Article

Not peer-reviewed version

Innovative Alignment-Based Method for Antiviral Peptide Prediction

Daniela De Llano García , Yovani Marrero-Ponce * , Guillermin Agüero-Chapin * , Francesc J. Ferri , Agostinho Antunes , Felix Martinez-Rios , Hortensia María Rodríguez-Cabrera

Posted Date: 18 July 2024

doi: 10.20944/preprints202407.1476.v1

Keywords: antiviral peptide; Multi Query Similarity Search; machine learning; StarPep toolbox; antiviral peptide dataset

Article

Innovative Alignment-Based Method for Antiviral Peptide Prediction

Daniela de Llano García [1], Yovani Marrero-Ponce [2,3,4,*], Guillermin Agüero-Chapin [5,6,*], Francesc J. Ferri [4], Agostinho Antunes [5,6], Felix Martinez-Rios [3] and Hortensia Rodriguez [1]

[1] School of Chemical Sciences and Engineering, Yachay Tech University, Hda. San José s/n y Proyecto Yachay, Urcuquí 100119, Ecuador
[2] 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
[3] Universidad Panamericana, Facultad de Ingeniería, Benito Juárez 03920, Ciudad de México, México
[4] Computer Science Department, Universitat de València, Burjassot 46100, Spain
[5] CIIMAR – Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450-208, Portugal
[6] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal
* Correspondence: ymarrero77@yahoo.es (Y.M.-P.); gchapin@ciimar.up.pt (G.A.-C.)

Abstract: Antiviral peptides (AVPs) represent a promising strategy for addressing the global challenge of viral infections and their growing resistance to traditional drugs. Lab-based AVP discovery methods are resource-intensive, highlighting the need for efficient computational alternatives. In this study, we developed five non-trained but supervised Multi-Query Similarity Search Models (MQSSMs) integrated into the StarPep toolbox. Rigorous testing and validation across diverse AVP datasets confirmed the models' robustness and reliability. The top-performing model, **M13+**, demonstrated impressive results with an accuracy of 0.969 and a Matthew's correlation coefficient of 0.71. To assess their competitiveness, the top five models were benchmarked against 14 publicly available machine learning and deep learning AVP predictors. The MQSSMs outperformed these predictors, highlighting their efficiency in terms of resource demand and public accessibility. Another significant achievement of this study is the creation of the most comprehensive dataset of antiviral sequences to date.

Keywords: antiviral peptide; Multi Query Similarity Search; machine learning; StarPep toolbox; antiviral peptide dataset

1. Introduction

Antiviral peptides (AVPs) represent a promising, innovative and unconventional approach in the ongoing fight against viral infections, a persistent global health concern [1]. The ever-present threat of viral outbreak underscores the critical need for effective antiviral treatments. AVPs have emerged as a promising alternative due to their unique ability to target various stages of viral infections [2].

Viruses, notorious for their vast genetic diversity and adept replication within host cells, present immense challenges in disease containment [3]. Traditional antiviral drugs, while effective in certain instances, often have limited reach and may become obsolete due to resistance [4]. This treatment void has catalysed research into AVPs as alternative antiviral agents. AVPs have shown efficacy against numerous viruses, including lethal ones like HIV [5], influenza [6], hepatitis [7], and emerging zoonotic viruses such as Ebola [8] and Zika [9].

 doi:10.20944/preprints202407.1476.v1

The therapeutic allure of AVPs is rooted in their distinctive mechanisms of action. They hinder viral attachment, fusion, and replication, offering a comprehensive strategy against viral infections [10]. Derived from synthetic libraries or sections of natural proteins, these peptides boast characteristics vital for their antimicrobial actions. Their low toxicity, high specificity, and efficiency marking them as potent contenders in medical applications [11].

Discovering and validating AVPs through lab experiments is resource-intensive, but computational strategies are gaining traction for their potential in identifying peptides with antimicrobial properties. Crafting robust computational models is imperative for effectively pinpointing potential AVPs. Currently, the strategy for processing expansive peptide data banks relies on machine learning (ML), enabling in-depth multidimensional data analysis. Conventional ML algorithms, such as SVM [12–14], kNN [15], RF [16–18], NN [19], and deep learning (DL) algorithms [20–22], have shown prowess in discerning patterns in peptide sequences and assessing new ones.

However, a point of contention in this domain is the efficacy of DL models in forecasting AVPs. Many DL techniques necessitate vast experimentally validated peptide sequence datasets, which are often lacking. One remedy is "data augmentation," although it is largely untapped [21]. As García-Jacas et al. highlighted [23], DL does not significantly outpace traditional ML, and their algorithmic outputs frequently intersect. Additional challenges include the overrepresentation of specific sequences and imbalanced data, which can distort evaluations based purely on accuracy. Moreover, reproducibility challenges persist, as not all scientists disclose their source code or datasets, obstructing these methods' widespread acceptance. Therefore, while ML-centric methods hold promise in predicting active peptides, their refinement is an ongoing endeavour [24,25].

Recent studies have showcased a non-trained supervised technique, the Multi-Query Similarity Search (MQSS), for predicting peptide bioactivities, including hemolysis [26], tumor-homing [27], and antiparasitic [28] activities, with impressive results. This method trumps conventional ML methods in several ways: it is user-friendly, does not rely on web server availability, consumes fewer computational resources, and processes sequences with non-standard amino acids or varying lengths. Remarkably, MQSS models (MQSSMs) function without extensive training, relying instead on fine-tuning certain parameters like sequence alignment type and the similarity cutoff value. They can be developed without needing a negative dataset, a significant advantage given the scarcity of validated negative sequences, ensuring the learning phase is not skewed by data imbalances.

In prior research, we explored the AVPs Chemical Space using interactive mining and complex networks via the StarPep toolbox ([29], https://github.com/Grupo-Medicina-Molecular-y-Traslacional/StarPep). We generated, disclosed, and made public diverse reducts of the AVPs Chemical Space by implementing scaffold extraction to Half-Space Proximal Networks formed from the predefined Chemical Space. These scaffolds now serve as a foundation for designing reference/query datasets for MQSSM creation [30]. As the MQSS technique has not been applied to AVPs, our primary objective is to create MQSSMs for AVPs that match or surpass existing predictors, while addressing the shortcomings of ML predictors. The MQSSMs were crafted using the available resources in StarPep *toolbox*. Our overarching goal is to enhance the methodologies for identifying and advancing AVPs, potentially revolutionizing antiviral therapies. Additionally, we have assembled the most extensive dataset of antiviral sequences to date, essential for developing effective AVP predictors and ensuring robust model validation.

2. Materials and Methods

2.1. The Multi-Query Similarity Search Model (MQSSM). The Overall Approach

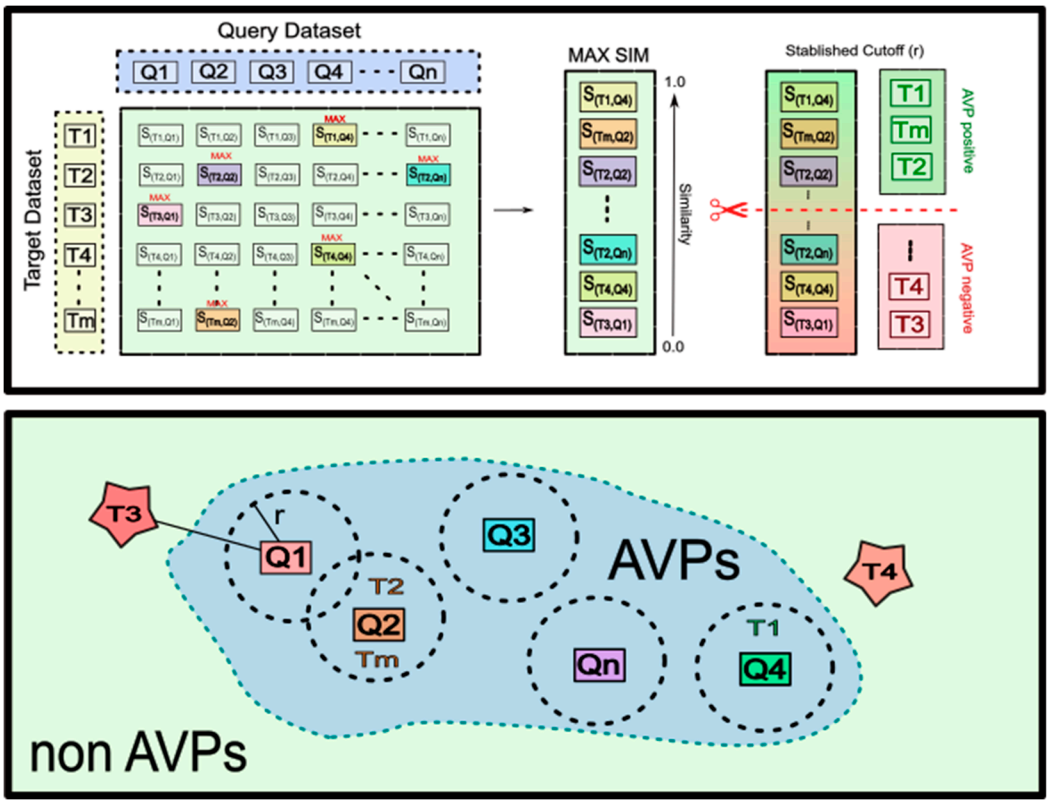
The models proposed here are based on similarity searching. Multiple positive sequences are employed as queries to predict the antiviral activity of a target dataset. This one-class model identifies potential positive sequences based on their similarity to the provided references. The construction of these models involves tuning three different parameters:

Query Dataset: This is defined as the reference dataset that the model uses for predictions.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 July 2024 doi:10.20944/preprints202407.1476.v1

Pairwise Sequence Alignment Algorithm: Smith-Waterman (local alignment) and Needleman-Wunsch (global alignment) algorithms are employed, using the Blossum-62 substitution matrix for calculating pairwise similarity scores.

Similarity Threshold: The MQSS is a fusion model, specifically a group fusion model. In a group fusion, a similarity is computed for each reference sequence *q* and target sequence *t*, denoted as *S* (*q*, *t*). All pairwise similarity scores are combined for each sequence *t*, and the MAX-SIM rule is applied. The similarity scores are then ranked in decreasing order. A Similarity Threshold (*c*) is set to determine which sequences are considered to have positive activity. For a better understanding of the process, a graphical representation is provided in Scheme 1.



Scheme 1. Overview of the proposed Multi-Query Similarity Search Model (MQSSM) for antiviral peptides (AVPs) prediction.

2.2. Construction of Query/Reference Datasets

When constructing a Query Dataset, a critical consideration is its ability to encompass a significant portion of the antiviral active chemical space without disproportionately representing any category of AVPs or neglecting rare sequences. To achieve this, we leveraged the reductions obtained in a previous study [30].

2.2.1. Recalling the Scaffold Extraction Procedure

To ensure a comprehensive coverage of antiviral active chemical space, we established a Half-Space Proximal Network (HSPN) using the 4,663 antiviral sequences stored in StarPepDB [31], recognized as one of the most extensive databases of biologically active peptides to date (36). Several scaffold (representative AVPs subsets) extractions were executed using the resources available in the StarPep toolbox. During this process, we adjusted the alignment algorithm type, the pairwise identity percentage, and the centrality measure (Community Hub-Bridge [32] and Harmonica [33,34]). These adjustments yielded a total of 20 scaffolds or representative AVPs subsets. Subsequently, we thoroughly examined these scaffolds using a Dover Analyzer [35] to assess the extent of similarity overlap between them.

 doi:10.20944/preprints202407.1476.v1

Based on these observations, we selected 10 scaffolds for their representativeness and diversity. Furthermore, an additional five scaffolds were crafted by combining some of the 20 previously generated scaffolds, considering their pairwise identity percentages. These 15 scaffolds served as the foundational datasets used as references for the MQSSMs. The particularities of each of the mentioned scaffolds are briefly explained in Table SI1.1 of the Supporting Information (SI) file 1 (FileSI1).

2.3. Target Datasets for Calibration and Validation of MQSSMs

We conducted a comprehensive literature review to gather a range of positive and negative sequences for testing our models. Ultimately, we selected 15 datasets, containing positive and negative sequences from synthetic and natural origin. These datasets from literature played a crucial role in various stages of model calibration and validation. All the mentioned datasets are depicted in Table 1.

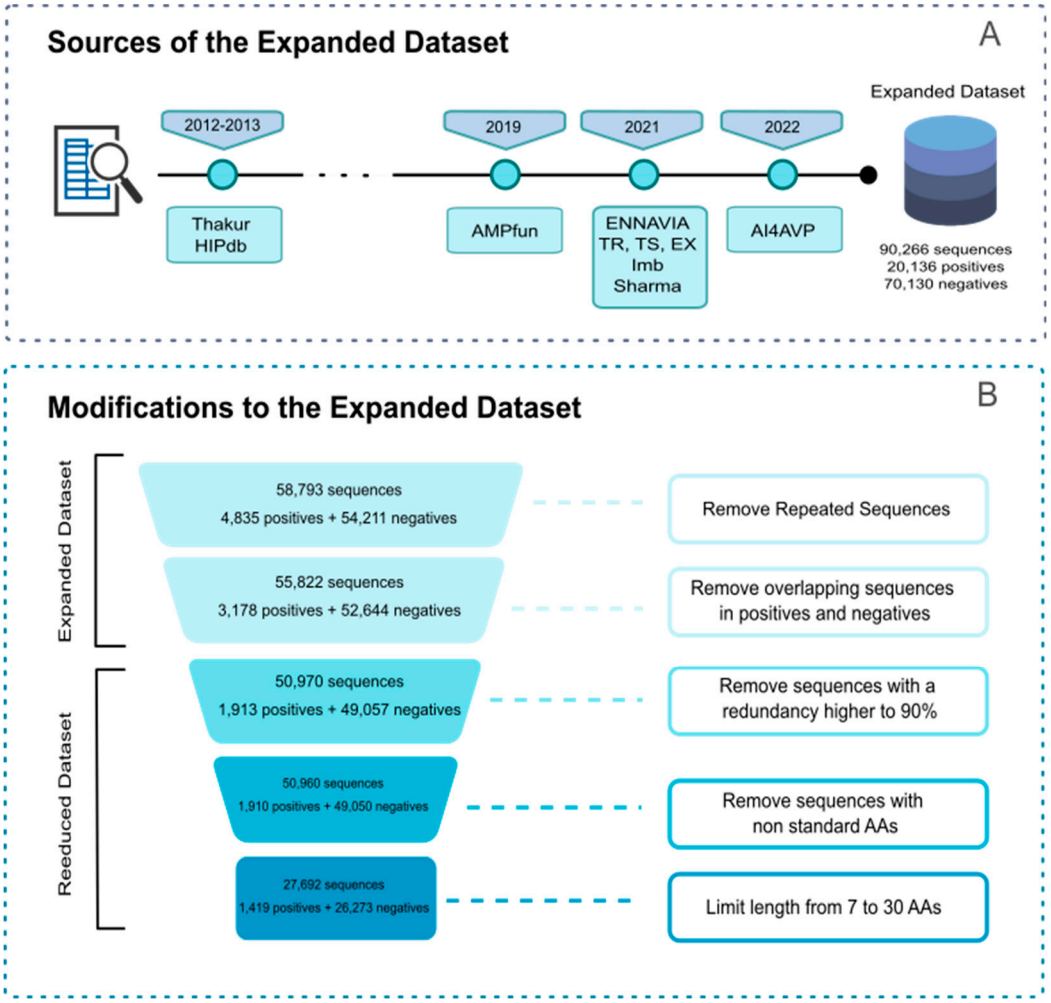
Table 1. List of All the Datasets Used for Calibration and Validation Stages.

Dataset	Size	Positives	Negatives	Ref
TR_StarPep	4,642	2,321	2,321	
TS_StarPep	1,246	623	623	[36]
Ex_Starpep	12,001	1,230	10,771	
AVPIden	53,113	2,662	51,116	[37]
AMPfun	5,826	2,001	3,825	[18]
ENNAVIA-A	974	557	420	
ENNAVIA-B	1,154	557	597	[19]
ENNAVIA-C	465	109	356	
ENNAVIA-D	469	110	359	
Imb	12,234	2,038 139 (Anti-CoV)	10,196	[38]
Thakur	1,056	604	452	[14]
Sharma	6,544	3,273	3,271	[39]
AI4AVP	20,222	2,934	17,288	[21]
Hipdb	981	981	-	[40]
Expanded	55,822	3,178	52,644	-
Reduced	27,692	1,419	26,273	-

Furthermore, we created two new datasets, amassing a total of 20,136 positive sequences and 70,130 negative sequences. These 90,266 sequences underwent several filters to eliminate redundancy, resulting in an "Expanded" Dataset comprising 55,822 sequences, including 3,178 positive and 52,644 negative sequences. The different sources for the “Expanded” Dataset are specified in Scheme 2.

Additionally, for benchmarking against state-of-the-art methods, we curated a "Reduced" Dataset with specific criteria, excluding non-standard amino acids and restricting sequence lengths to between 7 and 30 amino acids, as commonly specified by many existing predictors. Refer to Scheme 2 for a better depiction of the distinct filters that this dataset went through. The "Reduced" Dataset consists of 27,692 sequences, encompassing 1,419 positive and 26,273 negative sequences. Both, the “Expanded and Reduced” versions can be found at the supplementary materials (File SI2).

 doi:10.20944/preprints202407.1476.v1



Scheme 2. (A) Sources for the "Expanded" Dataset, (B) Filter Applied to Obtain "Reduced" Dataset.

2.4. Construction, Selection and, Improvement of MQSSMs

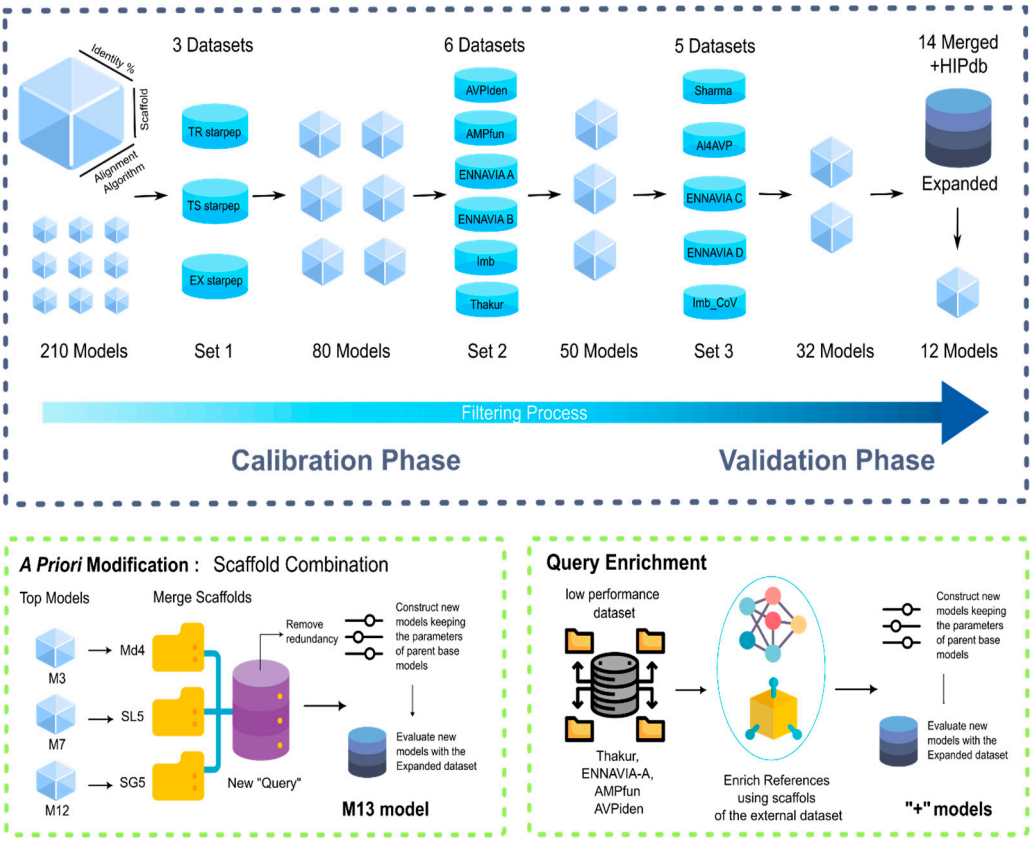
As mentioned earlier, three primary parameters require fine-tuning for the construction of Multi-Query Similarity Search Models (MQSSMs). The first parameter is the choice of the Query Dataset, for which 15 different scaffolds were evaluated. The second parameter is the selection of the alignment algorithm, alternating between global and local alignments. Lastly, the similarity threshold was varied from 0.3 to 0.9. These parameters collectively allowed for the potential creation of 210 variations of MQSSMs. Consequently, reducing this number of models was imperative.

To select the best-performing models, a two-phase approach was implemented, as depicted in Scheme 3. The first phase, known as the calibration phase, aimed to significantly reduce the number of models and identify key trends in parameter calibration. The calibration phase was further divided into two rounds. In the first round, 210 models were assessed using the datasets TS_StarPep, TR_StarPep, and Ex_StarPep, resulting in a reduction to 80 models. The second round of the calibration phase involved testing the models against six different datasets: AVPIden, AMPfun, ENNAVIA-A, ENNAVIA-B, Imb, and Thakur. After this evaluation, a set of 50 models remained.

These 50 models subsequently entered the validation stage, where the primary objective was to fine-tune parameter values while also assessing the models against more specific datasets. Once again, the validation phase was subdivided into two rounds. In the first round, the 50 models were tested against five datasets: Sharma, AI4AVP, ENNAVIA-C, ENNAVIA-D, and Imb_CoV, with the last three containing Anti-SARS-CoV sequences. This round led to the selection of 32 models, which were further evaluated against the Expanded Datasets. Following this analysis, a final set of 12 models was chosen, comprising six that utilized global alignment and six that employed local alignment (M1-M12).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 July 2024 doi:10.20944/preprints202407.1476.v1

As 12 models still represented a considerable number, an examination was conducted to assess the degree of overlap in the sequences recovered by these different models. After a cursory examination, three base models were selected referred as M3, M7, M12.



Scheme 3. Workflow for the Selection and Validation of Models. First Section of the Scheme Summarizes the Model's Selection Process. The Second Section Focuses on the Model's Improvement.

2.5. Scaffold Fusion

To enhance the model's performance, a scaffold fusion was performed with the aim of expanding the reference dataset. This fusion process involved combining the individual scaffolds from the best-performing base models (M3, M7, M12) to create a new query dataset. Once this new query dataset was constructed, 14 MQSSMs were generated using it, including variations in the alignment algorithm and the similarity threshold during the creation of these models.

Subsequently, these models underwent the previously described workflow of calibration and validation. Following this process, the best-performing model was identified and is now referred to as M13 (Scheme 3).

2.6. Scaffold Enrichment

To further enhance the base models, a query enrichment strategy was implemented. This involved gathering the positive sequences from the datasets used in Calibration Phase, round 2 (Thakur, ENNAVIA-A, AMPfun, and Imb) into a single dataset. This process resulted in 2403 unique sequences. To ensure the absence of these sequences from the currently top-performing model's scaffolds, a pairwise similarity comparison was conducted using Dover Analyzer (REF). Subsequently, these sequences were used to construct a Half-Space Proximal Network (HSPN), similarly to what was done for the StarPepDB AVPs sequences.

With the newly curated and validated dataset in hand, it was integrated into the StarPep toolbox to construct an HSPN, similarly to the HSPNs developed previously for the entire AVP space.

 doi:10.20944/preprints202407.1476.v1

Following HSPN construction, Scaffold Extraction was applied to the network, varying the centrality measure between Harmonica and Community Hub-Bridge. The alignment algorithm type was adjusted between local and global, considering only 80% and 90% as the sequence identity percentages. This process yielded 8 new scaffolds, designated as" external scaffolds." These scaffolds were utilized to develop new MQSSMs employing the same methodology as before, resulting in 112 new models. These newly generated models underwent testing against the Calibration and Validation workflow of 15 datasets to identify the most effective external scaffolds.

Finally, two models, named E1 and E2, were selected. The scaffolds from models E1 and E2 were added to those of models M3, M7, M12, and M13, ensuring that there were no repeated sequences in the enriched scaffolds. Once the enriched scaffolds were obtained, the models retained the alignment type and similarity cutoff of the base models, and they were tested against the 14 individual datasets and the expanded dataset (Scheme 3).

2.7. Performance Evaluation

The performance of all the models and state-of-the-art models was measure using the metrics of Accuracy (ACC), Sensitivity (SN), Specificity (SP), Mathews Correlation Coefficient (MCC) , False Positive Rate (FPR) and F1 Score [41]:

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

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

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

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

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

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

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

where, TP are the True Positives, TN are the True Negatives, the FP are the False Positives and FN are the False Negatives.

Additionally, the Friedman test was employed to analyse the performance rankings of the various MQSS models across all described metrics. The corresponding significance test was carried out using the KEEL software ([42], https://sci2s.ugr.es/keel/development.phpx1-2000) and the Non-parametric Statistical Analysis Module. This non-parametric test is well-suited for comparing multiple models across different datasets or conditions, especially when dealing with ranked data, as in the present case.

2.8. Comparison with State-of-the-Art

To assess the robustness of the MQSSMs, a selection of state-of-the-art predictors was made, primarily considering their availability through web services and ease of implementation. The Table 2 provides a summary of the predictors used in this section. For the comparison with the predictors the Reduced Dataset was employed as previously stated

 doi:10.20944/preprints202407.1476.v1

Table 2. Web Servers and Stand-Alone Predictors Used for Comparison.

Predictor	Year	Algorithm	Implementation	Ref
AI4AVP	2022	CNN	https://axp.iis.sinica.edu.tw/AI4AVP/	[21]
iACVP	2022	RF	http://kurata35.bio.kyutech.ac.jp/iACVP/	[16]
PTPAMP	2022	SVM	http://www.nipgr.ac.in/PTPAMP/	[12]
seqpros	2022	MLP, LSTM	https://github.com/cotovic/seqpropstherapeutic	[43]
ProDcal	2021	RF, RNN	https://biocom-ampdiscover.cicese.mx/	[36]
AMPfun	2020	RF	http://fdblab.csie.ncu.edu.tw/AMPfun/index.html	[18]
FIRM-AVP	2020	RF SVM, DL	https://github.com/pmartR/FIRM-AVP	[44]
Meta-iAVP	2019	hybrid	http://codes.bio/meta-iavp/	[17]
AntiVPP	2019	RF	https://github.com/bio-coding/AntiVPP	[45]
ClassAMP	2012	RF SVM	http://www.bicnirrh.res.in/classamp/	[13]
AVPpred	2012	SVM	http://erdd.osdd.net/servers/avppred/	[14]

3. Results and Discussion

3.1. Performance of MQSSMs at the Calibration Phase

This segment focuses on managing the initial amount of 210 models via a reduction process. In the preliminary phase of calibration, an evaluation of these 210 models was conducted utilizing three datasets: TR Starpep, TS StarPep, and EX Starpep. These datasets, predominantly derived from StarPepDB, demonstrated a relatively uniform behavior across models' performance, as depicted in Figure 1. Such an outcome is expected considering the considerable overlap in sequences shared between these datasets and the employed "Query."

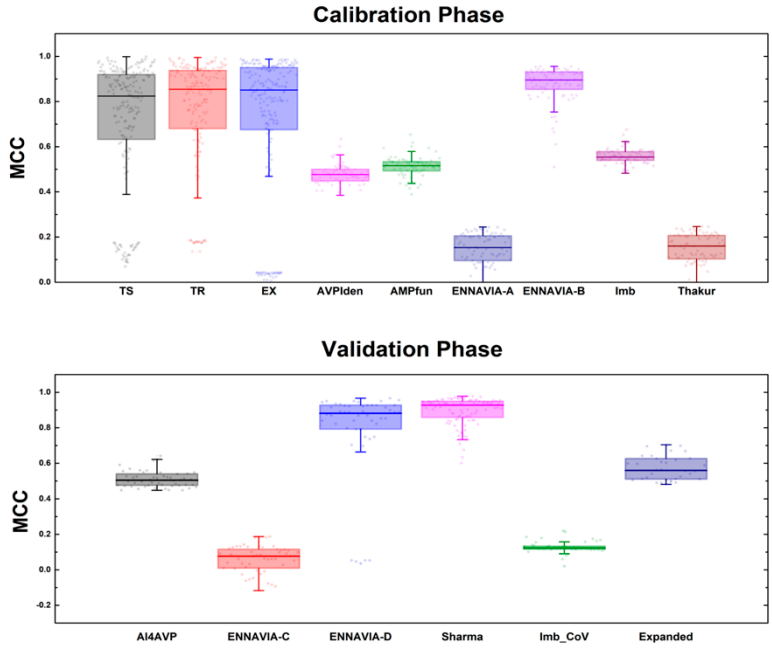


Figure 1. MQSS Models Mathew's Correlation Coefficient (MCC). Distribution in each of the 15 tested datasets in the different station of model selection.

During the next phase of calibration, six datasets not associated with StarPepDB were incorporated (Figure 1). Predictably, in this phase, the models reduced their effectiveness compared to the initial round. A critical revelation from this stage was a better performance of models when dealing with datasets composed of both randomly generated negative sequences and sequences yet to be experimentally tested. In contrast, when the datasets had experimentally validated sequences, the performance metrics dropped. This behaviour might be linked to the fact that a significant

 doi:10.20944/preprints202407.1476.v1

number of these negative sequences closely resembled the positive ones. Consequently, the methods based on alignment faced challenges in differentiating unique attributes of each category.

This problem was especially pronounced in datasets like ENNAVIA-A and Thakur, where experimental sequences were used as negative datasets. Conversely, with ENNAVIA-B, which contained the same positive sequences, the models more effectively identified non-antiviral sequences. It is important to emphasize that experimentally validated negative sequences are scarce in comparison to positive sequences. Hence, the predominant challenge in modelling lay in improving the detection rate (recall) of positive sequences overall.

From the calibration stage, clear trends in the MQSSMs parameters emerged. A notable observation was that models with larger scaffolds performed better. This improvement was attributed to the more detailed characterization of the AVPs' chemical space. Scaffolds such as Md4, Md5, SG4, SG5, SL4, and SL5 generated the most model variants (Table SI1.1). In terms of alignment, global alignment proved more effective with lower sequence identity percentages, while local alignments were more successful with higher percentages. For simpler scaffolds, global alignments consistently outperformed local alignments, regardless of the identity percentage.

3.2. Performance of MQSSMs at the Validation Phase

In the validation stage, we assessed the models against datasets with sequences targeting specific virus like SARS-CoV. Datasets including ENNAVIA-C, ENNAVIA-D, and Imb_CoV were used for this purpose. As shown in the Figure 1, the base models had poor performance with ENNAVIA-C and Imb_CoV. This result was expected because the model references are from 2019. This highlights the importance of the "Query" dataset's representation in model performance. However, the models showed better performance on the ENNAVIA-D dataset, which contained random negative sequences. This aligns with previous observations of model behaviour in similar datasets.

During the model selection process, 32 models chosen in the first round of the validation were tested against the Expanded dataset. From this, 12 models were identified as top performers using a multi-variable Friedman ranking method. These models, labelled M1 to M12, included 6 based on global alignment and 6 on local alignment strategies. The parameters for these models are summarized in the Table SI1.2. Additionally, from these 12 models, the top 3 performing ones were further singled out. This selection focused efforts on fine-tuning these models, particularly towards enhancing the recovery of positive (AVP) sequences.

3.3. Improving MQSSMs Performance by Fusing Scaffolds

Continuing from the focused analysis of the top models for positive sequence recovery, an initial approach to enhancement involved combining the scaffolds from models M3, M7, and M12 (md4, SL5, SG5) into a single, consolidated scaffold. This step included the removal of duplicate sequences, culminating in a scaffold that contained 3206 unique sequences. Subsequent testing of this modified scaffold indicated a slight improvement in the performance of models using global alignment with a 90% similarity threshold. Labelled as M13, a new model was crafted using these parameters. The improvements in M13 were primarily due to the expanded representation of sequence space, enriched by the increased number of sequences.

Although the modifications provided insights into the nuances of the MQSSMs, their overall efficacy remained unsatisfactory. The subsequent strategy aimed to leverage the performance information gathered during the calibration phase. As shown in Figure 1, the base models struggled with datasets such as Thakur, ENNAVIA-A, AMPfun, and AVPiden. This struggle was likely due to the inadequate representation of diverse sequences from these datasets in the MQSS scaffolds. Furthermore, the presence of many experimentally validated negative sequences in some datasets increased the difficulty of making accurate predictions.

 doi:10.20944/preprints202407.1476.v1

3.4. Improving MQSSMs Performance by Enriching the Best Scaffolds

In tackling the identified issues, a Half Space Proximal Network (HSPN) was made by pooling together positive sequences from the challenging datasets (Thakur, ENNAVIA-A, AMPfun, and AVPiden). A total of 2403 sequences were employed in constructing the HSPN. This effort resulted in the production of 8 scaffolds, out of which the top two were chosen to enhance the existing scaffolds. Consequently, this led to the introduction of new, improved models: M3+, M7+, M12+, and M13+. The "+" in their names signifies their references enrichment. Post-enhancement, these scaffolds contained 3155, 3437, 3472, and 3606 sequences, respectively. To incorporate unique scaffolds do not present in StarPepDB, we crafted E1 and E2 models using an external scaffold, comprising 1517 and 1261 sequences, respectively. This increased our analysis pool to 10 models, adding 6 newly enriched models to the pre-existing M3, M7, M12, and M13. Complete details of these models are outlined in Table SI1.2.

Following their development, the 10 models were extensively tested across 15 databases, the 14 datasets from our workflow plus the Expanded Dataset (File SI3). We used a Friedman ranking system to halve the number of top-performing models, evaluating them based on accuracy (ACC), specificity (SP), sensitivity (SN), Matthew's correlation coefficient (MCC), and F1 score. The ranking results identified M3+, M13+, M7, M12, and E1 as the most effective, as highlighted in grey in Table 3.

Table 3. Evaluation Metrics for the Top 10 Performing Models[a].

Model	ACC	SP	SN	MCC	FPR	F1
E1	0.966	0.995	0.481	0.624	0.005	0.614
E2	0.961	0.995	0.398	0.562	0.005	0.54
M12	0.958	0.962	0.891	0.704	0.038	0.708
M12+	0.736	0.724	0.937	0.33	0.276	0.288
M13	0.964	0.98	0.694	0.667	0.02	0.686
M13+	0.969	0.979	0.802	0.731	0.021	0.746
M3	0.935	0.944	0.782	0.568	0.056	0.577
M3+	0.935	0.939	0.876	0.609	0.061	0.606
M7	0.958	0.964	0.873	0.699	0.036	0.705
M7+	0.736	0.724	0.933	0.329	0.276	0.287

[a]ACC = accuracy, SP= specificity, SN =sensitivity, MCC =Mathew's correlation Coefficient, FPR= False Positive Rate, F1 Score.

Notably, while models with a greater number of reference sequences like M3+ and M13+ ranked high, models such as E1, M7, and M12 with fewer sequences performed comparably well. This suggests that the effectiveness of the references hinges more on their diversity and representational range than on their quantity.

3.5. Benchmarking the Best MQSSMs against State-of-the-Art Predictors

The top 5 models were subsequently benchmarked against existing predictors in the literature, providing a comparative assessment of their performance relative to other available tools. To ensure unbiased comparisons, any sequences shared between the scaffolds of models M3+, M7, M12, M13+, E1, and the Reduced dataset were eliminated. This action reduced the number of positive sequences to 116, while the number of negative sequences remained unchanged. It is important to note that many negative sequences in the Reduced dataset are part of the training sets for the external predictors, but these were not removed. This decision placed our models at a significant disadvantage in terms of performance evaluation. We tested a total of 14 external predictors (File SI4), using evaluation metrics such as ACC, SP, SN, MCC, FPR and F1 Score. MCC, which is unaffected by the dataset's imbalance, was the primary metric for our analysis.

 doi:10.20944/preprints202407.1476.v1

The alteration made to the Reduced dataset resulted in a marked decrease in the performance of the MQSSMs. This decline was especially pronounced in the SN and MCC values. However, ACC and SP remained relatively stable, a situation due to the substantial imbalance between positive and negative cases in the class distribution. The notable drop in sensitivity underscores a significant, consistent shortfall in correctly identifying positive sequences. This inability to accurately recall true positives notably affected the MCC, with a stark decrease observed, for example, in the M13+ model where MCC plummeted from 0.731 to 0.214, as detailed in Table 4. The F1 score, which depends on both recall and precision, also experienced a corresponding decline.

Table 4. Evaluation Metrics of the Top Five Performing MQSS Models and *state-of-the-art predictors*.[a]

Model	ACC	SP	SN	MCC	FPR	F1
M3+	0.929	0.93	0.603	0.137	0.07	0.069
M7	0.97	0.972	0.448	0.163	0.028	0.115
M12	0.968	0.971	0.466	0.165	0.029	0.114
M13+	0.983	0.986	0.422	**0.214**	0.014	0.18
E1	0.993	**0.996**	0.19	0.184	0.004	**0.187**
AI4AVP	0.387	0.385	0.905	0.039	0.615	0.013
AI4AVP(DA)	0.379	0.376	0.871	0.034	0.624	0.012
FIRM-AVP	0.647	0.647	0.595	0.034	0.353	0.015
Meta-iAVP	0.594	0.593	0.647	0.032	0.407	0.014
seqpros	0.119	0.116	**0.94**	0.011	0.884	0.009
AMPfun	0.463	0.462	0.784	0.033	0.538	0.013
iACVP	0.893	0.895	0.517	0.088	0.105	0.041
PTPAMP	0.825	0.827	0.336	0.028	0.173	0.017
ClassAMP	0.795	0.798	0.31	0.018	0.202	0.013
AntiVPP	0.732	0.734	0.457	0.028	0.266	0.015
ProtDcalRF	**0.995**	1	0	-0.001	**0**	0
ProtDcalHier	**0.995**	**0.999**	0	-0.002	0.001	0
ProtDcalRNN	0.95	0.954	0.034	-0.004	0.046	0.006
AVPpred	0.902	0.904	0.371	0.062	0.096	0.032

[a]ACC = accuracy, SP= specificity, SN =sensitivity, MCC =Mathew's correlation Coefficient, FPR= False Positive Rate, F1 Score.

Despite the less-than-ideal results, the MQSSMs still surpassed the external predictors in overall performance. Analysing the performance data of the external predictors, Figure 2 highlights two distinct trends. Some models are highly effective at identifying most positive sequences, resulting in a high SN but with the trade-off of a higher rate of false positives. Conversely, other models excel at correctly classifying all negative sequences but tend to misclassify many positive ones, a trait observed in the MQSSMs we developed. Typically, most models based on deep learning fall into the former category, demonstrating high sensitivity, whereas traditional ML models are more likely to be in the latter category, with a stronger emphasis on specificity.

 doi:10.20944/preprints202407.1476.v1

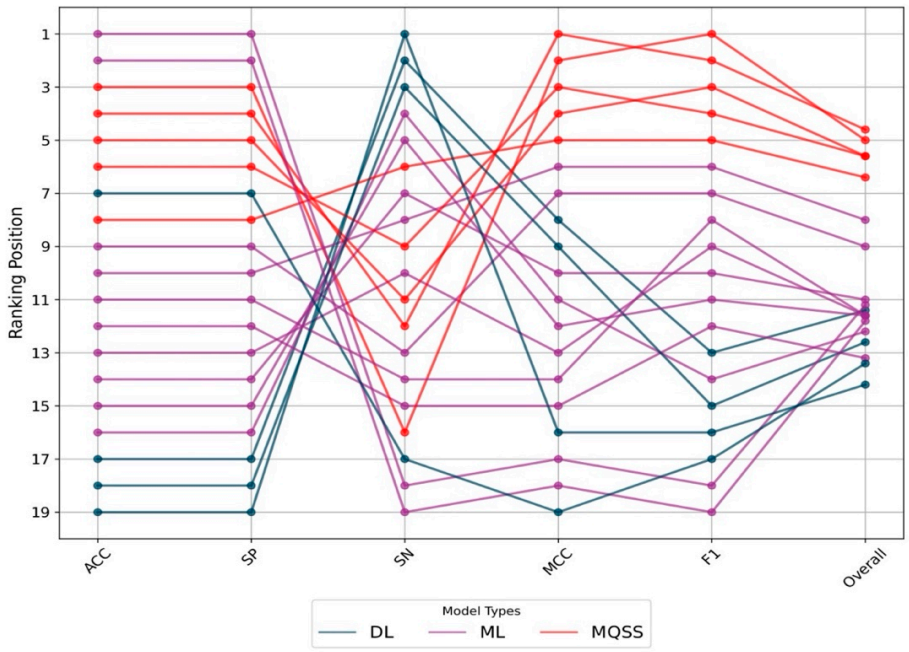


Figure 2. Ranking Fluctuation in Each of the Performance Metrics for Deep Learning (DL), Machine Learning (ML) and MQSS Model. ACC = accuracy, SP= specificity, SN =sensitivity, MCC =Mathew's correlation Coefficient, F1 Score.

The analysis revealed that no single predictor excelled in all evaluation categories, supporting Garcia-Jacas et al.'s conclusion that DL methods may not be the most effective for AVP prediction [36]. Extending this observation, it's evident that none of the ML models tested demonstrated satisfactory performance, suggesting a need for significant improvements. A central issue identified is the quality and representativeness of the training data. Most positive sequences used in training exhibit high similarity, often up to 90%. Moreover, the experimentally validated negative sequences often mirror their positive counterparts. The challenge, therefore, does not lie in the complexity of the models' architectures but rather in the data's availability and diversity. This highlights an ongoing challenge in gathering and effectively utilizing comprehensive data for such models.

Evaluating *state-of-the-art* predictors revealed a common issue of accessibility. Many predictors were difficult to assess for various reasons. A significant issue was the poor construction of several web servers, leading to operational failures or frequent malfunctions. This problem affected even relatively new servers, those less than 2 years old. Furthermore, a few research repositories lacked comprehensive instructions, complicating their use. This difficulty in accessing and implementing these tools echoes concerns previously noted by researchers [46] about the availability of source codes. Nonetheless a comprehensive summary of the prediction tools currently available is provided in the supplementary information (Table SI1.4).

In contrast to these challenges, the MQSSMs distinguish themselves with their accessibility. They are available through the StarPep toolbox standalone software, which boasts a user-friendly interface, making these models more approachable and easier to use for users.

Despite the need for improvements to address their shortcomings, the MQSSMs still have a performance edge over traditional ML models. This is substantiated by a Friedman MCC, ACC, SP, SN, and the F1 Score. An important aspect to note is that the MQSSMs require considerably fewer computational resources and are not constrained by sequence length limitations, presenting significant benefits.

When choosing a prediction model, it's essential to align with the specific needs of the researcher. Some may prioritize identifying a larger number of potential AVPs, while others might prefer a smaller, more accurate set to reduce false positives. Considering the resource-intensive nature of experimental procedures, the latter approach is often more practical for synthesizing

 doi:10.20944/preprints202407.1476.v1

potential AVPs, as it optimizes the balance between resource use and the likelihood of accurate predictions.

4. Conclusions

This research marks a significant advancement with the development of Multi-Query Similarity Search Models (MQSSMs). These models, devised from a deep understanding gained in the chemical space exploration and scaffold extraction phases, employ a novel approach that leverages structural similarities for predicting biological activities. This has been instrumental in improving model selection.

Another major achievement of this study is the assembly of the most extensive datasets of AVPs, referred to as the "Expanded" and "Reduced" datasets. These datasets to were designed to cater to a spectrum of research needs in the field AVP modelling and prediction and are available at the File SI2 in supplementary materials.

Such rich data diversity is essential for the development of effective AVP predictors. Following comprehensive evaluations and filtering processes, five MQSSMs were selected for their superior performance. These models were rigorously tested across various metrics, with the top model demonstrating impressive results: ACC=0.969, SP=0.979, SN=0.802, MCC=0.71, FPR=0.021, F1=0.746. In contrast, 14 contemporary ML-based predictors, though extensively promoted, were surpassed by the MQSSMs. This not only emphasizes the limitations of existing methods but also highlights the advanced capabilities of MQSSMs in predicting AVP sequences, effectively handling the challenges of variable-length sequences and imbalanced datasets. As these models notably eclipsed existing *state-of-the-art* predictors, they have set a new standard in the field of AVPs discovery.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. File SI1: Supporting Tables (Table SI1.1 Brief Description of the initial scaffolds used for the MQSSMs, TableSI1.2 Parameters used for selected MQSSMs, Table SI1.3 Final Overall Ranking, Table SI1.4 Web server and Stand-Alone Software for the prediction of AVPs. File SI2: Datasets used for calibration and validations satges of the MQSSMs. File SI3: Classification results of the 10-top perfoming MQSSMs on 14 datasets. File SI4: Classification results of 14 web-based AVP predictors on the "Reduced Dataset".

Author Contributions: Y.M.-P and G.A.-C contributed to the conceptualization, methodology, supervision, drafting, and reviewing of the manuscript. D.dL.-G was responsible for data curation, model building, validation, and data/results visualization. F.M.-R and FJ.F participated in programming tasks and data analysis. A.A. and H.R. were responsible for funding acquisition, supervision, and reviewing the manuscript. All authors have read and approved the final version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The presented MQSSMs can be easily run using the StarPep toolbox, which is freely available at: https://github.com/Grupo-Medicina-Molecular-y-Traslacional/StarPep. The 4,663 AVPs stored in StarPepDB can be accessed and downloaded directly through the StarPep toolbox. The Half Space Proximal Networks (HSPNs) are also constructed using the StarPep toolbox. The datasets used in the calibration and validation stages of the MQSSMs, as well as those used for performance comparison with existing tools, are included as part of the supplementary files in this work.

Acknowledgments: Y.M.-P. acknowledges support from USFQ "MED Grant 2023-4 (Project ID23234). M-P. Y also thanks the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to work at Valencia University 2024-2025. G.A.-C. and A.A. acknowledge the support of the FCT – Foundation for Science and Technology (Portugal) within the scope of UIDB/04423/913 2020 and UIDP/04423/2020.

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

References

1. Enquist, L.W. Virology in the 21st Century. *J Virol* **2009**, *83*, 5296–5308. https://doi.org/10.1128/JVI.00151-09.

2. 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*, 48. https://doi.org/10.1038/s41392-022-00904-4.
3. Sanjuán, R.; Domingo-Calap, P. Mechanisms of Viral Mutation. *Cell. Mol. Life Sci.* **2016**, *73*, 4433–4448. https://doi.org/10.1007/s00018-016-2299-6.
4. Mahmoud, A. New Vaccines: Challenges of Discovery. *Microb. Biotechnol.* **2016**, *9*, 549–552. https://doi.org/10.1111/1751-7915.12397.
5. P. Carter, E.; G. Ang, C.; M. Chaiken, I. Peptide Triazole Inhibitors of HIV-1: Hijackers of Env Metastability. *CPPS* **2023**, *24*, 59–77. https://doi.org/10.2174/1389203723666220610120927.
6. Agamennone, M.; Fantacuzzi, M.; Vivenzio, G.; Scala, M.C.; Campiglia, P.; Superti, F.; Sala, M. Antiviral Peptides as Anti-Influenza Agents. *IJMS* **2022**, *23*, 11433. https://doi.org/10.3390/ijms231911433.
7. Divyashree, M.; Mani, M.K.; Reddy, D.; Kumavath, R.; Ghosh, P.; Azevedo, V.; Barh, D. Clinical Applications of Antimicrobial Peptides (AMPs): Where Do We Stand Now? *PPL* **2020**, *27*, 120–134. https://doi.org/10.2174/0929866526666190925152957.
8. Yu, Y.; Cooper, C.L.; Wang, G.; Morwitzer, M.J.; Kota, K.; Tran, J.P.; Bradfute, S.B.; Liu, Y.; Shao, J.; Zhang, A.K.; et al. Engineered Human Cathelicidin Antimicrobial Peptides Inhibit Ebola Virus Infection. *iScience* **2020**, *23*, 100999. https://doi.org/10.1016/j.isci.2020.100999.
9. Jackman, J.A.; Costa, V.V.; Park, S.; Real, A.L.C.V.; Park, J.H.; Cardozo, P.L.; Ferhan, A.R.; Olmo, I.G.; Moreira, T.P.; Bambirra, J.L.; et al. Therapeutic Treatment of Zika Virus Infection Using a Brain-Penetrating Antiviral Peptide. *Nature Mater* **2018**, *17*, 971–977. https://doi.org/10.1038/s41563-018-0194-2.
10. Vilas Boas, L.C.P.; Campos, M.L.; Berlanda, R.L.A.; de Carvalho Neves, N.; Franco, O.L. Antiviral Peptides as Promising Therapeutic Drugs. *Cell. Mol. Life Sci.* **2019**, *76*, 3525–3542. https://doi.org/10.1007/s00018-019-03138-w.
11. Lau, J.L.; Dunn, M.K. Therapeutic Peptides: Historical Perspectives, Current Development Trends, and Future Directions. *Bioorganic & Medicinal Chemistry* **2018**, *26*, 2700–2707. https://doi.org/10.1016/j.bmc.2017.06.052.
12. Jaiswal, M.; Singh, A.; Kumar, S. PTPAMP: Prediction Tool for Plant-Derived Antimicrobial Peptides. *Amino Acids* **2023**, *55*, 1–17. https://doi.org/10.1007/s00726-022-03190-0.
13. Joseph, S.; Karnik, S.; Nilawe, P.; Jayaraman, V.K.; Idicula-Thomas, S. ClassAMP: A Prediction Tool for Classification of Antimicrobial Peptides. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **2012**, *9*, 1535–1538. https://doi.org/10.1109/TCBB.2012.89.
14. Thakur, N.; Qureshi, A.; Kumar, M. AVPpred: Collection and Prediction of Highly Effective Antiviral Peptides. *Nucleic Acids Research* **2012**, *40*, W199–W204. https://doi.org/10.1093/nar/gks450.
15. Xiao, X.; Wang, P.; Lin, W.-Z.; Jia, J.-H.; Chou, K.-C. iAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types. *Analytical Biochemistry* **2013**, *436*, 168–177. https://doi.org/10.1016/j.ab.2013.01.019.
16. Kurata, H.; Tsukiyama, S.; Manavalan, B. iACVP: Markedly Enhanced Identification of Anti-Coronavirus Peptides Using a Dataset-Specific Word2vec Model. *Briefings in Bioinformatics* **2022**, *23*, bbac265. https://doi.org/10.1093/bib/bbac265.
17. Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W. Meta-iAVP: A Sequence-Based Meta-Predictor for Improving the Prediction of Antiviral Peptides Using Effective Feature Representation. *IJMS* **2019**, *20*, 5743. https://doi.org/10.3390/ijms20225743.
18. 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. *Briefings in Bioinformatics* **2020**, *21*, 1098–1114. https://doi.org/10.1093/bib/bbz043.
19. Timmons, P.B.; Hewage, C.M. ENNAVIA Is a Novel Method Which Employs Neural Networks for Antiviral and Anti-Coronavirus Activity Prediction for Therapeutic Peptides. *Briefings in Bioinformatics* **2021**, *22*, bbab258. https://doi.org/10.1093/bib/bbab258.
20. Zhang, S.; Li, X. Pep-CNN: An Improved Convolutional Neural Network for Predicting Therapeutic Peptides. *Chemometrics and Intelligent Laboratory Systems* **2022**, *221*, 104490. https://doi.org/10.1016/j.chemolab.2022.104490.
21. Lin, T.-T.; Sun, Y.-Y.; Wang, C.-T.; Cheng, W.-C.; Lu, I.-H.; Lin, C.-Y.; Chen, S.-H. AI4AVP: An Antiviral Peptides Predictor in Deep Learning Approach with Generative Adversarial Network Data Augmentation. *Bioinformatics Advances* **2022**, *2*, vbac080. https://doi.org/10.1093/bioadv/vbac080.
22. Wei, L.; Zhou, C.; Su, R.; Zou, Q. PEPred-Suite: Improved and Robust Prediction of Therapeutic Peptides Using Adaptive Feature Representation Learning. *Bioinformatics* **2019**, *35*, 4272–4280. https://doi.org/10.1093/bioinformatics/btz246.
23. García-Jacas, C.R.; Pinacho-Castellanos, S.A.; García-González, L.A.; Brizuela, C.A. Do Deep Learning Models Make a Difference in the Identification of Antimicrobial Peptides? *Briefings in Bioinformatics* **2022**, *23*, bbac094. https://doi.org/10.1093/bib/bbac094.

24. Yan, J.; Cai, J.; Zhang, B.; Wang, Y.; Wong, D.F.; Siu, S.W.I. Recent Progress in the Discovery and Design of Antimicrobial Peptides Using Traditional Machine Learning and Deep Learning. *Antibiotics* **2022**, *11*, 1451. https://doi.org/10.3390/antibiotics11101451.
25. Basith, S.; Manavalan, B.; Hwan Shin, T.; Lee, G. Machine Intelligence in Peptide Therapeutics: A Next-generation Tool for Rapid Disease Screening. *Med Res Rev* **2020**, *40*, 1276–1314. https://doi.org/10.1002/med.21658.
26. Castillo-Mendieta, K.; Agüero-Chapin, G.; Santiago Vispo, N.; 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*; MATHEMATICS & COMPUTER SCIENCE, 2023;
27. 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*, 401. https://doi.org/10.3390/antibiotics11030401.
28. 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*, 46012–46036. https://doi.org/10.1021/acsomega.2c03398.
29. Aguilera-Mendoza, L.; Ayala-Ruano, S.; Martinez-Rios, F.; Chavez, E.; García-Jacas, C.R.; Brizuela, C.A.; Marrero-Ponce, Y. *StarPep Toolbox*: An Open-Source Software to Assist Chemical Space Analysis of Bioactive Peptides and Their Functions Using Complex Networks. *Bioinformatics* **2023**, *39*, btad506. https://doi.org/10.1093/bioinformatics/btad506.
30. De Llano García, D.; Rodríguez Cabrera, H.M.; Yachay, U. de I. de T.E. *A Novel Network Science and Similarity-Searching-Based Approach for Discovering Potential Antiviral Peptides* /; Urcuquí, 2023; http://repositorio.yachaytech.edu.ec/handle/123456789/698.
31. 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*, 4739–4747. https://doi.org/10.1093/bioinformatics/btz260.
32. 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*, 18074. https://doi.org/10.1038/s41598-020-75029-1.
33. Marchiori, M.; Latora, V. Harmony in the Small-World. *Physica A: Statistical Mechanics and its Applications* **2000**, *285*, 539–546. https://doi.org/10.1016/S0378-4371(00)00311-3.
34. Ruan, Y.; Tang, J.; Hu, Y.; Wang, H.; Bai, L. Efficient Algorithm for the Identification of Node Significance in Complex Network. *IEEE Access* **2020**, *8*, 28947–28955. https://doi.org/10.1109/ACCESS.2020.2972107.
35. 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*, 2553–2559. https://doi.org/10.1093/bioinformatics/btv180.
36. 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*, 3141–3157. https://doi.org/10.1021/acs.jcim.1c00251.
37. Pang, Y.; Yao, L.; Jhong, J.-H.; Wang, Z.; Lee, T.-Y. AVPIden: A New Scheme for Identification and Functional Prediction of Antiviral Peptides Based on Machine Learning Approaches. *Briefings in Bioinformatics* **2021**, *22*, bbab263. https://doi.org/10.1093/bib/bbab263.
38. Pang, Y.; Wang, Z.; Jhong, J.-H.; Lee, T.-Y. Identifying Anti-Coronavirus Peptides by Incorporating Different Negative Datasets and Imbalanced Learning Strategies. *Briefings in Bioinformatics* **2021**, *22*, 1085–1095. https://doi.org/10.1093/bib/bbaa423.
39. Sharma, R.; Shrivastava, S.; Singh, S.K.; Kumar, A.; Singh, A.K.; Saxena, S. Deep-AVPpred: Artificial Intelligence Driven Discovery of Peptide Drugs for Viral Infections. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 5067–5074. https://doi.org/10.1109/JBHI.2021.3130825.
40. Qureshi, A.; Thakur, N.; Kumar, M. HIPdb: A Database of Experimentally Validated HIV Inhibiting Peptides. *PLoS ONE* **2013**, *8*, e54908. https://doi.org/10.1371/journal.pone.0054908.
41. Chang, K.Y.; Yang, J.-R. Analysis and Prediction of Highly Effective Antiviral Peptides Based on Random Forests. *PLoS ONE* **2013**, *8*, e70166. https://doi.org/10.1371/journal.pone.0070166.
42. Alcalá-Fdez, J.; Sánchez, L.; García, S.; Del Jesus, M.J.; Ventura, S.; Garrell, J.M.; Otero, J.; Romero, C.; Bacardit, J.; Rivas, V.M.; et al. KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems. *Soft Comput* **2009**, *13*, 307–318. https://doi.org/10.1007/s00500-008-0323-y.
43. Otović, E.; Njirjak, M.; Kalafatovic, D.; Mauša, G. Sequential Properties Representation Scheme for Recurrent Neural Network-Based Prediction of Therapeutic Peptides. *J. Chem. Inf. Model.* **2022**, *62*, 2961–2972. https://doi.org/10.1021/acs.jcim.2c00526.

 doi:10.20944/preprints202407.1476.v1

44. Chowdhury, A.S.; Reehl, S.M.; Kehn-Hall, K.; Bishop, B.; Webb-Robertson, B.-J.M. Better Understanding and Prediction of Antiviral Peptides through Primary and Secondary Structure Feature Importance. *Sci Rep* **2020**, *10*, 19260. https://doi.org/10.1038/s41598-020-76161-8.
45. Beltrán Lissabet, J.F.; Belén, L.H.; Farias, J.G. AntiVPP 1.0: A Portable Tool for Prediction of Antiviral Peptides. *Computers in Biology and Medicine* **2019**, *107*, 127–130. https://doi.org/10.1016/j.compbiomed.2019.02.011.
46. Diakou, I.; Papakonstantinou, E.; Papageorgiou, L.; Pierouli, K.; Dragoumani, K.; Spandidos, D.; Bacopoulou, F.; Chrousos, G.; Eliopoulos, E.; Vlachakis, D. Novel Computational Pipelines in Antiviral Structure-based Drug Design (Review). *Biomed Rep* **2022**, *17*, 97. https://doi.org/10.3892/br.2022.1580.