
(8) van der Does, A. M.; Hiemstra, P. S.; Mookherjee, N. Antimicrobial Host Defence Peptides: Immunomodulatory Functions and Translational Prospects. In *Antimicrobial Peptides: Basics for Clinical Application*; Matsuzaki, K., Ed.; Advances in Experimental Medicine and Biology; Springer: Singapore, 2019; pp 149–171. https://doi.org/10.1007/978-981-13-3588-4_10.
(9) Browne, K.; Chakraborty, S.; Chen, R.; Willcox, M. D.; Black, D. S.; Walsh, W. R.; Kumar, N. A New Era of Antibiotics: The Clinical Potential of Antimicrobial Peptides. *Int. J. Mol. Sci.* **2020**, *21* (19), 7047. https://doi.org/10.3390/ijms21197047.
(10) Piyadasa, H.; Hemshekhar, M.; Altieri, A.; Basu, S.; Does, A. M. van der; Halayko, A. J.; Hiemstra, P. S.; Mookherjee, N. Immunomodulatory Innate Defence Regulator (IDR) Peptide Alleviates Airway Inflammation and Hyper-Responsiveness. *Thorax* **2018**, *73* (10), 908–917. https://doi.org/10.1136/thoraxjnl-2017-210739.
(11) Chow, L. N. Y.; Choi, K.-Y. (Grace); Piyadasa, H.; Bossert, M.; Uzonna, J.; Klonisch, T.; Mookherjee, N. Human Cathelicidin LL-37-Derived Peptide IG-19 Confers Protection in a Murine Model of Collagen-Induced Arthritis. *Mol. Immunol.* **2014**, *57* (2), 86–92. https://doi.org/10.1016/j.molimm.2013.08.011.
(12) Ho, S.; Pothoulakis, C.; Wai Koon, H. Antimicrobial Peptides and Colitis. *Curr. Pharm. Des.* **2013**, *19* (1), 40–47. https://doi.org/10.2174/138161213803903074.
(13) Roudi, R.; Syn, N. L.; Roudbary, M. Antimicrobial Peptides As Biologic and Immunotherapeutic Agents against Cancer: A Comprehensive Overview. *Front. Immunol.* **2017**, *8*. https://doi.org/10.3389/fimmu.2017.01320.
(14) Haney, E. F.; Straus, S. K.; Hancock, R. E. W. Reassessing the Host Defense Peptide Landscape. *Front. Chem.* **2019**, *7*. https://doi.org/10.3389/fchem.2019.00043.
(15) Torrent, M.; Pulido, D.; Rivas, L.; Andreu, D. Antimicrobial Peptide Action on Parasites. *Curr. Drug Targets* **2012**, *13* (9), 1138–1147. https://doi.org/10.2174/138945012802002393.
(16) Vagner, J.; Qu, H.; Hruby, V. J. Peptidomimetics, a Synthetic Tool of Drug Discovery. *Curr. Opin. Chem. Biol.* **2008**, *12* (3), 292–296. https://doi.org/10.1016/j.cbpa.2008.03.009.
(17) Pane, K.; Durante, L.; Crescenzi, O.; Cafaro, V.; Pizzo, E.; Varcamonti, M.; Zanfardino, A.; Izzo, V.; Di Donato, A.; Notomista, E. Antimicrobial Potency of Cationic Antimicrobial Peptides Can Be Predicted from Their Amino Acid Composition: Application to the

Detection of "Cryptic" Antimicrobial Peptides. *J. Theor. Biol.* **2017**, *419*, 254–265. https://doi.org/10.1016/j.jtbi.2017.02.012.
(18) Walsh, C. J.; Guinane, C. M.; Toole, P. W. O.; Cotter, P. D. A Profile Hidden Markov Model to Investigate the Distribution and Frequency of LanB-Encoding Lantibiotic Modification Genes in the Human Oral and Gut Microbiome. *PeerJ* **2017**, *5*, e3254. https://doi.org/10.7717/peerj.3254.
(19) Azkargorta, M.; Soria, J.; Ojeda, C.; Guzmán, F.; Acera, A.; Iloro, I.; Suárez, T.; Elortza, F. Human Basal Tear Peptidome Characterization by CID, HCD, and ETD Followed by in Silico and in Vitro Analyses for Antimicrobial Peptide Identification. *J. Proteome Res.* **2015**, *14* (6), 2649–2658. https://doi.org/10.1021/acs.jproteome.5b00179.
(20) Fuente-Nunez, C. de la. Toward Autonomous Antibiotic Discovery. *mSystems* **2019**, *4* (3). https://doi.org/10.1128/mSystems.00151-19.
(21) Porto, W. F.; Pires, A. S.; Franco, O. L. Computational Tools for Exploring Sequence Databases as a Resource for Antimicrobial Peptides. *Biotechnol. Adv.* **2017**, *35* (3), 337–349. https://doi.org/10.1016/j.biotechadv.2017.02.001.
(22) Capecchi, A.; Reymond, J.-L. Peptides in Chemical Space. *Med. Drug Discov.* **2021**, *9*, 100081. https://doi.org/10.1016/j.medidd.2021.100081.
(23) Agüero-Chapin, G.; Pérez-Machado, G.; Molina-Ruiz, R.; Pérez-Castillo, Y.; Morales-Helguera, A.; Vasconcelos, V.; Antunes, A. TI2BioP: Topological Indices to BioPolymers. Its Practical Use to Unravel Cryptic Bacteriocin-like Domains. *Amino Acids* **2011**, *2* (40), 431–442. https://doi.org/10.1007/s00726-010-0653-9.
(24) Xu, J.; Li, F.; Leier, A.; Xiang, D.; Shen, H.-H.; Marquez Lago, T. T.; Li, J.; Yu, D.-J.; Song, J. Comprehensive Assessment of Machine Learning-Based Methods for Predicting Antimicrobial Peptides. *Brief. Bioinform.* **2021**, No. bbab083. https://doi.org/10.1093/bib/bbab083.
(25) Cardoso, M. H.; Orozco, R. Q.; Rezende, S. B.; Rodrigues, G.; Oshiro, K. G. N.; Cândido, E. S.; Franco, O. L. Computer-Aided Design of Antimicrobial Peptides: Are We Generating Effective Drug Candidates? *Front. Microbiol.* **2020**, *10*. https://doi.org/10.3389/fmicb.2019.03097.
(26) Mitchell, M. *Complexity: A Guided Tour*; Oxford University Press, 2009.
(27) Barabási, A.-L. *Network Science*; Cambridge University Press, 2016.
(28) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996. https://doi.org/10.1021/ci9800211.
(29) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today* **2006**, *11* (23), 1046–1053. https://doi.org/10.1016/j.drudis.2006.10.005.
(30) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; García-Jacas, C. R.; Chavez, E.; Beltran, J. A.; Guillen-Ramirez, H. A.; Brizuela, C. A. Automatic Construction of Molecular Similarity Networks for Visual Graph Mining in Chemical Space of Bioactive Peptides: An Unsupervised Learning Approach. *Sci. Rep.* **2020**, *10* (1), 18074. https://doi.org/10.1038/s41598-020-75029-1.
(31) Pinacho-Castellanos, S. A.; García-Jacas, C. R.; Gilson, M. K.; Brizuela, C. A. Alignment-Free Antimicrobial Peptide Predictors: Improving Performance by a Thorough Analysis of the Largest Available Data Set. *J. Chem. Inf. Model.* **2021**, *61* (6), 3141–3157. https://doi.org/10.1021/acs.jcim.1c00251.

(32) Chung, C.-R.; Kuo, T.-R.; Wu, L.-C.; Lee, T.-Y.; Horng, J.-T. Characterization and Identification of Antimicrobial Peptides with Different Functional Activities. *Brief. Bioinform.* **2020**, *21* (3), 1098–1114. https://doi.org/10.1093/bib/bbz043.
(33) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Beltran, J. A.; Tellez Ibarra, R.; Guillen-Ramirez, H. A.; Brizuela, C. A. Graph-Based Data Integration from Bioactive Peptide Databases of Pharmaceutical Interest: Toward an Organized Collection Enabling Visual Network Analysis. *Bioinformatics* **2019**, *35* (22), 4739–4747. https://doi.org/10.1093/bioinformatics/btz260.
(34) Coscia, M. *The Atlas for the Aspiring Network Scientist*; arXiv [Preprint], 2021.
(35) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Beltran, J. A.; Tellez Ibarra, R.; Guillen-Ramirez, H. A.; Brizuela, C. A. Graph-Based Data Integration from Bioactive Peptide Databases of Pharmaceutical Interest: Toward an Organized Collection Enabling Visual Network Analysis. *Bioinformatics* **2019**, *35* (22), 4739–4747. https://doi.org/10.1093/bioinformatics/btz260.
(36) Zahoránszky-Kőhalmi, G.; Bologa, C. G.; Oprea, T. I. Impact of Similarity Threshold on the Topology of Molecular Similarity Networks and Clustering Outcomes. *J. Cheminformatics* **2016**, *8* (1), 16. https://doi.org/10.1186/s13321-016-0127-5.
(37) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B.-T. Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. *Mol. Divers.* **2006**, *10* (1), 39–79. https://doi.org/10.1007/s11030-006-8697-1.
(38) Zwierzyna, M.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design and Characterization of Chemical Space Networks for Different Compound Data Sets. *J. Comput. Aided Mol. Des.* **2015**, *29* (2), 113–125. https://doi.org/10.1007/s10822-014-9821-4.
(39) Chavez, E.; Dobrev, S.; Kranakis, E.; Opatrny, J.; Stacho, L.; Tejeda, H.; Urrutia, J. Half-Space Proximal: A New Local Test for Extracting a Bounded Dilation Spanner of a Unit Disk Graph. In *Principles of Distributed Systems*; Anderson, J. H., Prencipe, G., Wattenhofer, R., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2006; pp 235–245. https://doi.org/10.1007/11795490_19.
(40) Corral-Corral, R.; Chavez, E.; Del Rio, G. Machine Learnable Fold Space Representation Based on Residue Cluster Classes. *Comput. Biol. Chem.* **2015**, *59*, 1–7. https://doi.org/10.1016/j.compbiolchem.2015.07.010.
(41) Smith, T. F.; Waterman, M. S. Identification of Common Molecular Subsequences. *J. Mol. Biol.* **1981**, *147* (1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.
(42) Ashtiani, M.; Mirzaie, M.; Jafari, M. CINNA: An R/CRAN Package to Decipher Central Informative Nodes in Network Analysis. *Bioinformatics* **2019**, *35* (8), 1436–1437. https://doi.org/10.1093/bioinformatics/bty819.
(43) Cherven, K. *Network Graph Analysis and Visualization with Gephi*; Birmingham, UK, 2013.
(44) Fruchterman, T. M. J.; Reingold, E. M. Graph Drawing by Force-Directed Placement. *Softw. Pract. Exp.* **1991**, *21* (11), 1129–1164. https://doi.org/10.1002/spe.4380211102.
(45) Gilbert, E. N. Random Graphs. *Ann. Math. Stat.* **1959**, *30* (4), 1141–1144.
(46) Csárdi, G.; Nepusz, T. The Igraph Software Package for Complex Network Research. *Int J Complex Sys* **2006**, *1695*, 1–9.
(47) Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks; 2009.
(48) Inkscape. *Inkscape Project*; 2021.

(49) Newman, M. *Networks*; Oxford University Press, 2018.
(50) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast Unfolding of Communities in Large Networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008* (10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.
(51) Newman, M. E. J. Modularity and Community Structure in Networks. *Proc. Natl. Acad. Sci.* **2006**, *103* (23), 8577–8582. https://doi.org/10.1073/pnas.0601602103.
(52) Barrat, A.; Barthélemy, M.; Pastor-Satorras, R.; Vespignani, A. The Architecture of Complex Weighted Networks. *Proc. Natl. Acad. Sci.* **2004**, *101* (11), 3747–3752. https://doi.org/10.1073/pnas.0400087101.
(53) Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Use R!; Springer-Verlag: New York, 2009. https://doi.org/10.1007/978-0-387-98141-3.
(54) Ashtiani, M.; Salehzadeh-Yazdi, A.; Razaghi-Moghadam, Z.; Hennig, H.; Wolkenhauer, O.; Mirzaie, M.; Jafari, M. A Systematic Survey of Centrality Measures for Protein-Protein Interaction Networks. *BMC Syst. Biol.* **2018**, *12* (1), 80. https://doi.org/10.1186/s12918-018-0598-2.
(55) Link, R. *Corrmorant: Flexible Correlation Matrices Based on Ggplot2*; 2021.
(56) Lafita, A.; Bliven, S.; Prlić, A.; Guzenko, D.; Rose, P. W.; Bradley, A.; Pavan, P.; Myers-Turnbull, D.; Valasatava, Y.; Heuer, M.; Larson, M.; Burley, S. K.; Duarte, J. M. BioJava 5: A Community Driven Open-Source Bioinformatics Library. *PLOS Comput. Biol.* **2019**, *15* (2), e1006791. https://doi.org/10.1371/journal.pcbi.1006791.
(57) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1177–1185. https://doi.org/10.1021/ci034231b.
(58) Rivera-Borroto, O. M.; García-de la Vega, J. M.; Marrero-Ponce, Y.; Grau, R. Relational Agreement Measures for Similarity Searching of Cheminformatic Data Sets. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2016**, *13* (1), 158–167. https://doi.org/10.1109/TCBB.2015.2424435.
(59) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1177–1185. https://doi.org/10.1021/ci034231b.
(60) Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45* (4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
(61) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinforma. Oxf. Engl.* **2000**, *16* (5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412.
(62) Iman, R. L.; Davenport, J. M. *Approximations of the Critical Region of the Friedman Statistic*; SAND-79-0883C; CONF-790825-1; Sandia Labs., Albuquerque, NM (USA); Texas Tech Univ., Lubbock (USA), 1979.
(63) García, S.; Fernández, A.; Luengo, J.; Herrera, F. A Study of Statistical Techniques and Performance Measures for Genetics-Based Machine Learning: Accuracy and Interpretability. *Soft Comput. - Fusion Found. Methodol. Appl.* **2009**, *13* (10), 959–977. https://doi.org/10.1007/s00500-008-0392-y.
(64) Demsˇar, J.; Demsar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. 30.

(65) Gupta, S.; Kapoor, P.; Chaudhary, K.; Gautam, A.; Kumar, R.; Consortium, O. S. D. D.; Raghava, G. P. S. In Silico Approach for Predicting Toxicity of Peptides and Proteins. *PLOS ONE* **2013**, *8* (9), e73957. https://doi.org/10.1371/journal.pone.0073957.
(66) Chaudhary, K.; Kumar, R.; Singh, S.; Tuknait, A.; Gautam, A.; Mathur, D.; Anand, P.; Varshney, G. C.; Raghava, G. P. S. A Web Server and Mobile App for Computing Hemolytic Potency of Peptides. *Sci. Rep.* **2016**, *6* (1), 22843. https://doi.org/10.1038/srep22843.
(67) Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Res.* **2002**, *30* (14), 3059–3066. https://doi.org/10.1093/nar/gkf436.
(68) Edgar, R. C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.* **2004**, *32* (5), 1792–1797. https://doi.org/10.1093/nar/gkh340.
(69) Notredame, C.; Higgins, D. G.; Heringa, J. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment11Edited by J. Thornton. *J. Mol. Biol.* **2000**, *302* (1), 205–217. https://doi.org/10.1006/jmbi.2000.4042.
(70) Waterhouse, A. M.; Procter, J. B.; Martin, D. M. A.; Clamp, M.; Barton, G. J. Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* **2009**, *25* (9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.
(71) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16* (6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.
(72) Bailey, T. L. STREME: Accurate and Versatile Sequence Motif Discovery. *Bioinformatics* **2021**, *37* (18), 2834–2840. https://doi.org/10.1093/bioinformatics/btab203.
(73) Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S. MEME Suite: Tools for Motif Discovery and Searching. *Nucleic Acids Res.* **2009**, *37* (suppl_2), W202–W208. https://doi.org/10.1093/nar/gkp335.
(74) Kanehisa, M. Linking Databases and Organisms: GenomeNet Resources in Japan. *Trends Biochem. Sci.* **1997**, *22* (11), 442–444. https://doi.org/10.1016/S0968-0004(97)01130-4.
(75) Sigrist, C. J. A.; Cerutti, L.; Hulo, N.; Gattiker, A.; Falquet, L.; Pagni, M.; Bairoch, A.; Bucher, P. PROSITE: A Documented Database Using Patterns and Profiles as Motif Descriptors. *Brief. Bioinform.* **2002**, *3* (3), 265–274. https://doi.org/10.1093/bib/3.3.265.
(76) Stumpf, M. P. H.; Porter, M. A. Critical Truths About Power Laws. *Science* **2012**, *335* (6069), 665–666. https://doi.org/10.1126/science.1216142.
(77) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M. DBAASP v3: Database of Antimicrobial/Cytotoxic Activity and Structure of Peptides as a Resource for Development of New Therapeutics. *Nucleic Acids Res.* **2021**, *49* (D1), D288–D297. https://doi.org/10.1093/nar/gkaa991.
(78) Singh, S.; Chaudhary, K.; Dhanda, S. K.; Bhalla, S.; Usmani, S. S.; Gautam, A.; Tuknait, A.; Agrawal, P.; Mathur, D.; Raghava, G. P. S. SATPdb: A Database of Structurally Annotated Therapeutic Peptides. *Nucleic Acids Res.* **2016**, *44* (D1), D1119–D1126. https://doi.org/10.1093/nar/gkv1114.
(79) Mehta, D.; Anand, P.; Kumar, V.; Joshi, A.; Mathur, D.; Singh, S.; Tuknait, A.; Chaudhary, K.; Gautam, S. K.; Gautam, A.; Varshney, G. C.; Raghava, G. P. S. ParaPep: A Web

Resource for Experimentally Validated Antiparasitic Peptide Sequences and Their Structures. *Database* **2014**, *2014* (bau051). https://doi.org/10.1093/database/bau051.
(80) Wang, G.; Li, X.; Wang, Z. APD3: The Antimicrobial Peptide Database as a Tool for Research and Education. *Nucleic Acids Res.* **2016**, *44* (D1), D1087–D1093. https://doi.org/10.1093/nar/gkv1278.
(81) Mukaka, M. M. A Guide to Appropriate Use of Correlation Coefficient in Medical Research. *Malawi Med. J.* **2012**, *24* (3), 69–71. https://doi.org/10.4314/mmj.v24i3.
(82) Schober, P.; Boer, C.; Schwarte, L. A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126* (5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864.
(83) Zucchini, W. An Introduction to Model Selection. *J. Math. Psychol.* **2000**, *44* (1), 41–61. https://doi.org/10.1006/jmps.1999.1276.
(84) Shi, G.; Kang, X.; Dong, F.; Liu, Y.; Zhu, N.; Hu, Y.; Xu, H.; Lao, X.; Zheng, H. DRAMP 3.0: An Enhanced Comprehensive Data Repository of Antimicrobial Peptides. *Nucleic Acids Res.* **2021**, No. gkab651. https://doi.org/10.1093/nar/gkab651.
(85) Maclean, M. J.; Brinkworth, C. S.; Bilusich, D.; Bowie, J. H.; Doyle, J. R.; Llewellyn, L. E.; Tyler, M. J. New Caerin Antibiotic Peptides from the Skin Secretion of the Dainty Green Tree Frog Litoria Gracilenta. Identification Using Positive and Negative Ion Electrospray Mass Spectrometry. *Toxicon* **2006**, *47* (6), 664–675. https://doi.org/10.1016/j.toxicon.2006.01.019.
(86) VanCompernolle, S.; Smith, P. B.; Bowie, J. H.; Tyler, M. J.; Unutmaz, D.; Rollins-Smith, L. A. Inhibition of HIV Infection by Caerin 1 Antimicrobial Peptides. *Peptides* **2015**, *71*, 296–303. https://doi.org/10.1016/j.peptides.2015.05.004.
(87) Park, I. Y.; Cho, J. H.; Kim, K. S.; Kim, Y.-B.; Kim, M. S.; Kim, S. C. Helix Stability Confers Salt Resistance upon Helical Antimicrobial Peptides *. *J. Biol. Chem.* **2004**, *279* (14), 13896–13901. https://doi.org/10.1074/jbc.M311418200.
(88) Haney, E. F.; Nazmi, K.; Lau, F.; Bolscher, J. G. M.; Vogel, H. J. Novel Lactoferrampin Antimicrobial Peptides Derived from Human Lactoferrin. *Biochimie* **2009**, *91* (1), 141–154. https://doi.org/10.1016/j.biochi.2008.04.013.
(89) Mechkarska, M.; Coquet, L.; Leprince, J.; Auguste, R. J.; Jouenne, T.; Mangoni, M. L.; Conlon, J. M. Peptidomic Analysis of the Host-Defense Peptides in Skin Secretions of the Trinidadian Leaf Frog Phyllomedusa Trinitatis (Phyllomedusidae). *Comp. Biochem. Physiol. Part D Genomics Proteomics* **2018**, *28*, 72–79. https://doi.org/10.1016/j.cbd.2018.06.006.
(90) Rost, B. Twilight Zone of Protein Sequence Alignments. *Protein Eng. Des. Sel.* **1999**, *12* (2), 85–94. https://doi.org/10.1093/protein/12.2.85.
(91) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res.* **1994**, *22* (22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673.
(92) Thomsen, M. C. F.; Nielsen, M. Seq2Logo: A Method for Construction and Visualization of Amino Acid Binding Motifs and Sequence Profiles Including Sequence Weighting, Pseudo Counts and Two-Sided Representation of Amino Acid Enrichment and Depletion. *Nucleic Acids Res.* **2012**, *40* (W1), W281–W287. https://doi.org/10.1093/nar/gks469.
(93) Bailey, T. L. DREME: Motif Discovery in Transcription Factor ChIP-Seq Data. *Bioinformatics* **2011**, *27* (12), 1653–1659. https://doi.org/10.1093/bioinformatics/btr261.

(94) Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y. C.; Laslo, P.; Cheng, J. X.; Murre, C.; Singh, H.; Glass, C. K. Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **2010**, *38* (4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.