JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

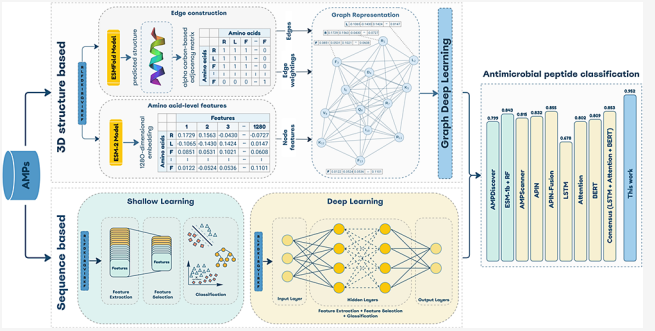
Predicting Antimicrobial Peptides Using ESMFold-Predicted Structures and ESM-2-Based Amino Acid Features with Graph Deep Learning

Greneter Cordoves-Delgado and César R. García-Jacas*

Cite This: J. Chem. Inf. Model. 2024, 64, 4310−4321 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Currently, antimicrobial resistance constitutes a serious threat to human health. Drugs based on antimicrobial peptides (AMPs) constitute one of the alternatives to address it. Shallow and deep learning (DL)-based models have mainly been built from amino acid sequences to predict AMPs. Recent advances in tertiary (3D) structure prediction have opened new opportunities in this field. In this sense, models based on graphs derived from predicted peptide structures have recently been proposed. However, these models are not in correspondence with state-of-the-art approaches to codify evolutionary information, and, in addition, they are memory- and time-consuming because depend on multiple sequence alignment. Herein, we presented a framework to create alignment-free models based on graph representations generated from ESMFold-predicted peptide structures, whose nodes are characterized with amino acid-level evolutionary information derived from the Evolutionary Scale Modeling (ESM-2) models. A graph attention network (GAT) was implemented to assess the usefulness of the framework in the AMP classification. To this end, a set comprised of 67,058 peptides was used. It was demonstrated that the proposed methodology allowed to build GAT models with generalization abilities consistently better than 20 state-of-the-art non-DL-based and DL-based models. The best GAT models were developed using evolutionary information derived from the 36- and 33-layer ESM-2 models. Similarity studies showed that the best-built GAT models codified different chemical spaces, and thus they were fused to significantly improve the classification. In general, the results suggest that esm-AxP-GDL is a promissory tool to develop good, structure-dependent, and alignment-free models that can be successfully applied in the screening of large data sets. This framework should not only be useful to classify AMPs but also for modeling other peptide and protein activities.



1. INTRODUCTION

Nowadays, antimicrobial resistance (AMR) constitutes a serious threat to human health,[1−3] killing at least 1.27 million people worldwide and related to approximately 5 million deaths in 2019.[4] According to data from 87 countries reporting in 2020 and 27 countries tracked since 2017,[1] AMR is causing life-threatening bloodstream infections and increasing treatment resistance for different ordinary infections. For example, compared with 2017, resistant *Escherichia coli* and *Salmonella* species, as well as *gonorrhea strains* led to bloodstream infections increased by at least 15% in 2020,[1] whereas resistance levels increased above 50% in bacteria that commonly trigger bloodstream infections, such as *Klebsiella pneumoniae* and *Acinetobacter* species.[1] Thus, several alternatives to conventional drugs/therapies are currently studied to address AMR, one of them being the discovery of drugs/therapies based on antimicrobial peptides (AMPs).[5−7]

Machine learning models have emerged as a time-saving and cost-effective tool for screening large data sets to identify potential AMPs. To date, machine learning models[8−11] based on amino acid sequences have mainly been built using traditional[12−17] and deep learning (DL)[18−28] techniques, as well as using similarity networks.[29,30] However, the outstanding results of deep neural network-based approaches, such as trRosetta,[31] AlphaFold,[32] RoseTTAFold,[33] ESMFold,[34] and HelixFold-Single,[35] in the prediction of tertiary (3D) structures of proteins from their amino acid sequences have unlocked new opportunities to build better predictive models. In this regard, non-DL based models using 3D protein descriptors[36−38] as well as Graph Neural Network-based models[39,40] (e.g., equivariant network) are promissory strategies to be developed.

Received: December 24, 2023
Revised: May 1, 2024
Accepted: May 1, 2024
Published: May 13, 2024

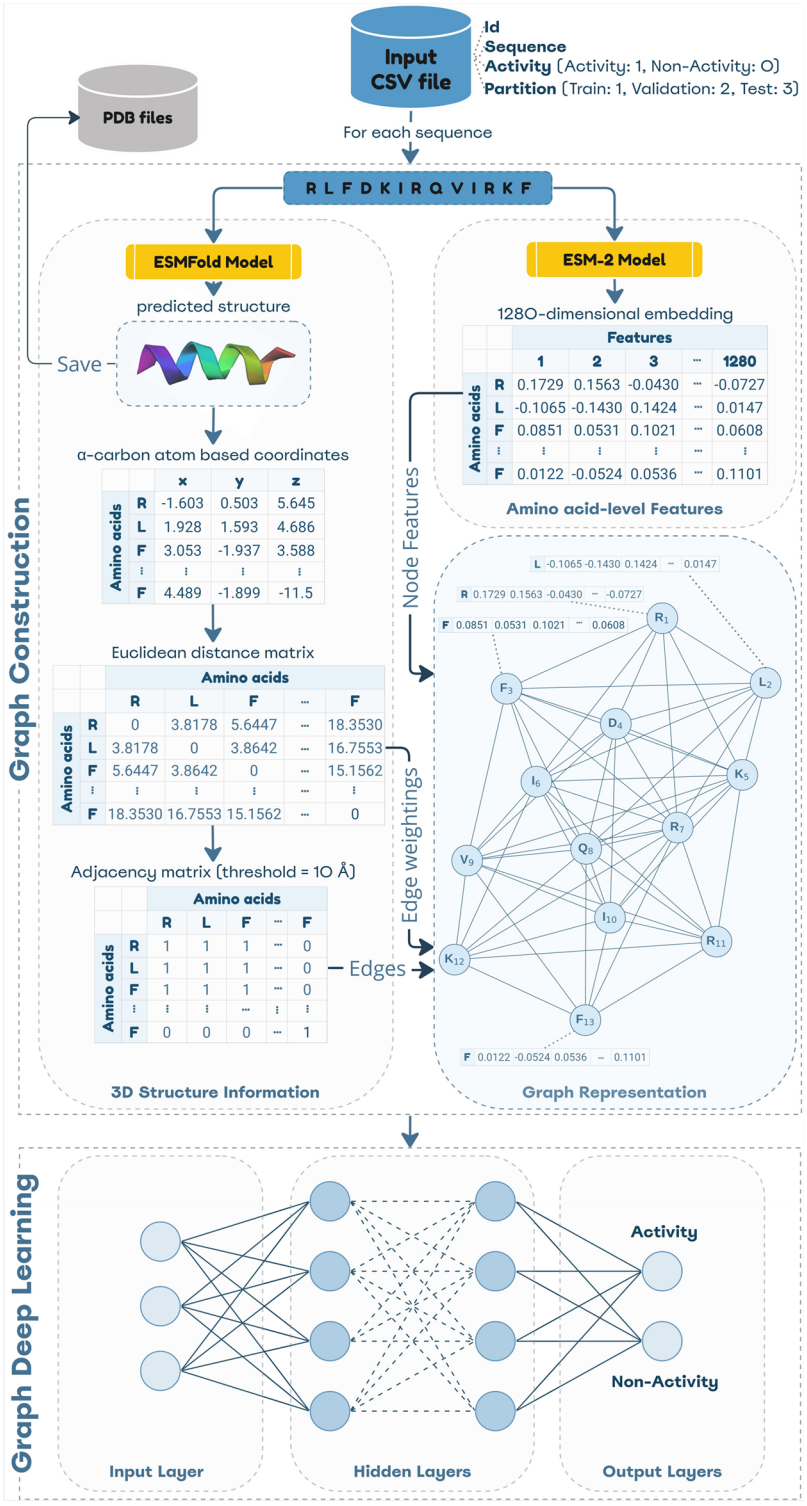


Figure 1. Overview of the workflow implemented in the esm-AxP-GDL framework.

Recently, Yan et al.[39] introduced a protocol (named sAMPpred-GAT) based on a Graph Attention Network (GAT) to classify general-AMPs, i.e., peptides that do not belong to a specific antimicrobial activity. This is the first one to leverage predicted peptide structures by applying a geometrical distance cutoff to build a graph-based representation per peptide structure. The created graphs, whose nodes (representing amino acids) are characterized with sequence information and evolutionary information, were used to feed a GAT architecture to learn discriminative features. However, this protocol has some disadvantages. On the one hand, sAMPpred-GAT uses amino acid-level evolutionary information derived from a Position-Specific Scoring Matrix (PSSM)[41] and, in addition, it is inputted with peptide structures predicted with the trRosetta[31] method. PSSM is obtained after applying a multiple sequence alignment (MSA), whereas trRosetta[31] uses MSA to derive 526 1-site and 2-site feature channels to feed a deep neural network to predict inter-amino acid geometries.

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

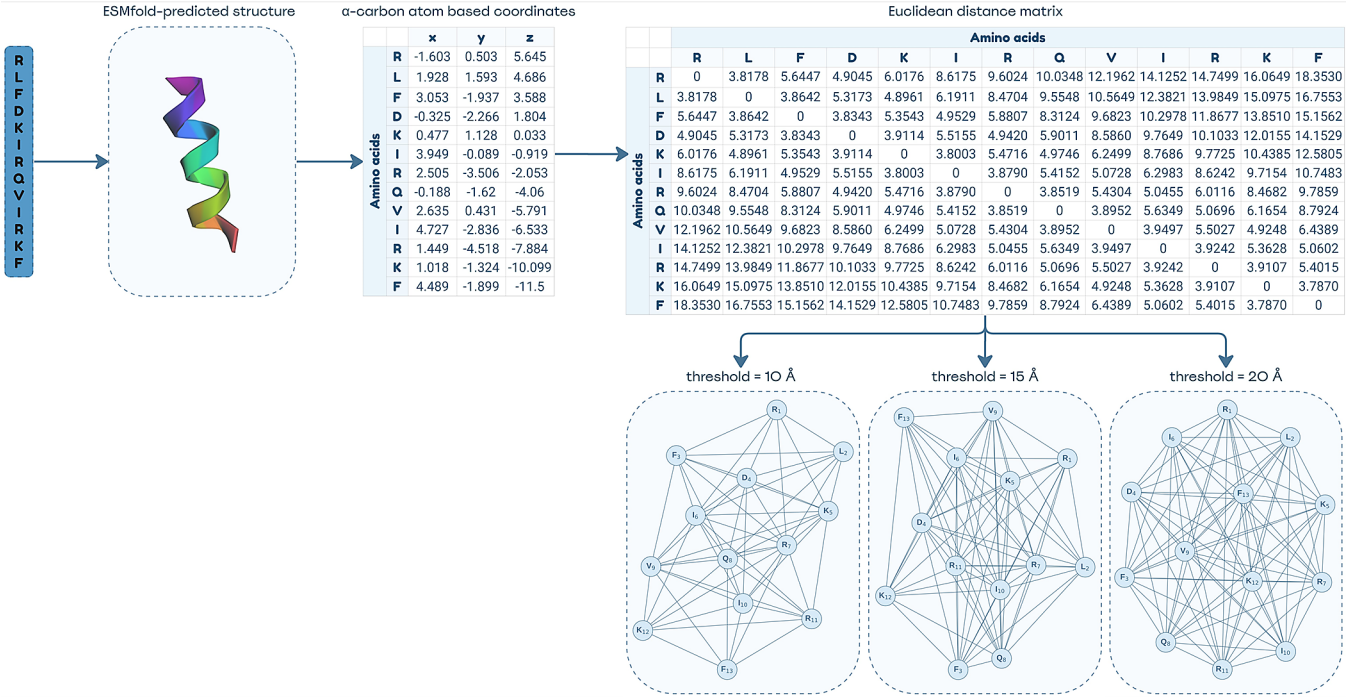


Figure 2. Examples of undirected, unweighted graphs built when applying three distance thresholds.

These characteristics become sAMPpred-GAT an alignment-dependent protocol, making it memory- and time-consuming.[42] Consequently, the prediction of peptide structures and calculation of PSSM-derived evolutionary information should be carried out as a standalone phase before using the sAMPpred-GAT model. This makes a little bit tricky the prospective use of this model for screening large data sets. On the other hand, sAMPpred-GAT is not in correspondence with state-of-the-art (SOTA) advances both to transfer evolutionary information and to predict protein tertiary structures.

Getting PSSM-derived amino acid-level evolutionary information can be considered ancient taking into account SOTA models such as the family of Evolutionary Scale Modeling-2 (ESM-2) models.[34] ESM-2 models, as well as their predecessor ESM-1b model (650 million parameters),[43] were trained on millions of protein sequences available freely in the UniRef database[44] by using a BERT (Bidirectional Encoder Representations from Transformers) architecture. Amino acid-level embeddings containing evolutionary information are obtained from the ESM models,[34,43] which have been successfully applied in predicting mutational effect,[43] secondary structure,[43] and tertiary structure[34] of proteins, as well as in predicting antihypertensive peptides[45] and allergenic proteins and peptides.[46] They also have improved SOTA features, such as the ones obtained from PSSMs, for the long-range residue−residue contact prediction.[43] These embeddings have also been applied in the classification of general AMPs and their antibacterial, antifungal, antiparasitic, and antiviral functional types,[16,17] demonstrating better modeling abilities than handcrafted features. Therefore, ESM model-derived amino acid-level embeddings are a time-saving, better alternative than MSA-based features to transfer evolutionary information when performing downstream tasks.

Moreover, the ESMFold model is a fully end-to-end single-sequence 3D structure predictor based on the 3-billion-parameter, 36-layer ESM-2 model.[34] Unlike trRosetta,[31] AlphaFold,[32] and RoseTTAFold,[33] ESMFold does not depend on MSAs and templates of similar protein structures to achieve a competitive performance. Consequently, ESMFold is computationally more efficient than the other aforementioned approaches. According to the TM-score metric,[47] ESMFold (0.83) achieved comparable performance to AlphaFold (0.88) and RoseTTAFold (0.83) on the CAMEO data set, but it (0.68) was remarkably inferior to AlphaFold (0.85) on the CASP14 data set. These performance gaps between the ESMFold and AlphaFold models are due to the 36-layer ESM-2 model perplexity. For proteins with low perplexity, ESMFold and AlphaFold performed evenly.[34] Nonetheless, despite these performance gaps, ESMFold can be valuable in implementing protocols demanding high-throughput requests, such as the screening of large data sets to discover potential AMPs using graph deep learning (GDL)-based models.

All in all, this manuscript aims to introduce a framework, termed esm-AxP-GDL, to build GDL-based models leveraging ESMFold-predicted peptide structures and amino acid featurization based on the ESM-2 models for the prediction of AMPs. We used a data set comprised of 67,058 peptides (22,461 general-AMPs, 44,597 non-AMPs) to preliminarily evaluate the usefulness of the built framework. We studied the importance of the different amino acid-level ESM-2 embeddings to develop GDL-based models. We compared the built models to non-DL-based models and DL-based models reported in the literature to predict general AMPs. The esm-AxP-GDL framework was designed and implemented to be easily extended to modeling any task related to the prediction of peptide and protein biological activities (or properties).

2. MATERIALS AND METHODS

2.1. Overview of the Framework Esm-AxP-GDL. The esm-AxP-GDL framework (see Figure 1) receives a comma-separated value (CSV) file as input, which contains the identifier, the amino acid sequence, the activity (0 and 1 for

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

negative and positive activities, respectively), and the partition of each peptide. We used the numbers 1, 2, and 3 to represent the training, validation, and test partitions, respectively. The tertiary structure of each sequence in the input file is predicted through the ESMFold model.[34] All the predicted structures are saved in Protein Data Bank (PDB) files. These PDB files can be reused to avoid the ESMFold phase if the same data set is used again for another modeling task. Likewise, PDB files containing 3D structures predicted by other methods[31−33] can be used to leverage the other characteristics of the framework. For each predicted structure, the geometrical distance between the α carbon atoms of every pair of amino acids is calculated using the Euclidean metric. If that distance is less than or equal to a given threshold, then a relationship (edge) is defined between those amino acids. Thereby, a graph-based representation per predicted structure is created, where the nodes represent the amino acids, and the edges represent the structural information. The edges could be weighted with the geometrical distance value. Figure 2 depicts examples of graphs created when applying three different distance thresholds.

The nodes in the built graphs are featured with evolutionary information using the family of ESM-2 models. This family of models is comprised of six pretrained 6-, 12-, 30-, 33-, 36-, and 48-layer models that scale up to 8 million, 35 million, 150 million, 650 million, 3 billion, and 15 billion parameters, respectively. These models were trained on 65 million unique sequences,[34] more than twice the number of unique sequences utilized to pretrain the predecessor ESM-1b model (27.1 million representative protein sequences).[43] The output of the ESM-2 models is a matrix N*x*M per peptide sequence, where *N* is the number of amino acids, and *M* is the embedding size. The larger the capacity of a model, the larger the embedding dimension. Hence, 320-, 480-, 640-, 1280-, 2560, and 5120-dimensional embeddings can be extracted from the 6-, 12-, 30-, 33-, 36- and 48-layer models to characterize each node, respectively. The distance threshold and the ESM-2 model to be used are specified as input parameters. Table SI1 (SI denotes Supporting Information) describes the entire list of framework parameters. The ESM-2 and ESMFold models are available freely at https://github.com/facebookresearch/esm.

Once all the graphs are created, those belonging to the training and validation partitions are selected to train a GDL-based model. If no validation data are given in the input file, then they are extracted from the training data by making a random splitting (80% training, 20% validation). The validation data are used to evaluate each trained model per epoch. We used the GAT architecture (see Section 2.5 in ref 39) implemented in the protocol sAMPpred-GAT because this work is not aimed to introduce a new GDL architecture to classify AMPs and their functional types (or other activities of peptides/proteins), but to present a framework that allows leveraging both the amino acid-level evolutionary information derived from the ESM-2 models and the use of the alignment-free ESMFold model to predict tertiary peptide structures. Therefore, any GDL architecture can be implemented instead of always using the GAT architecture provided. The test data set is used when running the test step only. The training step is performed for a specific number of epochs. For every epoch, the cross entropy-based loss value is calculated on the training and validation sets, respectively. The accuracy (ACC), Matthew correlation coefficient (MCC), area under the curve (AUC), and recall metrics are calculated on the validation and test sets, respectively (see Section SI1). The esm-AxP-GDL framework is freely available at https://github.com/cicese-biocom/esm-AxP-GDL.

2.2. Antimicrobial Peptide Benchmarking Data Set.

We used the AMPDiscover benchmarking set proposed by Pinacho-Castellanos et al.[12] to classify general AMPs. A detailed explanation of how this set was created can be found in Section 2.1 in ref 12. The AMPDiscover set is comprised of 19,548 training, 5125 validation, and 15,685 test peptide sequences (see Table 1 for a description). The positive sequences of the AMPDiscover set were obtained from the starPepDB database.[48,49] To the best of our knowledge, starPepDB is the AMP-related largest repository developed to date, containing a total of 22,642 non-redundant AMP sequences that were compiled from 40 different public databases (see Table 1 in ref 48). The negative sequences for the training and validation sets were created by Pinacho-Castellanos et al. following several criteria applied in the literature.[18,50−52] These authors obtained the negative sequences for the test set from Gabere and Noble.[50] From the test set, Pinacho-Castellanos et al. also created two reduced test sets to ensure fair comparisons regarding several SOTA models. The reduced test sets were built by removing duplicates with training sets of the literature. One of the reduced sets is comprised of sequences of up to 100 amino acids, whereas the other one contains sequences of up to 30 amino acids only. The set comprised of sequences of up to 30 amino acids was used to evaluate the performance of the models in the short-length AMP classification, mainly because they are easier and cheaper to synthesize, modify, and optimize than larger AMPs.

Table 1. Training, Validation, and Test Sets That Are Used in This Work for the General-AMP Classification

data set	total of sequences	positive sequences	negative sequences
training	19,548	9781	9767
validation	5125	2564	2561
test (whole)	15,685	4914	10,771
test (reduced-100)[a]	13,888	3117	10,771
test (reduced-30)[a,b]	7801	2133	5668
external	26,700	5202	21,498

[a]It contains sequences that are not duplicates with the training sets used by models reported in the literature. [b]It contains sequences of up to 30 amino acids.

Additionally, we created an external test set by joining the ABPDiscover,[12] AFPDiscover,[12] AVPDiscover,[12] AniAMPpred,[25] Deep-ABPpred,[23] Deep-AFPpred,[24] and Deep-AVPpred[26] sets. The ABPDiscover, AFPDiscover, and AVPDiscover data sets[12] were created in a similar way to AMPDiscover (see Section 2.1 in ref 12). As for the AniAMPpred data set,[25] their authors obtained the general-AMP sequences from the Protein[53] and starPepDB[48] databases. The antibacterial peptide sequences of the Deep-ABPpred set[23] were collected from the Antimicrobial Peptide Database (APD),[54] the Data Repository of Antimicrobial Peptides (DRAMP),[55] and the Milk Antimicrobial Peptides Database.[56] The antifungal sequences of the Deep-AFPpred set[24] were obtained from the CAMP,[57] DRAMP[55] and StarPepDB[48] databases; whereas the antiviral peptide sequences of the Deep-AVPpred set[26] were obtained from

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

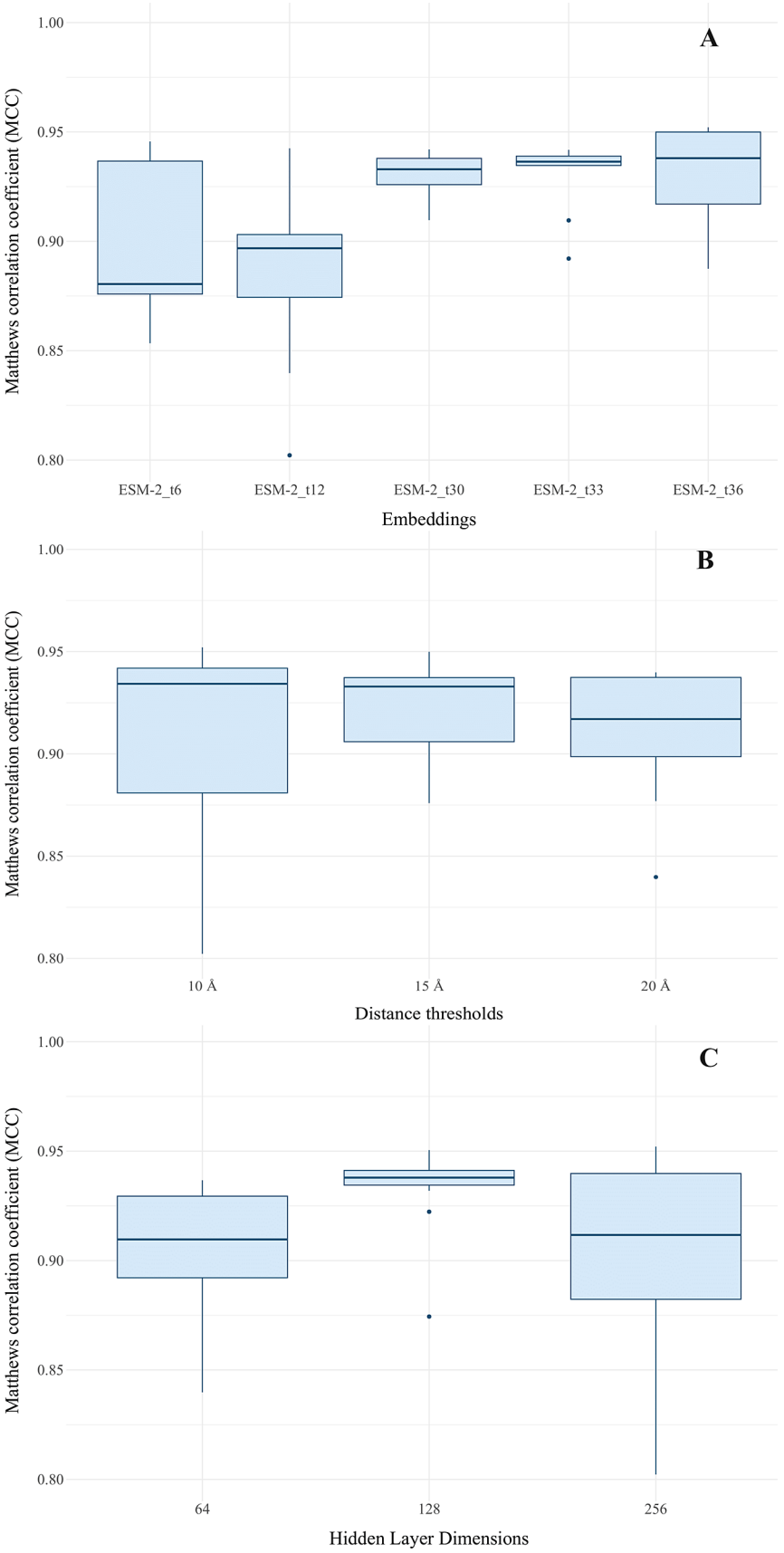


Figure 3. Boxplot graphics corresponding to the MCC_{test} values obtained by the GAT models built when using different (A) ESM-2 embeddings, (B) geometrical distance thresholds, and (C) hidden layer sizes.

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

Table 2. Comparison of the Generalization Ability Achieved on the Whole Test Set (See Table 1) by Models Built on the Training Set Used in This Work

model	SN_{test}	SP_{test}	ACC_{test}	MCC_{test}	AUC_{test}
(A) performance of the best models					
this work	**0.9961**	**0.9707**	**0.9786**	**0.9521**	**0.9942**
AMPDiscover (see Table 2 in ref 12)	0.9110	0.9090	0.9098	0.7990	0.9630
ESM-1b based Random Forest model—see Table 2 in ref 16	0.9040	0.9450	0.9319	0.8430	
AMPScanner (retrained) — see Table S1 in ref 62	0.8757	0.9407	0.9203	0.8151	0.9597
APIN (retrained) — see Table S12.1 in ref 16	0.9153	0.9306	0.9259	0.8317	
APIN-Fusion (retrained) — see Table S13.1 in ref 16	0.9011	0.9542	0.9376	0.8550	
(B) average performances					
this work	**0.9929** (0.005)	**0.9458** (0.026)	**0.9605** (0.017)	**0.9149** (0.033)	**0.9927** (0.001)
AMPScanner (retrained) — see Table S1 in ref 62	0.9207 (0.017)	0.8666 (0.039)	0.8835 (0.022)	0.7544 (0.033)	0.9611 (0.002)
APIN (retrained) — see Table S12.1 in ref 16	0.9301 (0.006)	0.9015 (0.015)	0.9105 (0.008)	0.8045 (0.015)	
APIN-Fusion (retrained) — see Table S13.1 in ref 16	0.9269 (0.013)	0.8932 (0.050)	0.9038 (0.030)	0.7942 (0.49)	
(C) performance of the worst models					
this work	**0.9992**	0.8528	**0.8986**	**0.8022**	**0.9921**
AMPScanner (retrained) — see Table S1 in ref 62	0.9528	0.7731	0.8294	0.6762	0.9624
APIN (retrained) — see Table S12.1 in ref 16	0.9359	**0.8734**	0.8930	0.7733	
APIN-Fusion (retrained) — see Table S13.1 in ref 16	0.9683	0.7225	0.7995	0.6408	

the AVPpred,[58] DBAASP,[59] DRAMP,[55] SATPDB,[60] and starPepDB[48] databases. The negative sequences of these sets were built from reviewed, manually annotated protein sequences available at the UniProt[44] database. Once all these sets were joined, we removed duplicates regarding the AMPDiscover set, getting a total of 26,700 unique sequences. Data SI1 contains the FASTA files of all the sets.

2.3. Methodological Setting To Assess the Usefulness of the Framework Esm-AxP-GDL. We evaluated the usefulness of the esm-AxP-GDL framework by building a total of 45 GAT-based models by using three hidden layer dimensions (i.e., 64, 128, and 256); by applying three distance thresholds (i.e., 10, 15, and 20) to build the graph edges; and by using 5 out of 6 ESM-2 models to characterize each node with evolutionary information. Because of hardware limitations (NVIDIA RTX A5500 24 GB), we were unable to use the 48-layer ESM-2 model because it needs a dedicated GPU memory greater than the one available. The created graphs are unweighted and undirected. Self-loops were considered in the Attention layers. A total of 200 epochs were iterated, the learning rate was set at 0.0001, the dropout value was set at 0.25, and the batch size was set at 512. The ADAM algorithm[61] was the one used for optimization purposes. The models obtained in the last epoch were used to perform the different analyses in this work.

3. RESULTS AND DISCUSSION

3.1. Impact of the ESM-2 Embeddings, Distance Thresholds, and Hidden Layer Dimensions in the Performance of the Built Models. We examined the impact of the ESM-2 embeddings, distance thresholds, and hidden layer dimensions regarding the generalization ability achieved on the whole test set by the created GAT models (see Table SI2). Figure 3A shows boxplot graphics corresponding to the MCC_{test} values obtained when using each ESM-2 model to characterize the nodes of the graphs. It can be noted that the GAT models based on the ESM-2_t36 (2560-dimensional) embeddings produced the highest MCC_{test} values. Specifically, these models achieved the highest (MCC_{test} = 0.9521), the second-highest (MCC_{test} = 0.9505), and the third-highest (MCC_{test} = 0.9499) performances of all the built models. It can also be observed that the distribution of MCC_{test} values corresponding to the ESM-2_t36 embeddings presents the median above the median of the other distributions. This thus indicates that the ESM-2_t36 embeddings contributed to deriving models performing better than most of the models developed with the other ESM-2 embeddings. Note additionally that the distribution of the ESM-2_t33 (1280-dimensional) embeddings presents the smallest interquartile range. This indicates that when they are used, models with good generalization abilities can always be built regardless of the topology of the input graphs and hidden layer dimension accounted for.

Moreover, Figure 3B depicts boxplot graphics corresponding to the MCC_{test} values obtained when applying different distance thresholds to build the graph representations from the predicted peptide structures. The best distribution of MCC_{test} values corresponds to the distance threshold of 15 Å according to the interquartile range, whereas the use of the distance threshold of 10 Å allowed to derive the models with the highest generalization abilities. Indeed, when analyzing the best-10 models according to their MCC_{test} values, 5 out of 10 best models (including the best two ones) were created using a distance threshold of 10 Å; 3 out of 10 best models were built using a distance threshold of 15 Å; and only 2 out of 10 best models (which were the worst ones) were created using a distance threshold of 20 Å. In this regard, notice that the median corresponding to the distribution of 20 Å is inferior to the median of the other two distributions. This implies that using dense graphs will not necessarily lead to yielding models with better generalization abilities than when using sparse graphs. The distance threshold of 20 Å was used in sAMPpred-GAT[39] without examining if it was the most suitable. Thus, according to the results of this analysis, it is important to study what is the most suitable threshold to build graph-based representations from predicted structures that lead to building discriminative models as good as possible.

Finally, Figure 3C depicts boxplot graphics corresponding to the MCC_{test} values achieved by the built models when varying the hidden layer dimension (HLD). As a result, it can be observed that a HLD equal to 128 allowed getting the best models. Nonetheless, the highest MCC_{test} value was yielded by

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

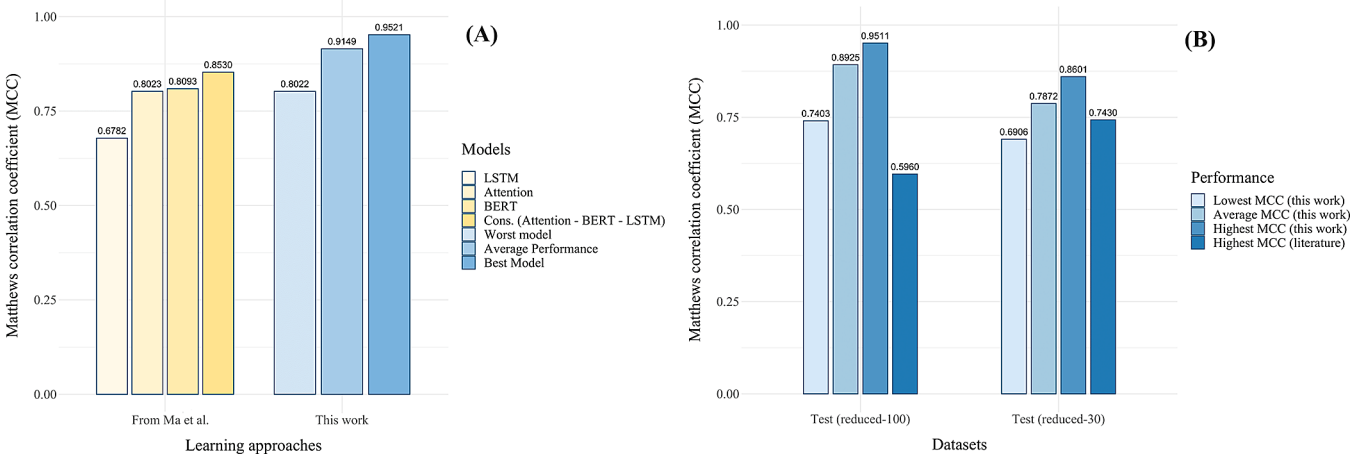


Figure 4. Comparison of the lowest, average, and highest MCC_{test} values achieved by the GAT models created in this work regarding (A) the MCC_{test} values yielded on the whole test set by the deep models proposed by Ma et al.,[64] and regarding (B) the highest MCC_{test} value obtained on the two reduced test sets by 11 models analyzed from the literature.

a model created with an HLD equal to 256. But the small performance difference (ΔMCC_{test} = 0.0016) between the best model and the second-best model (which used an HLD equal to 128) indicates that using an HLD equal to 256 does not contribute to represent better the input data and, consequently, models with notably better generalization abilities will not be built. In this sense, it can be analyzed that 4 out of 5 best models according to their MCC_{test} values were created using an HLD equal to 128. These findings also imply a more efficient training process because the smaller the HLD, the shorter the training time of the models.

3.2. Comparative Analysis Regarding State-of-the-Art Models Built on the Training Set Used in This Work. The built GAT models are compared to models reported in the literature (see Table 2) that were trained and assessed on the benchmarking set used in this work. In this way, fair comparisons are ensured. As for the models of the literature, AMPDiscover was the best model created where the benchmarking set was proposed.[12] The Random Forest model[16] based on the nonhandcrafted ESM-1b features was the best-developed model in a benchmarking study performed to evaluate the complementarity between nonhandcrafted and handcrafted features in the classification of AMPs. In that same study,[16] the AMPScanner,[18] APIN,[19] and APIN-Fusion[19] DL architectures were retrained 30 times, and each of those retrained models was evaluated on the whole test set.

On the one hand, it can be noted in Table 2A that the best GAT model (MCC_{test} = 0.9521) yielded SN_{test}, SP_{test}, ACC_{test}, and MCC_{test} values much better than the ones reported to date. The best model (MCC_{test} = 0.8550) of the literature is based on the APIN-Fusion architecture, which is a multiscale convolutional network fused with amino acid composition and dipeptide composition descriptors. As can be analyzed, according to the MCC_{test} metric, the best model based on the APIN-Fusion architecture was outperformed in absolute terms by the best GAT model in 0.0971. That is, the latter was better than the former by 11.34%. Additionally, it can be analyzed that the sensitivity (SN_{test} = 0.9961) and specificity (SP_{test} = 0.9707) values obtained by the best GAT model were better than the highest ones reported in the literature by 8.83 (SN_{test} = 0.9153) and 1.73% (SP_{test} = 0.9542), respectively.

On the other hand, as for the performance on average (see Table 2B), it can be seen that the 45 GAT models performed better than the models based on the deep architectures at least by 6.75, 4.91, 5.49 and 13.72% with regard to the SN_{test}, SP_{test}, ACC_{test}, and MCC_{test} metrics, respectively. Lastly, it can be observed in Table 2C that the worst GAT model also achieved better metrics than the worst models based on the deep architectures under analysis, except for the SP_{test} metric. In this case, the worst GAT model achieved a SP_{test} value equal to 0.8528, which was only outperformed by the worst model based on the APIN architecture. This APIN-based model achieved an SP_{test} value equal to 0.8734 (2.42% better). The SN_{test}, ACC_{test}, and MCC_{test} metrics obtained by the worst GAT model were better than the ones yielded by the worst models of the literature at least by 2.45, 0.62, and 3.74%, respectively.

Overall, these results suggest that both the use of ESMFold-predicted structures and the use of amino acid-level evolutionary information derived from the ESM-2 models allow to generate valuable graph-based representations, leading to build GDL models with performances consistently good, even better than several SOTA deep architectures (see also Figure 4A). These results seem to contradict the conclusions drawn by Garcia-Jacas et al.[62] in a study between non-DL-based and DL-based models in the classification of AMPs. In that study, several non-DL based models were built on data sets proposed to build DL-based models.[18−23,25−27] It was demonstrated that DL-based models do not outperform non-DL based models and that both types of models perform pretty similar predictions. However, the models analyzed in that study were fed with or learned information derived from amino acid sequences only. Consequently, we think that the outstanding results achieved with the models built in this work are due to the use of input data derived from the 3D information of the predicted structures, which has also been discussed elsewhere.[63]

3.3. Comparative Analysis Regarding State-of-the-Art Models Built on Other Training Sets. Firstly, we performed comparisons with respect to deep models based on the Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and Attention architectures. These models were introduced by Ma et al.[64] to identify general AMPs from the human gut microbiome. The consensus results of these models were also included in the comparisons (i.e., AMP whether the three models classify a

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

Table 3. Generalization Metrics Achieved on the External Test Set by the Best Built Models

models								
ID	HLD	distance threshold	embeddings	$SN_{external}$	$SP_{external}$	$ACC_{external}$	$MCC_{external}$	$AUC_{external}$
model 1	128	10	ESM-2_t6	**0.9485**	0.6226	0.6861	0.4525	0.9099
model 2	128	10	ESM-2_t33	0.9348	0.6842	0.7331	**0.4944**	**0.9346**
model 3	128	10	ESM-2_t36	0.8950	**0.7039**	**0.7412**	0.4819	0.8972
model 4	128	15	ESM-2_t12	0.9321	0.5275	0.6063	0.3669	0.8866
model 5	128	15	ESM-2_t30	0.9404	0.6789	0.7298	0.4939	0.9286
model 6	128	15	ESM-2_t36	0.9312	0.4976	0.5821	0.3448	0.8663

peptide as AMP, non-AMP otherwise). This study was carried out on the whole test set, and the results are shown in Figure 4A (and Table S15). Secondly, we performed comparisons regarding the AmPEP,[13] Deep-AmPEP30,[22] RF-AmPEP30,[22] iAMP-2L,[52] ADAM,[65] MLAMP,[66] AMPfun,[67] CAMPR3-ANN,[57] CAMPR3-RF,[57] CAMPR3-DA,[57] and CAMPR3-SVM[57] models. In this case, we used the reduced test sets. The results are shown in Figure 4B (and Table 5 in ref 12, and Tables SI3 and SI4).

It can first be seen in Figure 4A that the built GAT models yielded comparable-to-superior generalization abilities (MCC metric) to the approaches built by Ma et al.[64] Notice that the worst GAT model was better than the LSTM model by 18.28%, whereas it was inferior to the Attention- and BERT-, Consensus-based models by 0.01, 0.89, and 6.33%, respectively. However, the performance of the best GAT model and the performance on average of all the GAT models were always better than the performances reported by Ma et al. The best-built GAT model outperformed the LSTM-, Attention-, BERT- and Consensus-based models by 40.39, 18.67, 17.64, and 11.62%, respectively, whereas the performance on average of the GAT models was better than the results obtained by the LSTM-, Attention-, BERT-, and Consensus-based models by 34.9, 14.03, 13.05, and 7.26%, respectively. The comparisons between the GAT models and the Ma et al. models were performed without removing duplicates between the Ma et al. training set and the test set used here. Even so, the GAT models were significantly better.

Moreover, similar conclusions can be drawn from the results shown in Figure 4B. As for the reduced set comprised of sequences of up to 100 amino acids, it can be analyzed that the lowest, average, and highest performances achieved in this work were better than the highest performance of the literature by 24.21, 49.75, and 59.58%, respectively. Regarding the reduced test set comprised of short-AMPs, the best performance reported in the literature was obtained by the RF-AmPEP30 model[22] ($MCC_{reduced\text{-}30}$ = 0.743), which is inferior to the average and highest performances yielded in this work by 5.61 and 13.61%, respectively. The lowest performance obtained here to predict short-AMPs was inferior to the RF-AmPEP30 performance by 7.05%. All these results confirm that good graph-based models can be developed leveraging ESMFold-predicted structures and evolutionary information derived from the ESM-2 models. Additional studies are shown below to ascertain how the best GAT models perform in sequence sets other than the AMPDiscover set.

3.4. Validation of the Best Built Models on an External Test Set. From the results explained above, the best three GAT models built with distance thresholds of 10 and 15 Å were selected regarding their MCC_{test} values, respectively. These models were evaluated on the external test set (see Section 2.2), and the results are shown in Table 3. It can first be observed that 4 out of the best 6 models presented a moderated generalization ability since they achieved $MCC_{external}$ values between 0.45 and 0.5. Only two models achieved a poor performance, both presenting $MCC_{external}$ values less than 0.37. The MCC metric considers the true and false positives and negatives. Thus, when diving deeper into the $SN_{external}$ metric, it can be seen that all the models had an outstanding true positive rate. They mainly yielded $SN_{external}$ values above 0.93, except for a model that obtained an $SN_{external}$ value equal to 0.895. These are expected results because the training set was built from the StarPepDB database, which is the largest AMP-related repository to date. Thus, the GAT models learned discriminative features from AMPs representing almost all the universe of such peptides.

However, when analyzing the $SP_{external}$ metric, it can be observed that the models mainly presented a moderate-to-low true negative rate. In this regard, notice that 4 out of the best 6 models were only able to recover between 62 and 70.4% of non-AMPs. The other two models achieved $SP_{external}$ values pretty close to 0.5. These results indicate that the built graph representations were not enough to learn good discriminative features to identify non-AMPs. On the one hand, the latter can be because the negative data set used for training is not representative of the universe of non-AMPs, which has been discussed in ref 68 as an issue to address when building predictive models for this task. On the other hand, the results related to the $SP_{external}$ metric would also suggest improving the graphs representing non-AMPs with the purpose of getting better-learned features. An approach could be the use of topologically different graph representations per peptide-predicted structure to perform multi-instance learning.[69] This approach is because a single graph should not contain all the structural information of a peptide structure, which could adversely affect the construction of models with better true negative rates. Different distance-based criteria and/or physicochemical criteria can be used to build different graphs as has been detailed elsewhere.[70]

Due to the moderate-to-low true negative rates, we analyzed how different the predictions of non-AMPs are between models 2, 3, and 5 shown in Table 3. These models were the best three regarding their $MCC_{external}$ values. To carry out this study, we calculated the disagreement measure[71] for every model pair. This measure is the ratio between the number of predictions on which one model is correct and the other is incorrect regarding the total number of instances. The higher the disagreement, the higher the chance to get better predictions by combining models. As a result, there is a high disagreement (greater than 20%) between models 2 and 3 (0.2605), models 2 and 5 (0.277), and models 3 and 5 (0.2755). Consequently, we first combined models 2 and 5 (highest disagreement) and then we calculated the disagreement metric regarding model 3, getting a value equal to 0.2314

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

which indicates that the three models can be combined together. Table 4 shows the performance metrics when combining the models. Table SI5 shows the consensus predictions for all the negative instances of the external test set.

Table 4. Generalization Metrics Achieved on the External Test Set by the Consensus Models Created by Fusing the Best Three Individual Models

combined models	$SN_{external}$	$SP_{external}$	$ACC_{external}$	$MCC_{external}$
model 2 − model 3	0.8631	0.8244	0.8319	0.5890
model 2 − model 5	0.8933	0.8200	0.8343	0.6062
model 3 − model 5	0.8595	0.8291	0.8351	0.5924
model 2 − model 3 − model 5	0.8339	0.8777	0.8691	0.6417

As expected, the consensus predictions are much better than the predictions yielded by the best individual model between 19.13 and 29.79% regarding the $MCC_{external}$ metric. Notice that the highest $SP_{external}$ value obtained by an individual model ($SP_{external}$ = 0.7039) was significantly improved between 16.49 and 24.69%. Additionally, it can be observed in Table 4 that combining the three individual models improves the performance of combining any pair of them between 5.86 and 8.95% regarding the $MCC_{external}$ metric, which is supported by the disagreement metric-based analysis explained above. Models 2 and 3 were built with graphs derived from a distance threshold of 10 Å and their nodes were characterized with the ESM-2_t33 and ESM-2_t36 embeddings, respectively, whereas model 5 was built with graphs derived from a distance threshold of 15 Å and its nodes were characterized with the ESM-2_t30 embeddings. Thus, the results of combining these models demonstrate that using different ESM-2 embeddings and distance thresholds to get graph-based representations leads to developing models codifying different chemical spaces, which can be successfully used together in prospective screening tasks.

Finally, Figure 5 depicts the performance obtained by both the best three individual models and the consensus model of them in sets comprised of non-AMP pairs whose similarities are below a threshold. We used similarity thresholds ranging between 0.1 and 0.9 with a step equal to 0.1. The sets were created by removing similar non-AMP pairs between the training, validation, test, and external sets. Only the dissimilar non-AMPs belonging to the external set were used in this study. The total number of dissimilar non-AMPs per built data set (see Data SI2 for FASTA files) is also shown in Figure 5. Table SI6 shows the SP values achieved by each model. Overall, it can be seen that all the models achieved a performance as good as the one obtained on the (original) external test set. For the data sets built with similarity thresholds between 0.4 and 0.9, the SP values yielded by the models were comparable to the (original) $SP_{external}$ values (see Tables 3 and 4), whereas, for the data sets built with similarity thresholds between 0.1 and 0.3, the models achieved SP values greater than 0.8. These results suggest that the implemented methodology allows us to build complementary models, that when combined, perform well in sequence sets different from the ones used to train them, which is desirable in prospective applications.

4. CONCLUSIONS

This work aimed to introduce a framework, termed esm-AxP-GDL, to create alignment-free models based on graphs generated from ESMFold-predicted peptide structures, and whose nodes are characterized with amino acid-level evolutionary information derived from the ESM-2 models. We assessed the usefulness of the framework by building several models to predict general AMPs. As a result, it was

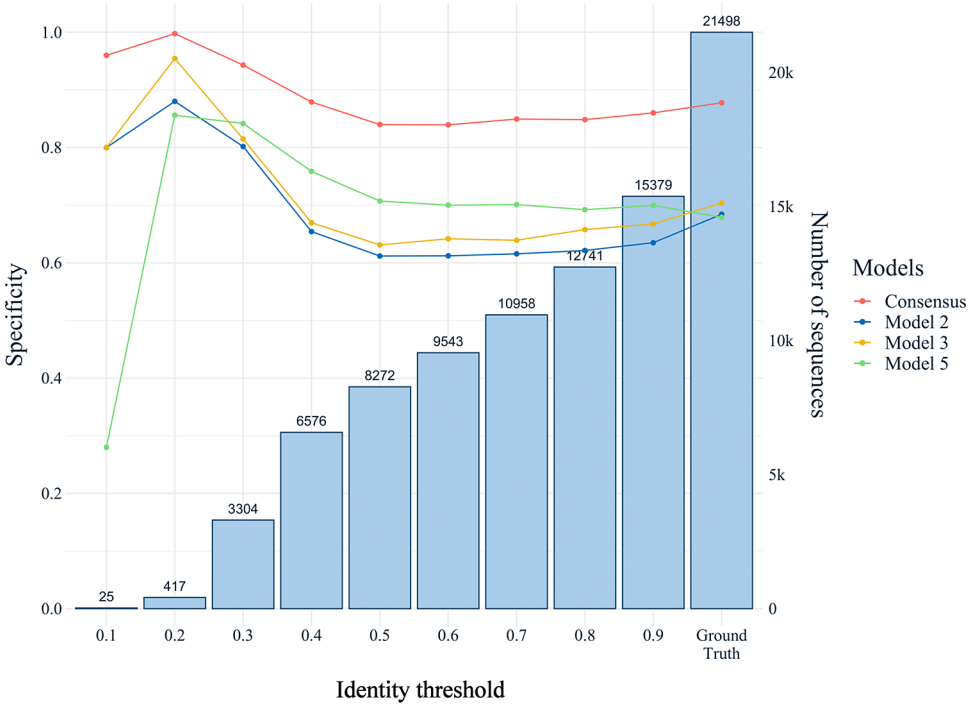


Figure 5. True negative rates obtained by the best three individual models and the consensus model of them in sets containing only non-AMP pairs whose similarities are below a threshold. Ground truth represents the specificity values in the external set.

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

demonstrated that the best performances were achieved when using amino acid-level evolutionary information derived from the ESM-2_t36 embeddings, followed by the ESM-2_t33 and ESM-2_t30 embeddings in that order. It was also demonstrated that all the built models achieved generalization abilities consistently better than 20 state-of-the-art non-DL based and DL-based models that were created from amino acid sequences only. Thus, we conclude that the good results in this work are due to the use of input data derived from the 3D information of the predicted structures. Additionally, it was shown that the built models codify different chemical spaces, and thus they can be combined to significantly improve the classification, mainly the true negative rate. This is particularly useful for prospective screening tasks. In general, the results suggest that the esm-AxP-GDL framework constitutes a prominent tool for building good, structure-dependent, and alignment-free discriminative models that can be utilized to implement screening protocols. This framework should not only be useful to classify AMPs but also for modeling other peptide and protein activities.

5. FUTURE OUTLOOK

We will include the MSA-free HelixFold-Single[35] method to predict 3D peptide structures as another alternative to ESMFold.[34] Additionally, it will be added new strategies to build the graph-based representations, such as the use of contact maps predicted by the ESM-2 models, the use of distance functions other than the Euclidean distance, the use of the criteria implemented in the Graphein library,[70] among others.[72] In this way, it can be analyzed if topologically different graphs could be built to improve the performance of GDL-based models. Finally, it will be implemented the perplexity metric and centrality measures to determine the domain applicability of the models. The domain applicability is to know how trustworthy the predictions of a model are.

■ ASSOCIATED CONTENT

Data Availability Statement

The esm-AxP-GDL framework is available-freely at https://github.com/cicese-biocom/esm-AxP-GDL. The best three built models to classify general-AMPs and the CSV files containing the training, validation, test, and external sets are freely available in the repository specified above into the "best_models" and "datasets" directories, respectively. A small input set is also given into the "example" directory for a quick test of the framework.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c02061.

It contains the links where the FASTA files corresponding to the data sets can be downloaded. Mathematical definition of the performance metrics, list of parameters of the framework, and tables with the performance metrics of all the built models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

César R. García-Jacas − *Cátedras CONAHCYT − Departamento de Ciencias de la Computación, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), 22860 Ensenada, Baja California, México;* orcid.org/0000-0002-3962-7658; Email: cesarrjacas1985@gmail.com

Author

Greneter Cordoves-Delgado − *Departamento de Ciencias de la Computación, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), 22860 Ensenada, Baja California, México*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c02061

Author Contributions

G.C.-D. built the framework esm-AxP-GDL. G.C.-D. and C.R.G.-J. validated the framework. C.R.G.-J. wrote the manuscript and provided guidance and advised to G.C.-D. All authors read, reviewed, and approved the final manuscript.

Funding

CONACYT-funded project 320658 under the program "Ciencia Básica y/o Ciencia de Frontera, Modalidad: Paradigmas y Controversias de la Ciencia 2022".

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

C.R.G.-J. acknowledges to the program "Cátedras CONACYT" from "Consejo Nacional de Ciencia y Tecnología (CONACYT), México" by the support to the endowed chair 501/2018 at "Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE)". C.R.G.-J. also acknowledges the support of CONACYT under grant 320658.

■ REFERENCES

(1) Larkin, H. Increasing Antimicrobial Resistance Poses Global Threat, WHO Says. *JAMA* **2023**, *329*, 200.

(2) WHO Antimicrobial resistance. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance (Oct 10, 2023).

(3) Li, T.; Wang, Z.; Guo, J.; de la Fuente-Nunez, C.; Wang, J.; Han, B.; Tao, H.; Liu, J.; Wang, X. Bacterial resistance to antibacterial agents: Mechanisms, control strategies, and implications for global health. *Sci. Total Environ.* **2023**, *860*, No. 160461.

(4) CDC About Antimicrobial Resistance. https://www.cdc.gov/drugresistance/about.html (Oct 10, 2023).

(5) Hancock, R. E. W.; Haney, E. F.; Gill, E. E. The immunology of host defence peptides: beyond antimicrobial activity. *Nat. Rev. Immunol.* **2016**, *16*, 321−334.

(6) Ahmed, A.; Siman-Tov, G.; Hall, G.; Bhalla, N.; Narayanan, A. Human antimicrobial peptides as therapeutics for viral infections. *Viruses* **2019**, *11*, 704.

(7) 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*, No. e0181748.

(8) Agüero-Chapin, G.; Antunes, A.; Marrero-Ponce, Y. A 2022 Update on Computational Approaches to the Discovery and Design of Antimicrobial Peptides. *Antibiotics* **2023**, *12*, 1011.

(9) Aguilera-Puga, M. D. C.; Cancelarich, N. L.; Marani, M. M.; de la Fuente-Nunez, C.; Plisson, F. Accelerating the Discovery and Design of Antimicrobial Peptides with Artificial Intelligence. In *Computational Drug Discovery and Design*; Gore, M.; Jagtap, U. B., Eds.; Springer: US: New York, NY, 2024; pp 329−352.

(10) Szymczak, P.; Szczurek, E. Artificial intelligence-driven antimicrobial peptide discovery. *Curr. Opin. Struct. Biol.* **2023**, *83*, No. 102733.

(11) Du, Z.; Comer, J.; Li, Y. Bioinformatics approaches to discovering food-derived bioactive peptides: Reviews and perspectives. *Trac-Trends Anal. Chem.* **2023**, *162*, No. 117051.
(12) 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.
(13) Bhadra, P.; Yan, J.; Li, J.; Fong, S.; Siu, S. W. I. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci. Rep.* **2018**, *8*, 1697.
(14) Ruiz-Blanco, Y. B.; Agüero-Chapin, G.; Romero-Molina, S.; Antunes, A.; Olari, L.-R.; Spellerberg, B.; Münch, J.; Sanchez-Garcia, E. ABP-Finder: A Tool to Identify Antibacterial Peptides and the Gram-Staining Type of Targeted Bacteria. *Antibiotics* **2022**, *11*, 1708 DOI: 10.3390/antibiotics11121708.
(15) Meher, P. K.; Sahu, T. K.; Saini, V.; Rao, A. R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **2017**, *7*, 42362.
(16) García-Jacas, C. R.; García-González, L. A.; Martinez-Rios, F.; Tapia-Contreras, I. P.; Brizuela, C. A. Handcrafted versus non-handcrafted (self-supervised) features for the classification of antimicrobial peptides: complementary or redundant? *Brief. Bioinf.* **2022**, *23*, bbac428 DOI: 10.1093/bib/bbac428.
(17) Martínez-Mauricio, K. L.; García-Jacas, C. R.; Cordoves-Delgado, G. Examining evolutionary scale modeling-derived different-dimensional embeddings in the antimicrobial peptide classification through a KNIME workflow. *Protein Sci.* **2024**, *33*, No. e4928.
(18) Veltri, D.; Kamath, U.; Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **2018**, *34*, 2740−2747.
(19) Su, X.; Xu, J.; Yin, Y.; Quan, X.; Zhang, H. Antimicrobial peptide identification using multi-scale convolutional network. *BMC Bioinf.* **2019**, *20*, 730.
(20) Fu, H.; Cao, Z.; Li, M.; Wang, S. ACEP: improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding. *BMC Genomics* **2020**, *21*, 597.
(21) Li, J.; Pu, Y.; Tang, J.; Zou, Q.; Guo, F. DeepAVP: A Dual-Channel Deep Neural Network for Identifying Variable-Length Antiviral Peptides. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3012−3019.
(22) Yan, J.; Bhadra, P.; Li, A.; Sethiya, P.; Qin, L.; Tai, H. K.; Wong, K. H.; Siu, S. W. I. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Mol. Ther.–Nucleic Acids* **2020**, *20*, 882−894.
(23) Sharma, R.; Shrivastava, S.; Kumar Singh, S.; Kumar, A.; Saxena, S.; Kumar Singh, R. Deep-ABPpred: identifying antibacterial peptides in protein sequences using bidirectional LSTM with word2vec. *Briefings Bioinf.* **2021**, *22*, bbab065.
(24) Sharma, R.; Shrivastava, S.; Kumar Singh, S.; Kumar, A.; Saxena, S.; Kumar Singh, R. Deep-AFPpred: identifying novel antifungal peptides using pretrained embeddings from seq2vec with 1DCNN-BiLSTM. *Briefings Bioinf.* **2021**, *23*, bbab422.
(25) Sharma, R.; Shrivastava, S.; Kumar Singh, S.; Kumar, A.; Saxena, S.; Kumar Singh, R. AniAMPpred: artificial intelligence guided discovery of novel antimicrobial peptides in animal kingdom. *Briefings Bioinf.* **2021**, *22*, bbab242.
(26) 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.
(27) Zhang, Y.; Lin, J.; Zhao, L.; Zeng, X.; Liu, X. A Novel Antibacterial Peptide Recognition Algorithm Based on BERT. *Briefings Bioinf.* **2021**, *22*, bbab200 DOI: 10.1093/bib/bbab200.
(28) Du, Z.; Ding, X.; Xu, Y.; Li, Y. UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. *Briefings Bioinf.* **2023**, *24*, bbad135.
(29) 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. *Antibiotics* **2023**, *12*, 747 DOI: 10.3390/antibiotics12040747.
(30) 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.
(31) Du, Z.; Su, H.; Wang, W.; Ye, L.; Wei, H.; Peng, Z.; Anishchenko, I.; Baker, D.; Yang, J. The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **2021**, *16*, 5634−5651.
(32) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(33) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.
(34) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.
(35) Fang, X.; Wang, F.; Liu, L.; He, J.; Lin, D.; Xiang, Y.; Zhu, K.; Zhang, X.; Wu, H.; Li, H.; Song, L. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat. Mach. Intell.* **2023**, *5*, 1087−1096.
(36) Terán, J. E.; Marrero-Ponce, Y.; Contreras-Torres, E.; García-Jacas, C. R.; Vivas-Reyes, R.; Terán, E.; Torres, F. J. Tensor Algebra-based Geometrical (3D) Biomacro-Molecular Descriptors for Protein Research: Theory, Applications and Comparison with other Methods. *Sci. Rep.* **2019**, *9*, 11391.
(37) Marrero-Ponce, Y.; Teran, J. E.; Contreras-Torres, E.; García-Jacas, C. R.; Perez-Castillo, Y.; Cubillan, N.; Peréz-Giménez, F.; Valdés-Martini, J. R. LEGO-based generalized set of two linear algebraic 3D bio-macro-molecular descriptors: Theory and validation by QSARs. *J. Theor. Biol.* **2020**, *485*, No. 110039.
(38) Emonts, J.; Buyel, J. F. An overview of descriptors to capture protein properties − Tools and perspectives in the context of QSAR modeling. *Comp. Struct. Biotechnol. J.* **2023**, *21*, 3234−3247.
(39) Yan, K.; Lv, H.; Guo, Y.; Peng, W.; Liu, B. sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. *Bioinformatics* **2022**, *39*, btac715 DOI: 10.1093/bioinformatics/btac715.
(40) Le, T.; Noe, F.; Clevert, D.-A., In *Proceedings of the First Learning on Graphs Conference*; Bastian, R.; Razvan, P., Eds.; PMLR: Proceedings of Machine Learning Research, 2022; Vol. *198*, pp 30:1−30:17.
(41) Mohammadi, A.; Zahiri, J.; Mohammadi, S.; Khodarahmi, M.; Arab, S. S. PSSMCOOL: a comprehensive R package for generating evolutionary-based descriptors of protein sequences from PSSM profiles. *Biol. Methods Protoc.* **2022**, *7*, bpac008.

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321

(42) Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* **2017**, *18*, 186.
(43) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2016239118.
(44) UniPort Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506−D515.
(45) Du, Z.; Ding, X.; Hsu, W.; Munir, A.; Xu, Y.; Li, Y. pLM4ACE: A protein language model based predictor for antihypertensive peptide screening. *Food Chem.* **2024**, *431*, No. 137162.
(46) Du, Z.; Xu, Y.; Liu, C.; Li, Y. pLM4Alg: Protein Language Model-Based Predictors for Allergenic Proteins and Peptides. *J. Agric. Food Chem.* **2024**, *72*, 752−760.
(47) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889−895.
(48) 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.
(49) 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 DOI: 10.1093/bioinformatics/btad506.
(50) Gabere, M. N.; Noble, W. S. Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics* **2017**, *33*, 1921−1929.
(51) Torrent, M.; Andreu, D.; Nogués, V. M.; Boix, E. Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model. *PLoS One* **2011**, *6*, No. e16968.
(52) 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. *Anal. Biochem.* **2013**, *436*, 168−177.
(53) Schoch, C. L.; Ciufo, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; Sharma, S.; Soussov, V.; Sullivan, J. P.; Sun, L.; Turner, S.; Karsch-Mizrachi, I. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, *2020*, baaa062 DOI: 10.1093/database/baaa062.
(54) Wang, G.; Li, X.; Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **2016**, *44*, D1087−D1093.
(55) Kang, X.; Dong, F.; Shi, C.; Liu, S.; Sun, J.; Chen, J.; Li, H.; Xu, H.; Lao, X.; Zheng, H. DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci. Data* **2019**, *6*, 148.
(56) Théolier, J.; Fliss, I.; Jean, J.; Hammami, R. MilkAMP: a comprehensive database of antimicrobial peptides of dairy origin. *Dairy Sci. Technol.* **2014**, *94*, 181−193.
(57) Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.* **2016**, *44*, D1094−D1097.
(58) Thakur, N.; Qureshi, A.; Kumar, M. AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res.* **2012**, *40*, W199−W204.
(59) Pirtskhalava, M.; Gabrielian, A.; Cruz, P.; Griggs, H. L.; Squires, R. B.; Hurt, D. E.; Grigolava, M.; Chubinidze, M.; Gogoladze, G.; Vishnepolsky, B.; Alekseev, V.; Rosenthal, A.; Tartakovsky, M. DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.* **2016**, *44*, D1104−D1112.
(60) Singh, S.; Chaudhary, K.; Dhanda, S. K.; Bhalla, S.; Usmani, S. S.; Gautam, A.; Tuknait, A.; Agrawal, P.; Mathur, D.; Raghava, G. P. S. SATPdb: a database of structurally annotated therapeutic peptides. *Nucleic Acids Res.* **2016**, *44*, D1119−D1126.
(61) Kingma, D. P.; Ba, J., Adam: A Method for Stochastic Optimization. *CoRR* **2014**, abs/1412.6980.
(62) Garcia-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 Bioinf.* **2022**, *23*, bbac094 DOI: 10.1093/bib/bbac094.
(63) Durairaj, J.; de Ridder, D.; van Dijk, A. D. J. Beyond sequence: Structure-based machine learning. *Comp. Struct. Biotechnol. J.* **2023**, *21*, 630−643.
(64) Ma, Y.; Guo, Z.; Xia, B.; Zhang, Y.; Liu, X.; Yu, Y.; Tang, N.; Tong, X.; Wang, M.; Ye, X.; Feng, J.; Chen, Y.; Wang, J. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* **2022**, *40*, 921−931.
(65) Lee, H.-T.; Lee, C.-C.; Yang, J.-R.; Lai, J. Z. C.; Chang, K. Y. A Large-Scale Structural Classification of Antimicrobial Peptides. *BioMed Res. Int.* **2015**, *2015*, No. 475062.
(66) Lin, W.; Xu, D. Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types. *Bioinformatics* **2016**, *32*, 3745−3752.
(67) 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 Bioinf.* **2020**, *21*, 1098−1114.
(68) 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. *Briefings Bioinf.* **2022**, *23*, bbac343 DOI: 10.1093/bib/bbac343.
(69) Zankov, D.; Madzhidov, T.; Varnek, A.; Polishchuk, P. Chemical complexity challenge: Is multi-instance machine learning a solution? *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2024**, *14*, No. e1698.
(70) Jamasb, A.; Viñas Torné, R.; Ma, E.; Du, Y.; Harris, C.; Huang, K.; Hall, D.; Lió, P.; Blundell, T.Graphein-a python library for geometric deep learning and network analysis on biomolecular structures and interaction networks. In *Advances in Neural Information Processing Systems*; 2022; Vol. *35*, pp 27153−27167.
(71) Kuncheva, L. I.; Whitaker, C. J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* **2003**, *51*, 181−207.
(72) Zheng, L.; Shi, S.; Lu, M.; Fang, P.; Pan, Z.; Zhang, H.; Zhou, Z.; Zhang, H.; Mou, M.; Huang, S.; Tao, L.; Xia, W.; Li, H.; Zeng, Z.; Zhang, S.; Chen, Y.; Li, Z.; Zhu, F. AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. *Genome Biol.* **2024**, *25*, 41.

https://doi.org/10.1021/acs.jcim.3c02061
J. Chem. Inf. Model. 2024, 64, 4310−4321