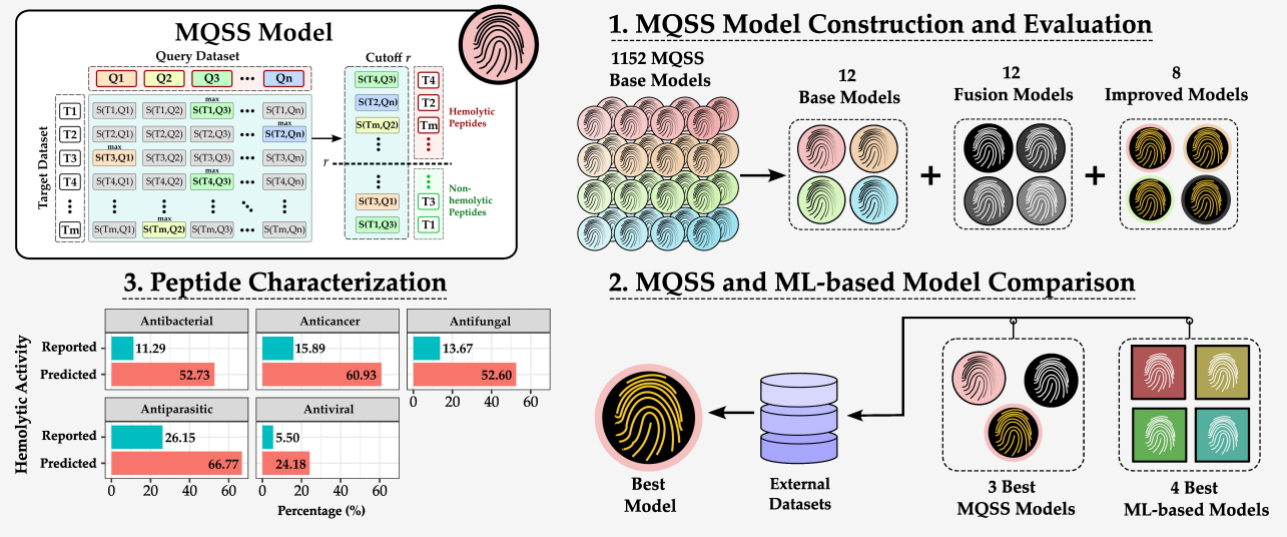
A New Robust Method for Predicting Hemolytic Toxicity from Peptide Sequence

Kevin Castillo-Mendieta[1], Guillermin Agüero-Chapin[2,3,*], Edgar A. Marquez[4], Yunierkis Perez-Castillo[5], Stephen J. Barigye[6], Mariela Pérez-Cárdenas[1], Facundo Peréz-Giménez[7], and Yovani Marrero-Ponce[7,8,*]

[1] School of Biological Sciences and Engineering, Yachay Tech University, Hda. San José s/n y Proyecto Yachay, Urcuquí 100119, Ecuador.
[2] 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; gchapin@ciimar.up.pt (G.A.-C.).
[3] Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal.
[4] Grupo de Investigaciones en Química y Biología, Departamento de Química y Biología, Facultad de Ciencias Básicas, Universidad del Norte, Carrera 51B, Km 5, vía Puerto Colombia, Barranquilla 081007, Colombia.
[5] Bio-Chemoinformatics Research Group and Escuela de Ciencias Físicas y Matemáticas. Universidad de Las Américas, Quito 170504. Ecuador.
[6] Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), 28049, Madrid, Spain.
[7] Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain.
[8] Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Edificio de Especialidades Médicas; and Instituto de Simulación Computacional (ISC-USFQ), Diego de Robles y vía Interoceánica, Quito, 170157, Pichincha, Ecuador (Y.M.-P).

***Correspondence:** gchapin@ciimar.up.pt (G.A.-C) and **Y. Marrero-Ponce** (Y.M.-P)**,** 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 GRAPHICS



ABSTRACT

The desirable pharmacological properties and a broad number of therapeutic activities have placed peptides as promising drugs over small organic molecules and antibody drugs. Nevertheless, toxic effects such as hemolysis have hampered the development of such promising drugs. Hence, a reliable computational tool to predict peptide hemolytic toxicity is enormously useful before synthesis and experimental evaluation. Currently, four web servers that predict hemolytic activity using Machine Learning (ML) algorithms; however, they exhibit some limitations such as the need for a reliable negative set and limited application domain. Hence, we developed a robust model based on a novel theoretical approach that combines network science and a multi-query similarity searching (MQSS) method. A total of 1152 initial models were constructed from 144 scaffolds generated in a previous report. These were evaluated on external datasets, and the best models were fused and improved. Our best MQSS model **I1** outperformed all *state-of-the-art* ML-based models and was used to characterize the prevalence of hemolytic toxicity on therapeutic peptides. Based on our model's estimation, the number of hemolytic peptides might be 3.9-fold higher than the reported.

Keywords: Hemolysis; Antimicrobial peptides; Drug design; Similarity searching; StarPep toolbox.

1. INTRODUCTION

Peptides are becoming as relevant as antibodies and small organic molecules in developing therapeutic drugs [1,2] because they own a combination of desirable pharmacological features. Peptides not only have higher efficacy, target selectivity and good tolerability than small organic molecules, but are at the same time less immunogenic and have higher tissue penetration capacity than proteins and antibodies [3–6]. Moreover, peptides have been shown to display a variety of therapeutic activities such as antimicrobial, antiviral, anticancer, antihypertensive, antioxidant, among others [5,7]. Mainly, antimicrobial peptides (AMPs) are of special interest because they are a promising alternative to overcome antimicrobial resistance caused by a serious misuse of conventional antibiotics [8,9]. Nevertheless, peptide-based drugs do not reach the market because they have some inherent weaknesses, as they are susceptible to peptidase activity and may show toxic effects like hemolysis [9].

Hemolysis involves the premature disruption of Red Blood Cells (RBCs) before the expected 4-month life-span releasing iron, heme and hemoglobin into the vasculature [3,10]. Such products cause NO scavenging, oxidative effects and promote inflammatory responses [11–13]. In consequence, an increased risk of thrombosis [14], atherosclerosis [15], and kidney injury [16] have been associated with hemolysis in many clinical situations [14]. For these reasons, designing a peptide-based therapeutic agent prone to cause hemolysis must be avoided.

Peptide-associated hemolysis can be evaluated experimentally, allowing scientists to screen for low or non-hemolytic peptides that retain efficient therapeutic activity [17–20]. Nevertheless, the lack of standardization among the protocols can lead to inaccurate results. For instance, it has been demonstrated that hemolytic activity assessment strongly depends on the source of RBCs (e.g., human, sheep, rat and rabbit) and the type of buffer or amount of DMSO used [21,22]. In addition, these experimental procedures can become laborious and relatively expensive when evaluating a high

number of peptides [10]. Therefore, a computational approach where peptide hemolytic toxicity can be determined by just providing the amino acid (AA) sequence would be relevant.

The availability of peptide databases containing information on hemolytic activity highly benefits the development of such computational prediction models. Currently, the main databases include: i) Hemolytik [23], with 2651 experimentally validated hemolytic peptides (accessed on March 25, 2023); ii) Database of Antimicrobial Activity and Structure of Peptides (DBAASP v3) [24] (accessed on March 28, 2023), with 11414 entries showing information on hemolytic and cytotoxic activities of AMPs; and iii) StarPep*DB* [25] (accessed on March 25, 2023), a graph-based database containing information of 2004 hemolytic peptides.

Nine machine learning (ML)-based methods for hemolytic toxicity prediction have been described so far [1–5,10,26–28], from which four have implemented a web server system, namely: HemoPI [1], HemoPred [10], HAPPENN [26] and HLPpred-Fuse [3]. Although the models fairly predict peptide hemolytic activity, important limitations are identifiable. Some of them require a specific peptide length range to process the peptides, do neither accept non-conventional natural AAs (e.g., pyrrolysine) nor D-AAs. Moreover, these models highly depend on a reliable negative set selection [29], which is hard to accomplish since there is limited agreement on which is the most appropriate metric to quantify hemolysis and on the definition of the minimum hemolytic concentration (MHC at 5%, 10%, 50% or 100% hemolysis) [26]. In other words, according to the selected criteria, a peptide might be classified as non-hemolytic despite showing moderate or low hemolytic activity. In consequence, a true negative set consisting of actual non-hemolytic peptides is not available.

Here, we propose a new and straightforward computational approach that overcomes such limitations. First, our models can handle peptides of virtually any length, with non-standard AAs and D-AAs. Second, our models do not need to be trained; instead, they rely on the fine-tuning of the sequence alignment type and the similarity cutoff value *r*. Third, since our models are one-class classification models, a negative dataset is not needed for model construction, thus it does not

influence on the learning phase of our algorithm. Finally, the robustness of our best models was deeply assessed, outperforming the *state-of-the-art* ML-based prediction models.

Our method is based on complex network science and multi-query similarity searching (MQSS) models. This procedure has demonstrated not only to accurately predict tumor-homing [30] and antiparasitic peptide activities [31], but also to get a deeper insight into the peptide chemical space [32,33], hence improving peptide-based drug design. Furthermore, a software specifically designed for this purpose, named StarPep toolbox v0.8.5 has been developed to facilitate the workflow [34].

In a previous report [33], we explored the chemical space of the StarPep*DB* hemolytic peptides and generated a variety of subsets (scaffolds) that represent the entire space but retaining a relatively low number of peptides avoiding overrepresentation. Such scaffolds were derived from Half-Space Proximal Networks (HSPNs) constructed by using five different similarity measures. These scaffolds were used in this study as queries for the construction of new and robust MQSS models.

2. MATERIALS AND METHODS

2.1 Databases and Web Server Predictors

2.1.1 Databases and Datasets

StarPepDB. It is a graph-based database containing 45120 peptides with annotated activities retrieved from 40 bioactive databases embedded in the StarPep toolbox software [25]. A subset consisting of 2004 hemolytic peptides was used in ref. [33] for generating scaffolds. Here, we constructed our multi-query similarity searching models based on the 144 scaffolds derived from HSPNs built with 3 similarity metrics: angular separation (AS), Chebyshev (Ch) and Euclidean (Eu) distance. See **SM4.2** in ref. [33] for more details. In addition, this database was used to characterize the prevalence of hemolytic toxicity in antibacterial, anticancer, antifungal, antiparasitic, and antiviral peptides.

***HemoPI-1 Main*.** It consists of 442 experimentally validated highly hemolytic peptides (positive samples) and 442 randomly extracted peptides from Swiss-Prot [1,35] (negative samples). This dataset was used to calibrate the models' parameters by assessing the performance of our 1152 base models.

***HemoPI-1 Validation*.** This dataset contains 110 experimentally validated highly hemolytic peptides (positive samples) and 110 randomly extracted peptides from Swiss-Prot (negative samples) [1]. *HemoPI-1 Validation* was used to confirm the patterns identified using *HemoPI-1 Main* and further select the best 24 and final 12 base (individual) models.

HemoPI-1 NRS1. It encompasses 234 positive and 552 negative samples. This dataset resulted from merging *HemoPI-1 Main* and *Validation* datasets, then peptides redundant with any of the 24 base models were removed. The dataset was used to select the best 12 base models, to validate the robustness and prediction power of these MQSS models, and to carry out performance comparisons with ML-based models.

***HemoPI-1 NRS2*.** This dataset consists of 211 positive and 552 negative samples. It is a subset of *HemoPI-1 NRS1* in which peptides redundant with any of our models (base, fusion, and improved models) were removed. This dataset was involved in validating the best MQSS and ML-based models.

***Big-Hemo*.** This dataset was created to overcome the lack of consensus on adequately quantifying hemolysis. It only contains non-redundant highly hemolytic peptides (positive samples) retrieved from the datasets: i) *HemoPI-2 Main* and *Validation* [1], ii) *HemoPI-3 Main* and *Validation* [1], iii) *HAPPENN* [26], iv) *HLPpred-Fuse Layer 2 Training* and *Independent* [3] and v) *HemoNet* [2]. This dataset is important to assess the ability of the models to correctly identify highly hemolytic peptides, which are more concerning when designing therapeutic drugs. In this dataset, peptides containing 'X' several times in a sequence and Nphe or Nleu in their sequences were discarded. The resulting *Big-Hemo* dataset contains 2196 highly hemolytic peptides [33].

***Big-Hemo NRS1*.** It is a subset of the *Big-Hemo* dataset in which any redundant peptide with all our models (base, fusion, and improved models) has been removed. This dataset contains 1279 peptides and was involved in the final validation of all MQSS and ML-based models.

THPdb. This database contains information about FDA-approved peptides and protein therapeutics [36]. Such peptides should have a low hemolytic toxic profile; hence, this database was used to validate our best model.

The datasets *HemoPI-1 NRS1*, *HemoPI-1 NRS2*, and *Big-Hemo NRS1* were generated to avoid bias from our models when validating their performance and make fair comparisons with ML-based models. Hence, these datasets do not contain peptides already included in any of our MQSS models. It is worth mentioning that ML-based models still have an advantage on these datasets since they were trained with most of these peptides. All datasets are available at **SM1**.

2.1.2 Web Server Predictors

Four ML-based web server predictors have been reported for peptide hemolytic activity: HemoPI [1] (https://webs.iiitd.edu.in/raghava/hemopi/index.php), HemoPred [10] (http://codes.bio/hemopred/), HAPPENN [26] (https://research.timmons.eu/happenn) and HLPpred-Fuse [3] (http://thegleelab.org/HLPpred-Fuse/index.html).

HemoPI (2016) is based on SVM models that use AA and dipeptide composition, binary profiles and motifs as input features [1]. This web server provides five prediction models: "SVM (HemoPI-1) based", "SVM + Motif (HemoPI-1) based", "SVM (HemoPI-2) based", "SVM + Motif (HemoPI-2) based" and "SVM (HemoPI-3) based".

HemoPred (2017) is based on random forest models whose input features are amino acid and dipeptide composition [10]. It only provides a single model in its web server: "HemoPred".

HAPPENN (2020) is based on an artificial neural network (NNs) classifier that has 1024 and 64 nodes in the first and second hidden layers, respectively [26]. Its web server provides three NN models: "HAPPENN-MAIN", "HAPPENN-RR90" and HAPPENN-HARD".

HLPpred-Fuse (2020), unlike the aforementioned prediction models, the final model relies on a fusion of 54 single feature-based models using a meta-predictor approach [3]. It only provides a single model on its web server: "HLPpred-Fuse".

In total, there are ten prediction models reported in four web servers. We compared our models' performance with these *state-of-the-art* prediction models.

2.2 Multi-Query Similarity Searching Method

Our prediction models rely on the premise that similar peptides have similar properties [37]. Therefore, they were constructed using the multi-query similarity searching (MQSS) method, which broadly consists of looking for peptides in a target dataset that are the most similar ones to the query peptides (**Figure 1**). The elements of a MQSS model are a query dataset, a similarity measure, and a similarity cutoff *r*. In this report, 144 scaffolds generated from HSPNs built with 3 similarity metrics, Angular separation (AS), Chebyshev distance (Ch) and Euclidean distance (Eu) were used as query datasets, see **SM4.2 in Ref.** [33]. Regarding the similarity measure, we tested the use of Smith-Waterman local alignment (L) [38] and Needleman-Wunsch global alignment (G) [39] both with BLOSUM-62 substitution matrix [40]. The similarity cutoff values *r* applied in this experiment were 0.4, 0.5, 0.6 and 0.7. The process of constructing MQSS models is described below.

First, using any of the alignment-based algorithms (G or L), a pairwise similarity measure, $S(T_i, Q_j)$, is calculated between each of the peptides in the query (Q) and the target (T) datasets. Then we applied the MAX-SIM rule [41] defined as: $\max\{S(T_i, Q_j)\}$; $\forall\, T_i \in T$. The resulting similarity values were grouped and ranked from the most to the least similar (Group Fusion). Subsequently, a similarity cutoff value *r* was fixed, and target peptides with higher similarity values than *r* are considered as positive samples (hemolytic). MQSS models with the best parameters are selected by evaluating their performance on an external dataset. Model construction and evaluation is easily done using the StarPep toolbox. **Figure 1A** shows a scheme of the MQSS method and **Figure 1B** shows a geometrical interpretation of this method.

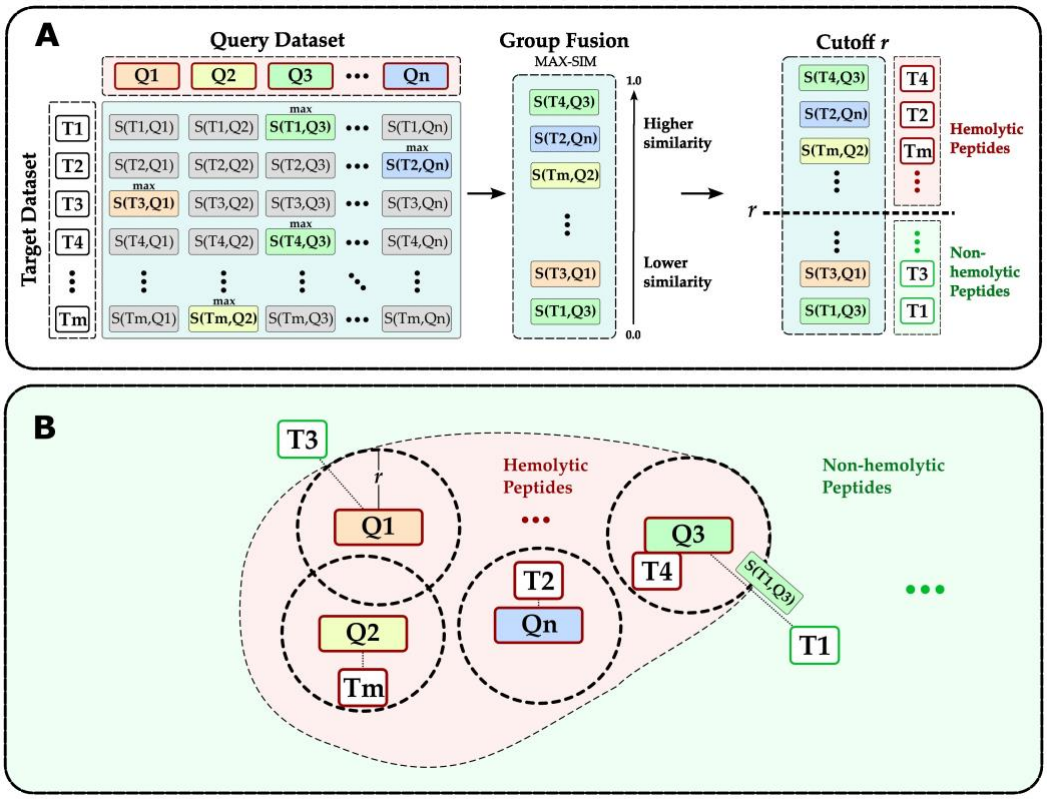


Figure 1. Description of the MQSS method. **A.** A pairwise similarity measure, $S(T_i, Q_j)$, is calculated between each of the peptides in the query dataset (Q) and each of the peptides in the target dataset (T). Local (L) and global (G) alignments can be used to calculate the similarity between peptides. Then for each target peptide, we applied the MAX-SIM rule defined as: $\max\{S(T_i, Q_j)\}$; $\forall\, T_i \in T$. The resulting similarity values were grouped and ranked from the most to the least similar (Group Fusion). Then a similarity cutoff value *r* was fixed, and target peptides with higher similarity values than *r* are considered as positive samples (hemolytic). **B.** Geometrical representation of the MQSS method. This method looks for the target peptides that are the nearest neighbors to our query peptides whose distance is smaller than the radius *r*. i.e., that are inside of the hemolytic space which is defined as the union of all the circles of radius *r* of the query peptides. Figure created with Inkscape [42].

2.3 Construction and Validation of MQQS Models

The workflow consists of four stages: (i) model exploration and selection, (ii) construction of fusion models, (iii) model improvement, and (iv) model validation. These steps were conducted using the StarPep toolbox [34], SeqKit Toolkit [43], and aided with *in-house* Python scripts.

In the first step, we constructed our models from the 144 scaffolds reported in **SM4.2** from Ref. [33]. For each scaffold, we built models with G and L alignments and cutoff values *r* from 0.4 – 0.7 in steps of 0.1. In total, 1152 base MQSS models were generated (**SM3**). Afterwards, we explored the relation between the combination of different parameters in MQSS models on *HemoPI-1 Main* dataset, from which 288 models were filtered for further validation on *HemoPI-1 Validation* dataset. Then, 24 base models were selected based on performance on this dataset (**Figure 2A**). After that,

we selected 12 base models (**B1–B12**) that best performed on the three datasets: *HemoPI-1 Main*, *Validation* and *HemoPI-1 NRS1* (**Figure 2B**).

In the second step, we selected the best three out of the twelve base models and fused their scaffolds, removing redundant peptides (**SM2.2**). It resulted in 4 new scaffolds from which 12 MQSS fusion models (**F1–F12**) were constructed using the similarity measures G or L, and cutoff *r*: 0.4, 0.5 and 0.6 (**Figure 2C** and **SM3**). The robustness of the base and fusion models was assessed on *HemoPI-1 NRS1* and *Big-Hemo* datasets. Furthermore, the performance of ML-based models was assessed using the same datasets (**Figure 2D** and **Figure 3A**).

The third step consists in improving the representativeness of the 8 best MQSS models. In this process, we extracted 413 peptides from the *Big-Hemo* dataset that were incorrectly predicted by more than 4 out of the 8 best models. Using this set of peptides, we constructed HSPNs and scaffolds with the same parameters used for extracting the corresponding scaffolds of the models to be improved (network generation and scaffold extraction processes are detailed in ref. [33]). The improved MQSS models (**I1–I8**) were constructed by fusing their initial scaffolds with the scaffolds obtained from the *Big-Hemo* dataset (**SM2.3**). The similarity measure and cutoff *r* were the same as the precursor model's (**Figure 3B** and **SM3).**

In the fourth step, we validated the robustness of our three best models and the best ML-based model implemented at each web server on the datasets *HemoPI-1 NRS2* and *Big-Hemo NRS1*. Finally, we obtained the best-performing model after ranking the models using the Friedman test calculated using KEEL 3.0 [44]. This test considered the model's performance on the datasets: *HemoPI-1 NRS1*, *HemoPI-1 NRS2*, *Big-Hemo* and *Big-Hemo NRS1*.

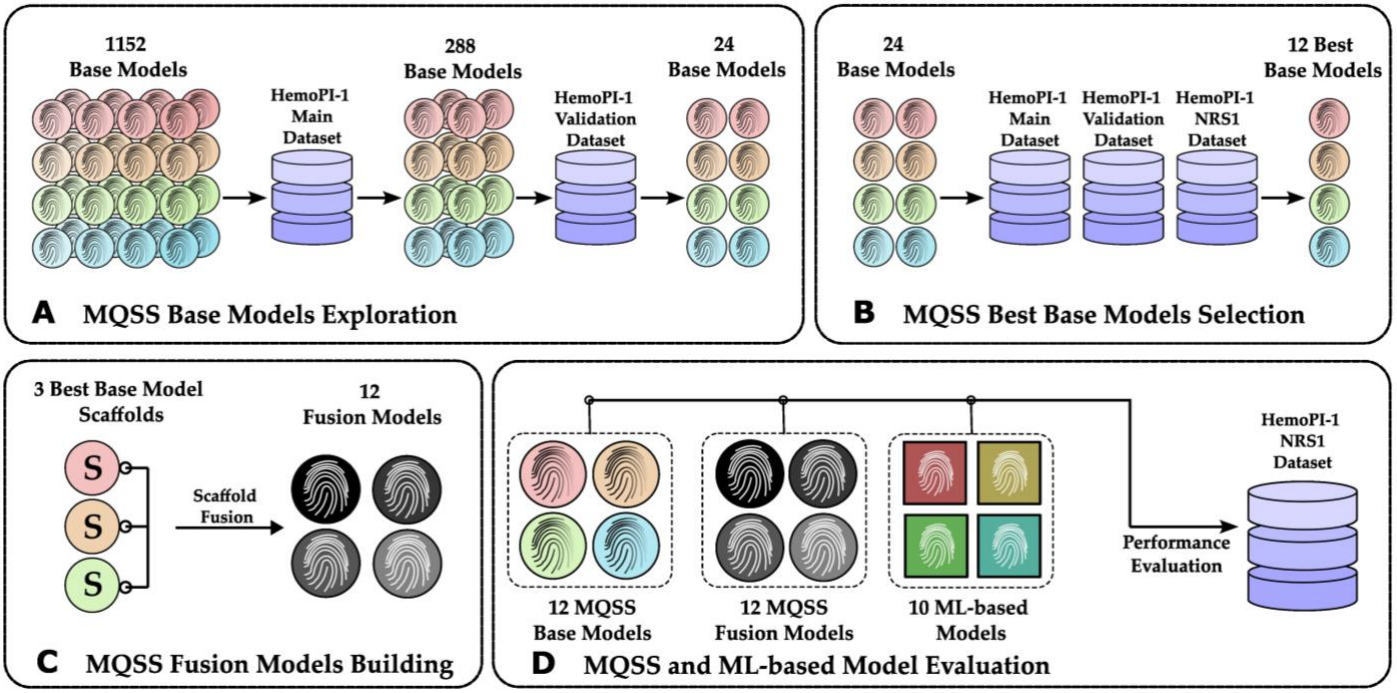


Figure 2. Experimental procedure for exploring and selecting the best MQSS models. **A.** The initial step involves evaluating our models on *HemoPI-1 Main* and *Validation* datasets to find relevant patterns of the combination of parameters used to build the models and parental scaffolds and HSPNs. **B.** The second step consists of selecting the 12 base MQSS models that best perform in the three datasets: *HemoPI-1 Main*, *Validation* and *HemoPI-1 NRS1*. **C.** The third step consists of fusing the scaffolds of the three base MQSS models and subsequently generating new models (fusion models). **D.** In step four, we evaluate our MQSS models and compare them with external models using the dataset *HemoPI-1 NRS1*. Circles and squares with fingerprints represent MQSS models and ML models, respectively. Circles with an "S" inside represent models' scaffolds. Figure created with Inkscape [42].

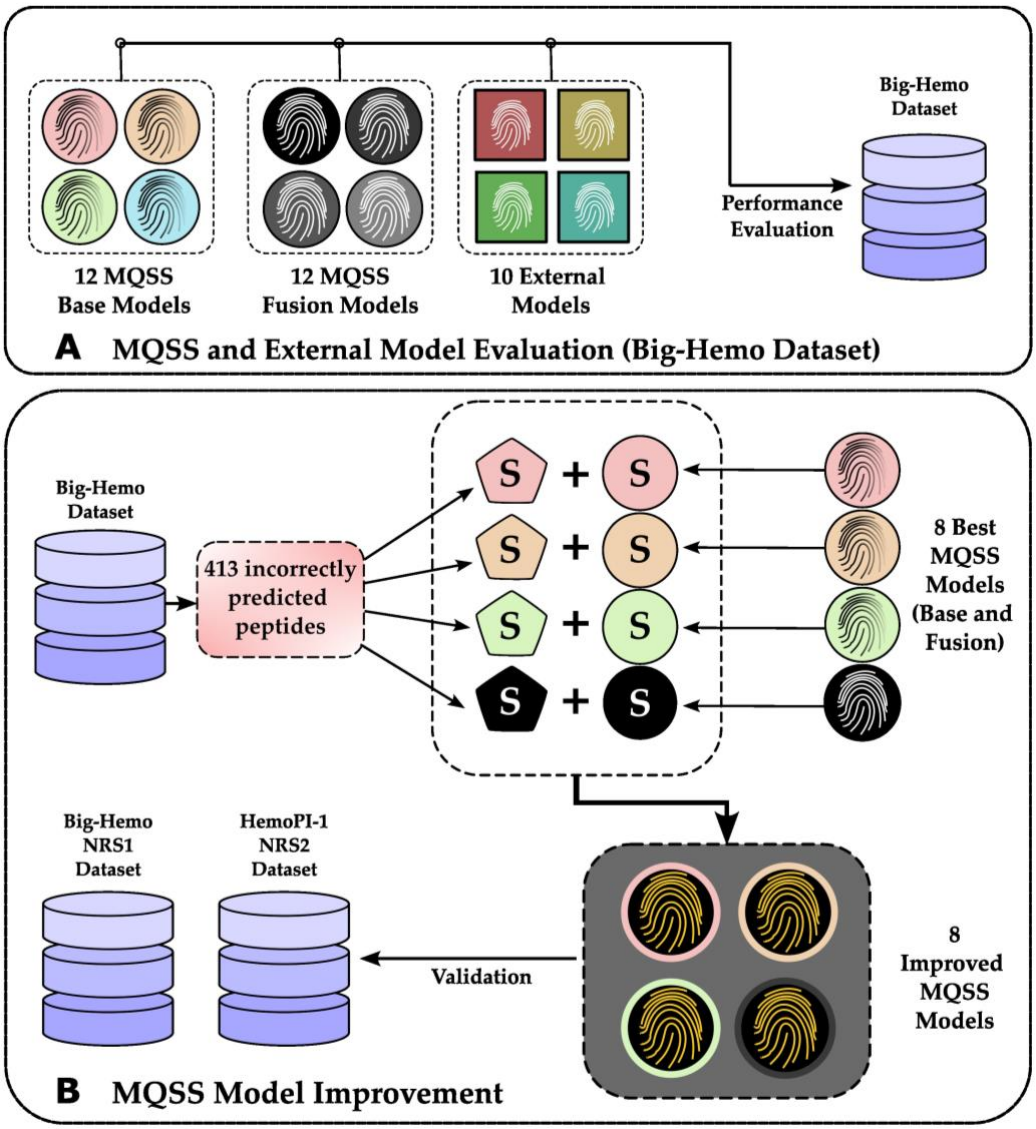


Figure 3. Experimental procedure for improving base and fusion MQSS models. **A.** Initial model validation on *Big-Hemo* dataset to assess the model's ability to recognize highly hemolytic peptides. Circles and squares with fingerprints represent MQSS and ML models, respectively. **B.** Process to improve the representativeness of our best 8 MQSS models by adding scaffolds from the wrongly predicted peptides in the *Big-Hemo* dataset. Circles with an "S" inside represent models' scaffolds, whereas pentagons with an "S" inside represent scaffolds extracted from the set of 413 incorrectly predicted peptides of the *Big-Hemo* dataset. Figure created with Inkscape [42].

2.4 Performance Evaluation

The robustness of MQSS models and the 10 external ML-based prediction models found in four web server predictors was evaluated by calculating their accuracy (Acc), kappa statistics (κ), sensitivity (Sn), specificity (Sp), the precision of positives and negatives (P_{pos} and P_{neg}, respectively), the model's coverage (Cov) and the Matthews correlation coefficient (MCC), being the latter the most important parameter [45,46]. Here, it is worth noting that since *Big-Hemo* and *Big-Hemo NRS1* contain only positive samples, the defining parameter in these datasets was the Sensitivity (Sn).

These parameters are defined as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}, \quad (1)$$

$$\kappa = \frac{Po-Pc}{1-Pc}, \quad (2)$$

$$Sn = \frac{TP}{TP+FN}, \quad (3)$$

$$Sp = \frac{TN}{TN+FP}, \quad (4)$$

$$P_{pos} = \frac{TP}{TP+FP}, \quad (5)$$

$$P_{neg} = \frac{TN}{TN+FN}, \quad (6)$$

$$Cov = \frac{T_M}{T_D}, \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}, \quad (8)$$

where, *TP*, *TN*, *FP*, *FN* are the number of true positives, true negatives, false positives, and false negatives, respectively. *Po* is the relative observed agreement between the observers and *Pc* is the expected change agreement calculated by the formula:

$$Pc = \frac{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}{(TP+TN+FP+FN)^2}, \quad (9)$$

T_M represents the total number of correctly and incorrectly predicted peptides by the model and T_D represents the total number of peptides in the benchmark dataset.

2.5 Characterization of Therapeutic Peptides

Finally, aided by our best model **I1**, we characterized the prevalence of hemolytic toxicity in therapeutic peptides with various endpoints. Peptides from the StarPep*DB* were chosen as this database is one of the most comprehensive reported [25].

Using the StarPep toolbox, we filtered the peptides by function: i) antibacterial, ii) antifungal, iii) antiviral, iv) anticancer and v) antiparasitic. Then we calculated for each endpoint the total number of peptides, the number of peptides reported as hemolytic, and the number of peptides predicted as hemolytic using the **I1** model. Finally, fasta files containing non-hemolytic peptides for each activity were provided.

3. RESULTS AND DISCUSSION

3.1 MQSS Model Exploration and Selection

A total of 1152 initial MQSS models were generated from the combination of 144 scaffolds extracted from Ref. [33], two sequence alignment algorithms (G and L) and four similarity values of *r*: 0.4, 0.5, 0.6 and 0.7. By exploring the behavior of these models on the *HemoPI-1 Main* dataset (**SM6.1**), combinations of parameters that resulted in higher performance were found. The most important factors were the selection of optimal values of cutoffs *s* and *r*. Models based on scaffolds extracted using a similarity cutoff *s* equal to 0.7 or 0.8, showed an average MCC of 0.933 and 0.964, respectively, showing a low variability (**Figure 4A**). Although, models with *s* = 0.9 showed a slightly better performance; this slight improvement does not compensate the higher number of peptides in the scaffolds, about 335 more peptides (**SM6.2**). Regarding the selection of *r*, models with values equal to 0.4, 0.5 and 0.6 tend to have less performance variability and higher robustness than models with *r* = 0.7 (**Figure 4B**).

By filtering models meeting this combination of parameters (*s* = 0.7 or 0.8; *r* = 0.4, 0.5 or 0.6), the number of models was reduced to 288. A further validation on the *HemoPI-1 Validation* dataset allowed to reduce the number of models to 24 (**SM3** and **SM4.2**). An additional dataset, *HemoPI-1 NRS1*, was used to fairly evaluate model performance since it has removed peptides found in any of the scaffolds of our 24 models. Finally, after multiple comparisons of the 24 models on the datasets

HemoPI-1 Main, *Validation* and *HemoPI-1 NRS1*, low performing and redundant models were removed. Only 12 base models (**B1–B12**) were retained for further analysis (**SM4.2**).

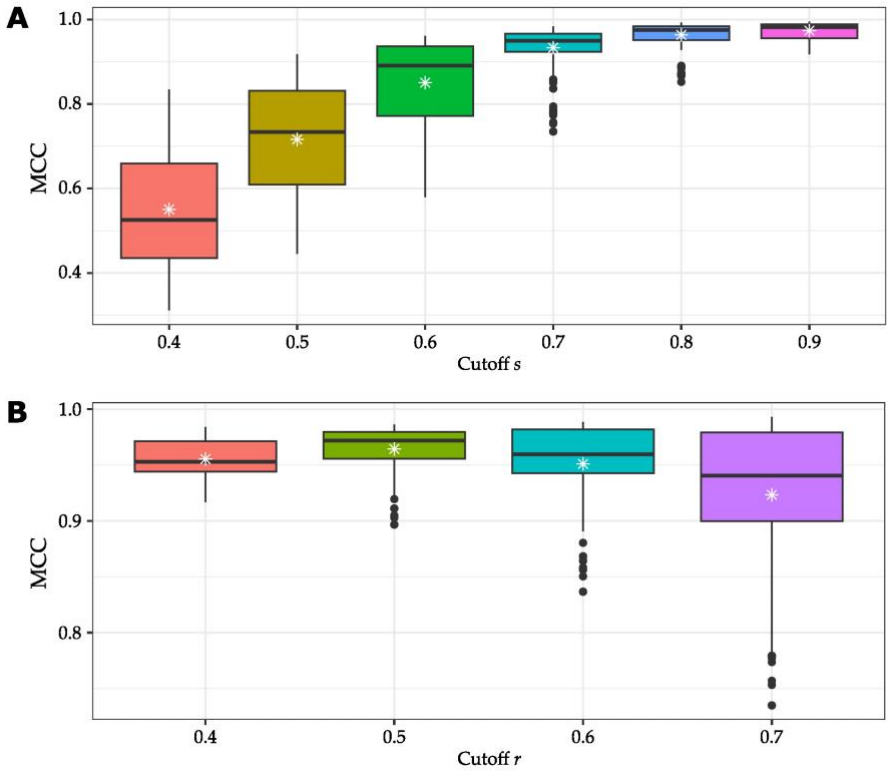


Figure 4. Boxplots showing **(A)** the Matthews correlation coefficient (MCC) of MQSS models built with scaffolds extracted using different values of cutoff *s* and **(B)** from MQSS models using different similarity values of cutoff *r*. White asterisks represent the average MCC. This figure was created with ggplot2 R package [47] and edited with Inkscape [42].

3.2 MQSS Fusion Models

To assess whether the fusion of scaffolds from MQSS base models increases the representativeness and model performance, we selected three base models from the 12 models reported in Section 3.1. Models **B6**, **B8** and **B11**, were chosen since they show high performance when evaluated on *HemoPI-1 Main*, *Validation* and *HemoPI-1 NRS1* datasets. Moreover, they were generated from scaffolds extracted from HSPNs built using different metrics (Ch, Eu, and AS, respectively).

Three new scaffolds were generated by pairwise combination (fusion) of scaffolds from **B6**, **B8** and **B11** and one additional scaffold resulted from merging all three base scaffolds. In total 12 fusion models (**F1–F12**) were created from these scaffolds (**SM3**). Only **F7** and **F9** models showed

performance improvement on *HemoPI-1 NRS1* compared to the 12 base models (**SM4.3**). However, all 12 fusion models were kept for further comparisons.

3.3 Model Improvement.

We compared the performance of base and fusion MQSS models with reported ML-based models on the dataset *HemoPI-1 NRS1* (**SM4.3**). **Table 1** shows the performance of five base and five fusion models. All MQSS models outperformed *state-of-the-art* ML-based models. The best model on this dataset was **F9** (MCC = 0.985), which resulted from the fusion of scaffolds belonging to the models **B8** and **B11**. The optimal parameters for **F9** are local alignment and r = 0.6. On the other hand, the best ML-based model on this dataset was HLPpred-Fuse (MCC = 0.942).

However, when comparing our model's ability to identify highly hemolytic peptides by assessing them on the *Big-Hemo* dataset (**Table 1** and **SM4.3**), models' performance decreased. Our best model on this dataset, **F11,** was ranked fourth (Sn = 0.862). Nevertheless, all our models still outperformed six ML-based models (all HAPPENN models and three of the five HemoPI-1 models).

Table 1. Performance of MQSS and ML-based models on the datasets *HemoPI-1 NRS1* and *Big-Hemo*. It only shows the five base models and five fusion models that performed better on both datasets. For the complete list of the 24 MQSS models and for results of the statistics P_{pos} and P_{neg}, refer to **SM4.3.**

Model	***HemoPI-1 NRS1***						***Big-Hemo***	
	Acc	κ	Sn	Sp	MCC	Cov	Sn	Cov
B1	0.992	0.982	0.983	0.996	0.982	1.00	0.808	1.00
B2	0.991	0.979	0.979	0.996	0.979	1.00	0.812	1.00
B4	0.991	0.979	0.979	0.996	0.979	1.00	0.803	1.00
B6	0.992	0.982	0.996	0.991	0.982	1.00	0.794	1.00
B11	0.992	0.982	0.996	0.991	0.982	1.00	0.854	1.00
F8	0.991	0.979	0.996	0.989	0.979	1.00	0.858	1.00
F9	0.994	0.985	0.996	0.993	0.985	1.00	0.811	1.00
F10	0.992	0.982	0.983	0.996	0.982	1.00	0.818	1.00
F11	0.990	0.976	0.996	0.987	0.976	1.00	0.862	1.00
F12	0.992	0.982	0.996	0.991	0.982	1.00	0.813	1.00
HAPPENN-MAIN	0.861	0.644	0.647	0.952	0.656	0.95	0.766	0.95
HAPPENN-RR90	0.833	0.566	0.576	0.943	0.581	0.95	0.714	0.95
HAPPENN-HARD	0.782	0.445	0.522	0.893	0.452	0.95	0.689	0.95
HemoPI(SVM_HemoPI-1based)	0.968	0.925	0.970	0.967	0.925	1.00	0.897	1.00
HemoPI(SVM_HemoPI-2based)	0.947	0.876	0.962	0.940	0.878	1.00	0.580	1.00
HemoPI(SVM_HemoPI-3based)	0.785	0.542	0.850	0.757	0.562	1.00	0.764	1.00

HemoPI(SVM+Motif_HemoPI-1based)	0.973	0.937	0.974	0.973	0.937	1.00	0.897	1.00
HemoPI(SVM+Motif_HemoPI-2based)	0.925	0.829	0.962	0.909	0.835	1.00	0.614	1.00
HemoPred	0.758	0.493	0.838	0.725	0.518	1.00	0.828	0.91
HLPpred-Fuse	0.975	0.941	1.000	0.964	0.942	1.00	0.955	0.81

Table 2. Sensitivity improvement for high hemolytic peptides evaluated on the *Big-Hemo* dataset. The left side shows the models before improvement, whereas the right side shows the renamed improved models.

Before Improvement		**After Improvement**	
Model	Sn	Model	Sn
B1	0.808	**I1**	0.991
B2	0.812	**I2**	0.992
B6	0.794	**I3**	0.969
B8	0.791	**I4**	0.981
B11	0.854	**I5**	0.994
B12	0.812	**I6**	0.992
F7	0.805	**I7**	0.992
F11	0.862	**I8**	0.997

Since our best models failed to retrieve some peptides from the *Big-Hemo* dataset, we improved their representativeness by adding scaffolds from highly hemolytic peptides. The models **B1**, **B2**, **B6**, **B8**, **B11**, **B12**, **F7**, and **F11** were enhanced by adding a scaffold from the *Big-Hemo* dataset. These Big-Hemo-based scaffolds were extracted using the parameters shown in **SM3**. When comparing the percentage of identical and similar peptides between scaffolds using the Dover Analyzer software [48], Big-Hemo-based scaffolds exhibited a different peptide representativeness compared to the initial scaffolds (**SM2.3.4**). Therefore, MQSS models were positively benefited from this scaffold addition.

Eight improved MQSS models were built, namely **I1–I8**. The alignment and cutoff *r* were the same used for building the precursor model. All improved models showed an enhancement in recognizing highly hemolytic peptides (**Table 2** and **SM4.3**). In fact, the improved version of **F11**, namely **I8**, reached a sensitivity of 0.997 on the *Big-Hemo* dataset, surpassing HLPpred-Fuse's performance (Sn = 0.955).

To ensure that adding new peptides to our models did not affect the overall performance, we assessed them on the *HemoPI-1 NRS1* dataset (**SM4.3**). All improved models were top ranked on this dataset.

3.4 Final Models Evaluation.

We selected one final MQSS model from each group (base, fusion and improved) based on their performance on *HemoPI-1 NRS2* and *Big-Hemo NRS1* datasets, and by considering the most diverse models following the disagreement and double-fault measures [49,50]. The final MQSS models (**B2**, **F11** and **I1**) along with the best ML-based model of each web server were tested on the datasets *HemoPI-1 NRS2* and *Big-Hemo NRS1* (**Table 3**). These datasets allow fair comparisons since they do not contain redundant peptides with any of the MQSS models (improved models included).

Table 3. Performance comparison between our three best MQSS models and the best ML-based model of each web server.

Model	***HemoPI-1 NRS2***						***Big-Hemo NRS1***	
	Acc	κ	Average Recall*	Average Precision*	MCC	Cov	Sn	Cov
B2	0.997	0.993	0.998	0.995	0.993	1.00	0.933	1.00
F11	0.991	0.977	0.994	0.984	0.978	1.00	0.941	1.00
I1	0.996	0.990	0.997	0.993	0.990	1.00	0.999	1.00
HLPpred-Fuse	0.974	0.936	0.982	0.957	0.938	1.00	0.977	0.85
HemoPI(SVM+Motif_HemoPI-1based)	0.975	0.939	0.977	0.963	0.939	1.00	0.892	1.00
HemoPred	0.747	0.463	0.772	0.722	0.492	1.00	0.855	0.85
HAPPENN-MAIN	0.859	0.621	0.786	0.850	0.633	0.96	0.762	0.97

*The average recall is the mean between Sn and Sp, whereas the average precision is the mean between P_{pos} and P_{neg}.

Our three models outperformed ML-based models on the *HemoPI-1 NRS2* dataset, and model **I1** showed the highest sensitivity on *Big-Hemo NRS1* dataset (0.999). HLPpred-Fuse and HemoPI(SVM+Motif_HemoPI-1 based) models also performed well on both datasets. However, HLPpred-Fuse was not able to process peptides with length less than 4 AAs and those containing D-AAs, thus showing one of the lowest coverages on *Big-Hemo* (**Table 1**) and *Big-Hemo NRS1* (**Table 3**) datasets.

On the other hand, HemoPred and HAPPENN-MAIN performed poorly on *HemoPI-1 NRS2* and *Big-Hemo NRS1* datasets; failing also in handling 15% of the peptides from *Big-Hemo NRS1* dataset. Furthermore, HAPPENN models showed several limitations regarding the application domain, which is restricted to peptides of 7–35 AAs in length and do not admit non-standard AAs. In addition, the web server only admits up to 20 sequences per run, which may become tedious when a high number of peptides is intended for prediction.

Finally, we performed the Friedman test to rank the final models based on their performance on the datasets *HemoPI-1 NRS1*, *HemoPI-1 NRS2*, *Big-Hemo* and *Big-Hemo NRS1*. **Figure 5** shows the average ranking scores in the Friedman test. Model **I1** was ranked first, followed by the other two MQSS models, **B2** and **F11**. The post-hoc comparison (**SM7**) revealed no statistical difference among the MQSS models. However, there was a statistical difference between **I1** and any of the ML-based models.

The scaffold of the **I1** model is provided in **SM2.4**. This model uses global alignment (G) and cutoff r = 0.4 as optimal parameters. The model can easily be utilized by importing the scaffold and setting the alignment and cutoff r in the StarPep toolbox.

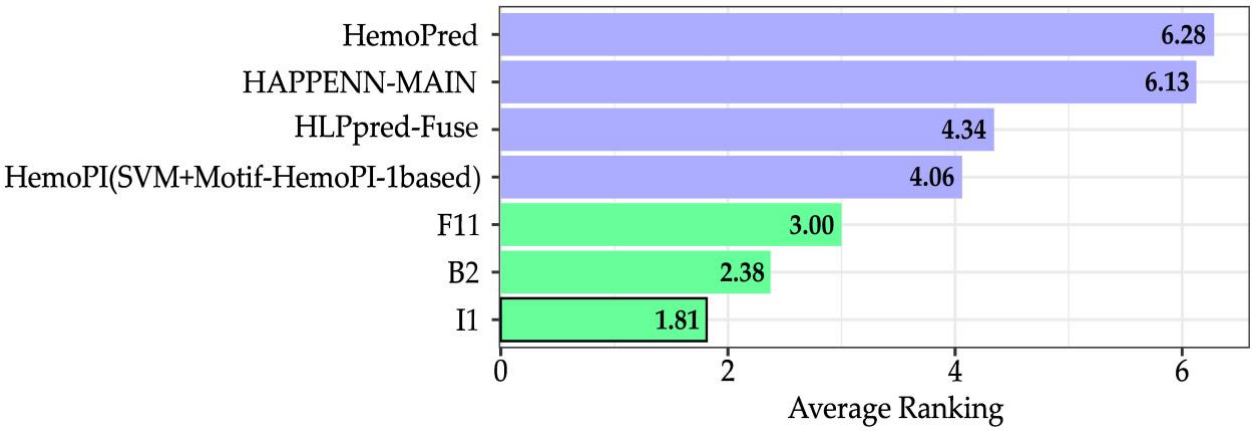


Figure 5. Average ranking scores obtained in the Friedman test. Friedman statistic (distributed according to chi-square with 6 degrees of freedom): 62.631696. P-value computed by Friedman test: 0. This figure was created with ggplot2 R package [47] and edited with Inkscape [42].

Finally, to corroborate our best model's usefulness, we tested it on the curated *THPdb* database (**SM1.7.3**), which includes FDA-approved peptides and protein therapeutics [36]. In principle, our model should not retrieve any of the approved peptides unless a drug reports hemolytic toxicity. Our model **I1** retrieved only three hemolytic peptides from the 183 sequences of the curated *THPdb* dataset, namely Th1024, Th1146 and Th1113. As expected, our model correctly predicted the hemolytic activity of Th1024 (Gramicidin D), an effective antibacterial used for treatment of skin lesions, surface wounds and eye infections. It is reported that due to their highly hemolytic activity of this drug, it cannot be administered internally and hence can only be applied on the skin [51]. The other two FDA-approved drugs, Th1146 (Lucinactant) and Th1113 (Glatiramer acetate) are used as

a pulmonary surfactant and an immunomodulator, respectively [52,53]. Unfortunately, we could not find any information related to the hemolytic toxicity of these peptides.

3.5 Characterization of Therapeutic Peptides

One interesting application for our best MQSS prediction model is to estimate the prevalence of hemolytic toxicity in reported peptides with different endpoints. We analyzed antibacterial, antifungal, antiviral, anticancer and antiparasitic peptides. **Figure 6** shows the number of peptides reported in the StarPep*DB* for each of these specific functions. When estimating the prevalence of hemolytic toxicity in such peptides, we realized that a high proportion of potentially hemolytic peptides still have not been evaluated for this undesired activity. Hence the number of reported hemolytic peptides is underestimated (**Figure 7** and **SM5.6**). For instance, only 11.29 % of the 14376 antibacterial peptides have reported hemolytic toxicity; however, our model predicts there might be about 7580 antibacterial peptides (52.73 %) with hemolytic toxicity.

Interestingly, antiviral peptides seem to have a low prevalence of hemolytic toxicity compared to the other endpoints as only 24.18 % of the peptides have been predicted as hemolytic. In general, our model predicts a 3.9-fold increase in the actual number of hemolytic peptides for these five endpoints. This result shows that caution should be taken when considering peptides whose information about hemolytic toxicity is not provided, as this does not imply that peptides are not hemolytic. In such scenarios, a prediction model such as the presented in this report could be useful.

Finally, to facilitate the reliable exploration of non-hemolytic therapeutic peptides, we provided a list of peptides sorted by function, in which peptides either reported or predicted as hemolytic have been removed (fasta files are available at **SM5**). Additional information about these peptides can be retrieved using the StarPep toolbox.

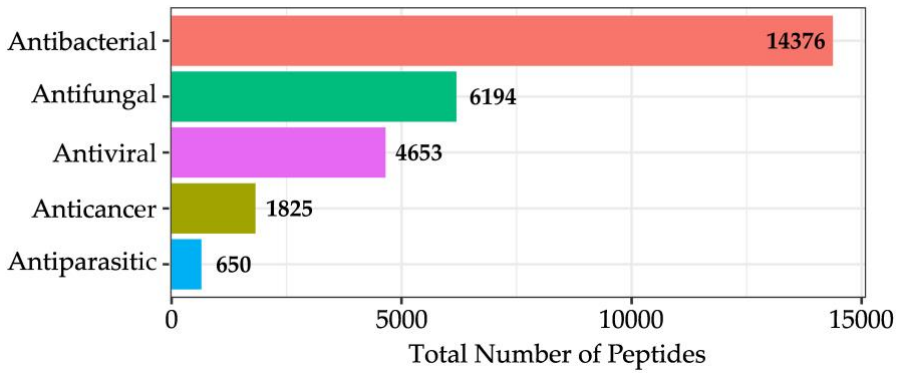


Figure 6. Number of peptides with different functions reported in StarPep*DB*. It is worth noting that these classes are not mutually exclusive as some peptides might have more than one activity. This figure was created with ggplot2 R package [47] and edited with Inkscape [42].

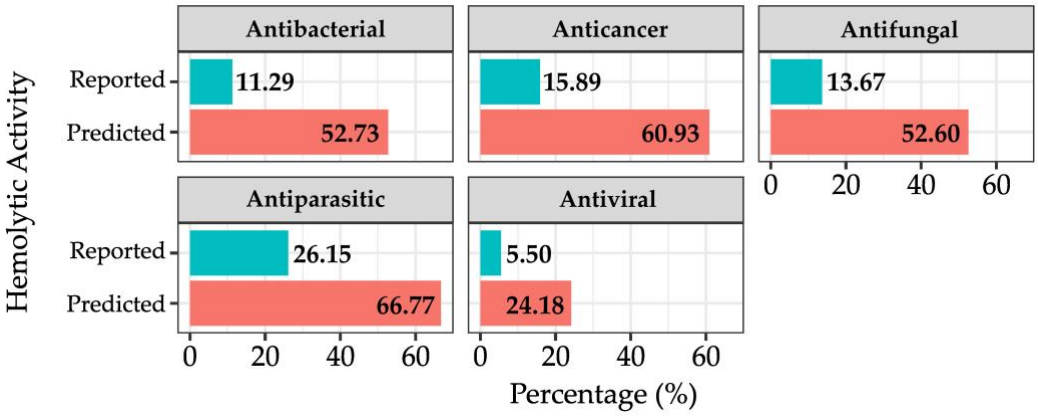


Figure 7. Percentage of peptides presenting reported and predicted hemolytic activity in antimicrobial, anticancer, antifungal, antiparasitic and antiviral peptides. This figure was created with ggplot2 R package [47] and edited with Inkscape [42].

4. CONCLUSIONS

An *in silico* model able to reliably predict hemolytic toxicity from peptide sequences is a highly useful tool that can help accelerate the development and approval of new peptide drugs cost-efficiently. Currently, nine prediction models are available for this task, all based on ML. In this report, we presented a more robust model based on a novel approach that uses network science and the MQSS method. Our best model not only outperformed *state-of-the-art* prediction models implemented in web servers but also overcomes some of their pitfalls, such as a limited generalization conditioned by an application domain. Furthermore, MQSS models can be easily implemented by uploading the model's scaffold and setting the appropriate alignment algorithm and cutoff *r* at the StarPep toolbox. As a valid application of our model, the prevalence of hemolytic toxicity on

therapeutic peptides was estimated, finding that the actual number of hemolytic peptides is underrepresented and there might be 3.9-fold more hemolytic peptides than the reported.

ASSOCIATED CONTENT

Supporting Information: The Supporting Information is available free of charge at https://pubs.acs.org/...

SM1 – Datasets used in this study for validating prediction models. **SM2** – Scaffolds used to build MQSS models. **SM3** – Information about the parameters of each MQSS model (scaffold, alignment type and cutoff *r*). **SM4** – Performance evaluation of the models on all benchmark datasets. **SM5** – Datasets only containing non-hemolytic peptides, quite useful for exploring/selecting non-toxic therapeutic peptides. **SM6** –Additional information of the parameter exploration in the initial models. **SM7** – Friedman test results.

Notes

The authors declare no competing financial interest. The authors declare no conflicts of interest. A noncommercial and fully cross-platform StarPep toolbox software and the respective user manual are freely available online at http://mobiosd-hub.com/starpep.

ACKNOWLEDGMENTS

Y.M.-P. thanks to the USFQ Collaboration Grant (Project **ID**16897) and Med-Grant (Project **ID**17601). G.A.-C. was supported by national funds through FCT - Foundation for Science and Technology within the scope of UIDB/04423/2020 and UIDP/04423/2020.

5. REFERENCES

(1) 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.

(2) Yaseen, A.; Gull, S.; Akhtar, N.; Amin, I.; Minhas, F. HemoNet: Predicting Hemolytic Activity of Peptides with Integrated Feature Learning. *J Bioinform Comput Biol* **2021**, *19* (5), 2150021. https://doi.org/10.1142/S0219720021500219.

(3) Hasan, M. M.; Schaduangrat, N.; Basith, S.; Lee, G.; Shoombuatong, W.; Manavalan, B. HLPpred-Fuse: Improved and Robust Prediction of Hemolytic Peptide and Its Activity by Fusing Multiple Feature Representation. *Bioinformatics* **2020**, *36* (11), 3350–3356. https://doi.org/10.1093/bioinformatics/btaa160.

(4) Kumar, V.; Kumar, R.; Agrawal, P.; Patiyal, S.; Raghava, G. P. S. A Method for Predicting Hemolytic Potency of Chemically Modified Peptides From Its Structure. *Front Pharmacol* **2020**, *11*, 54. https://doi.org/10.3389/fphar.2020.00054.

(5) Plisson, F.; Ramírez-Sánchez, O.; Martínez-Hernández, C. Machine Learning-Guided Discovery and Design of Non-Hemolytic Peptides. *Sci Rep* **2020**, *10* (1), 16581. https://doi.org/10.1038/s41598-020-73644-6.

(6) Vlieghe, P.; Lisowski, V.; Martinez, J.; Khrestchatisky, M. Synthetic Therapeutic Peptides: Science and Market. *Drug Discovery Today* **2010**, *15* (1), 40–56. https://doi.org/10.1016/j.drudis.2009.10.009.

(7) Wang, L.; Wang, N.; Zhang, W.; Cheng, X.; Yan, Z.; Shao, G.; Wang, X.; Wang, R.; Fu, C. Therapeutic Peptides: Current Applications and Future Directions. *Sig Transduct Target Ther* **2022**, *7* (1), 1–27. https://doi.org/10.1038/s41392-022-00904-4.

(8) Frieri, M.; Kumar, K.; Boutin, A. Antibiotic Resistance. *Journal of Infection and Public Health* **2017**, *10* (4), 369–378. https://doi.org/10.1016/j.jiph.2016.08.007.

(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. *International Journal of Molecular Sciences* **2020**, *21* (19), 7047. https://doi.org/10.3390/ijms21197047.

(10) Win, T. S.; Malik, A. A.; Prachayasittikul, V.; S Wikberg, J. E.; Nantasenamat, C.; Shoombuatong, W. HemoPred: A Web Server for Predicting the Hemolytic Activity of Peptides. *Future Medicinal Chemistry* **2017**, *9* (3), 275–291. https://doi.org/10.4155/fmc-2016-0188.

(11) Vinchi, F.; Sparla, R.; Passos, S. T.; Sharma, R.; Vance, S. Z.; Zreid, H. S.; Juaidi, H.; Manwani, D.; Yazdanbakhsh, K.; Nandi, V.; Silva, A. M. N.; Agarvas, A. R.; Fibach, E.; Belcher, J. D.; Vercellotti, G. M.; Ghoti, H.; Muckenthaler, M. U. Vasculo-Toxic and pro-Inflammatory Action of Unbound Haemoglobin, Haem and Iron in Transfusion-Dependent Patients with Haemolytic Anaemias. *British Journal of Haematology* **2021**, *193* (3), 637–658. https://doi.org/10.1111/bjh.17361.

(12) Rapido, F. The Potential Adverse Effects of Haemolysis. *Blood Transfus* **2017**, *15* (3), 218–221. https://doi.org/10.2450/2017.0311-16.

(13) Immenschuh, S.; Vijayan, V.; Janciauskiene, S.; Gueler, F. Heme as a Target for Therapeutic Interventions. *Frontiers in Pharmacology* **2017**, *8*.

(14) L'Acqua, C.; Hod, E. New Perspectives on the Thrombotic Complications of Haemolysis. *British Journal of Haematology* **2015**, *168* (2), 175–185. https://doi.org/10.1111/bjh.13183.

(15) Smith, A.; McCulloh, R. J. Mechanisms of Haem Toxicity in Haemolysis and Protection by the Haem-Binding Protein, Haemopexin. *ISBT Science Series* **2017**, *12* (1), 119–133. https://doi.org/10.1111/voxs.12340.

(16) Van Avondt, K.; Nur, E.; Zeerleder, S. Mechanisms of Haemolysis-Induced Kidney Injury. *Nat Rev Nephrol* **2019**, *15* (11), 671–692. https://doi.org/10.1038/s41581-019-0181-0.

(17) DeGrado, W. F.; Musso, G. F.; Lieber, M.; Kaiser, E. T.; Kézdy, F. J. Kinetics and Mechanism of Hemolysis Induced by Melittin and by a Synthetic Melittin Analogue. *Biophysical Journal* **1982**, *37* (1), 329–338. https://doi.org/10.1016/S0006-3495(82)84681-X.

(18) Li, Q.; Dong, C.; Deng, A.; Katsumata, M.; Nakadai, A.; Kawada, T.; Okada, S.; Clayberger, C.; Krensky, A. M. Hemolysis of Erythrocytes by Granulysin-Derived Peptides but Not by Granulysin. *Antimicrobial Agents and Chemotherapy* **2005**, *49* (1), 388–397. https://doi.org/10.1128/AAC.49.1.388-397.2005.

(19) Belokoneva, O. S.; Satake, H.; Mal'tseva, E. L.; Pal'mina, N. P.; Villegas, E.; Nakajima, T.; Corzo, G. Pore Formation of Phospholipid Membranes by the Action of Two Hemolytic Arachnid Peptides of Different Size. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2004**, *1664* (2), 182–188. https://doi.org/10.1016/j.bbamem.2004.05.007.

(20) Feder, R.; Dagan, A.; Mor, A. Structure-Activity Relationship Study of Antimicrobial Dermaseptin S4 Showing the Consequences of Peptide Oligomerization on Selective Cytotoxicity. *J Biol Chem* **2000**, *275* (6), 4230–4238. https://doi.org/10.1074/jbc.275.6.4230.

(21) Oddo, A.; Hansen, P. R. Hemolytic Activity of Antimicrobial Peptides. In *Antimicrobial Peptides: Methods and Protocols*; Hansen, P. R., Ed.; Methods in Molecular Biology; Springer: New York, NY, 2017; pp 427–435. https://doi.org/10.1007/978-1-4939-6737-7_31.

(22) Greco, I.; Molchanova, N.; Holmedal, E.; Jenssen, H.; Hummel, B. D.; Watts, J. L.; Håkansson, J.; Hansen, P. R.; Svenson, J. Correlation between Hemolytic Activity, Cytotoxicity and Systemic in Vivo Toxicity of Synthetic Antimicrobial Peptides. *Sci Rep* **2020**, *10* (1), 13206. https://doi.org/10.1038/s41598-020-69995-9.

(23) Gautam, A.; Chaudhary, K.; Singh, S.; Joshi, A.; Anand, P.; Tuknait, A.; Mathur, D.; Varshney, G. C.; Raghava, G. P. S. Hemolytik: A Database of Experimentally Determined Hemolytic and Non-Hemolytic Peptides. *Nucleic Acids Research* **2014**, *42* (D1), D444–D449. https://doi.org/10.1093/nar/gkt1008.

(24) 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 Research* **2021**, *49* (D1), D288–D297. https://doi.org/10.1093/nar/gkaa991.

(25) 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.

(26) Timmons, P. B.; Hewage, C. M. HAPPENN Is a Novel Tool for Hemolytic Activity Prediction for Therapeutic Peptides Which Employs Neural Networks. *Sci Rep* **2020**, *10* (1), 10869. https://doi.org/10.1038/s41598-020-67701-3.

(27) Capecchi, A.; Cai, X.; Personne, H.; Köhler, T.; Delden, C. van; Reymond, J.-L. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. *Chem. Sci.* **2021**, *12* (26), 9221–9232. https://doi.org/10.1039/D1SC01713F.

(28) Salem, M.; Keshavarzi Arshadi, A.; Yuan, J. S. AMPDeep: Hemolytic Activity Prediction of Antimicrobial Peptides Using Transfer Learning. *BMC Bioinformatics* **2022**, *23* (1), 389. https://doi.org/10.1186/s12859-022-04952-z.

(29) Sidorczuk, K.; Gagat, P.; Pietluch, F.; Kała, J.; Rafacz, D.; Bąkała, L.; Słowik, J.; Kolenda, R.; Rödiger, S.; Fingerhut, L. C. H. W.; Cooke, I. R.; Mackiewicz, P.; Burdukiewicz, M. Benchmarks in Antimicrobial Peptide Prediction Are Biased Due to the Selection of Negative Data. *Brief Bioinform* **2022**, *23* (5), bbac343. https://doi.org/10.1093/bib/bbac343.

(30) Romero, M.; Marrero-Ponce, Y.; Rodríguez, H.; Agüero-Chapin, G.; Antunes, A.; Aguilera-Mendoza, L.; Martinez-Rios, F. A Novel Network Science and Similarity-Searching-Based Approach for Discovering Potential Tumor-Homing Peptides from Antimicrobials. *Antibiotics* **2022**, *11* (3), 401. https://doi.org/10.3390/antibiotics11030401.

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

(32) Agüero-Chapin, G.; Antunes, A.; Mora, J. R.; Pérez, N.; Contreras-Torres, E.; Valdes-Martini, J. R.; Martinez-Rios, F.; Zambrano, C. H.; Marrero-Ponce, Y. Complex Networks Analyses of Antibiofilm Peptides: An Emerging Tool for Next Generation Antimicrobials Discovery. Preprints March 10, 2023. https://doi.org/10.20944/preprints202303.0193.v1.

(33) Castillo-Mendieta, K.; Agüero-Chapin, G.; Vispo, N. S.; Márquez, E. A.; Perez-Castillo, Y.; Barigye, S. J.; Marrero-Ponce, Y. Peptide Hemolytic Activity Analysis Using Visual Data Mining of Similarity-Based Complex Networks. Preprints March 17, 2023. https://doi.org/10.20944/preprints202303.0322.v1.

(34) 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.

(35) The UniProt Consortium. Update on Activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research* **2013**, *41* (D1), D43–D47. https://doi.org/10.1093/nar/gks1068.

(36) Usmani, S. S.; Bedi, G.; Samuel, J. S.; Singh, S.; Kalra, S.; Kumar, P.; Ahuja, A. A.; Sharma, M.; Gautam, A.; Raghava, G. P. S. THPdb: Database of FDA-Approved Peptide and Protein Therapeutics. *PLoS One* **2017**, *12* (7), e0181748. https://doi.org/10.1371/journal.pone.0181748.

(37) 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.

(38) Smith, T. F.; Waterman, M. S. Identification of Common Molecular Subsequences. *Journal of Molecular Biology* **1981**, *147* (1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.

(39) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J Mol Biol* **1970**, *48* (3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4.

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

(41) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11* (23), 1046–1053. https://doi.org/10.1016/j.drudis.2006.10.005.

(42) Inkscape. Inkscape Project, 2023. https://github.com/inkscape/inkscape (accessed 2023-03-01).

(43) Shen, W.; Le, S.; Li, Y.; Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* **2016**, *11* (10), e0163962. https://doi.org/10.1371/journal.pone.0163962.

(44) Triguero, I.; González, S.; Moyano, J. M.; García López, S.; Alcalá Fernández, J.; Luengo Martín, J.; Fernández Hilario, A. L.; Jesús Díaz, M. J. del; Sánchez, L.; Herrera Triguero, F. KEEL 3.0: An Open Source Software for Multi-Stage Analysis in Data Mining. *International Journal of Computational Intelligence Systems* **2017**, *10*, 1238–1249. https://doi.org/10.2991/ijcis.10.1.82.

(45) Chicco, D.; Tötsch, N.; Jurman, G. The Matthews Correlation Coefficient (MCC) Is More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation. *BioData Mining* **2021**, *14* (1), 13. https://doi.org/10.1186/s13040-021-00244-z.

(46) Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics* **2020**, *21* (1), 6. https://doi.org/10.1186/s12864-019-6413-7.

(47) Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*, 1st ed.; Use R!; Springer New York, NY, 2009.

(48) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Tellez-Ibarra, R.; Llorente-Quesada, M. T.; Salgado, J.; Barigye, S. J.; Liu, J. Overlap and Diversity in Antimicrobial Peptide Databases: Compiling a Non-Redundant Set of Sequences. *Bioinformatics* **2015**, *31* (15), 2553–2559. https://doi.org/10.1093/bioinformatics/btv180.

(49) Kuncheva, L. I.; Whitaker, C. J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning* **2003**, *51* (2), 181–207. https://doi.org/10.1023/A:1022859003006.

(50) Tang, E. K.; Suganthan, P. N.; Yao, X. An Analysis of Diversity Measures. *Mach Learn* **2006**, *65* (1), 247–271. https://doi.org/10.1007/s10994-006-9449-2.

(51) *Gramicidin D*. https://go.drugbank.com/drugs/DB00027 (accessed 2023-04-05).

(52) *Lucinactant*. https://go.drugbank.com/drugs/DB04897 (accessed 2023-04-05).

(53) *Copaxone (Glatiramer Acetate): Uses, Dosage, Side Effects, Interactions, Warning*. RxList. https://www.rxlist.com/copaxone-drug.htm (accessed 2023-04-06).