AutoPeptideML: Automated Machine Learning for Building Trustworthy Peptide Bioactivity Predictors

Raúl Fernández-Díaz[1,2,3,*], **Rodrigo Cossio-Pérez**[1,2,4], **Clement Agoni**[1,2], **Hoang Thanh Lam**[3], **Vanessa Lopez**[3], **Denis C. Shields**[1,2,*]
*raul.fernandezdiaz@ucdconnect.ie, *denis.shields@ucd.ie

[1] School of Medicine. University College Dublin
[2] Conway Institute of Biomolecular and Biomedical Science. University College Dublin
[3] IBM Research, Dublin
[4] Department of Science and Technology. Quilmes National University

Abstract

Automated machine learning (AutoML) solutions can bridge the gap between new computational advances and their real-world applications by enabling experimental scientists to build trustworthy models. Here, we consider the design of such a tool for developing peptide bioactivity predictors. We analyse different design choices concerning data acquisition and negative class definition, homology partitioning for the construction of independent evaluation sets, the use of protein language models as a general sequence representation method, and model selection and hyperparameter optimisation. Finally, we integrate the conclusions drawn from this study into AutoPeptideML, an end-to-end, user-friendly application that enables experimental researchers to build trustworthy models, facilitating compliance with community guidelines.

The source code, documentation, and data are available in the project GitHub repository: `https://github.com/IBM/AutoPeptideML`. Additionally, we have established a dedicated web-server, accessible at:`http://peptide.ucd.ie/AutoPeptideML`.

Keywords Bioactive peptides · Machine Learning · AutoML · Protein Language Models · Trustworthy AI

1 Introduction

Recent machine learning (ML) and deep learning (DL) algorithms have had tremendous impact on biochemistry research, particularly in the areas of drug discovery [1] and structural bioinformatics [2]. However, there are still technical challenges that need to be addressed before the full potential of these advances can be realised. They include working with small dataset sizes [3], sparsely labelled data [4], minimising sequence homology between training and evaluation sets to properly test the model generalisation capabilities [5, 6, 7], and a lack of general-purpose sequence representation methods [8].

One task that epitomises most of the key issues present in the prediction of biological function from sequence is peptide bioactivity prediction. Most available datasets are small and lacking in explicit negative labels, there is no standard representation method for peptide sequences, nor is there any standardised practice regarding control of sequence homology between training and evaluation sets [4, 3, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. Additionally, there is significant demand for custom ML predictors as there are many different bioactivities relevant for diverse applications ranging from drug discovery [19] to understanding metabolic or inflammatory signalling pathways [20, 21]. Currently, the development of models for new bioactivities is a tricky process that involves both ML expertise and domain knowledge [3].

Automated machine learning (AutoML) seeks to address these challenges by incorporating strategies that enable users with little or no ML knowledge to build high-quality custom predictors [22, 23]. Such a system would be able to assist researchers in designing more efficient experimental workflows informed by robust and reliable predictions [24]

accelerating scientific discovery. A key feature of such automation is that it should integrate best practices and follow community guidelines in ML implementation [5] leading to more trustworthy models [25] that avoid the typical pitfalls of model over-prediction and lack of generalisability and produce reproducible and interpretable outputs.

Data acquisition and curation. The first step in the development of any ML model is obtaining and curating the data. The main challenge regarding data acquisition for peptide bioactivity prediction is the lack of explicit negative samples. There is a variety of strategies proposed for addressing this problem that include drawing random fragments from proteins in databases like UniProt or SwissProt [4, 16, 9], drawing actual random peptides from the same databases [18, 26, 12, 4], and using peptides with a specific bioactivity as the negative class [4, 10, 17]. In all cases, there are important tradeoffs regarding the generalizability of the model. In the first and second approaches, the differential features between positive and negative classes could be explained by a myriad of confounding factors that do not have a direct bearing on their bioactivity. In the third approach, the opposite is true, positive and negative classes are so similar to each other that the model hyper-focuses on the specific differential features between both bioactivities, which hinders its generalizability. If the evaluation set has been constructed following any of these strategies, it will inherit the same biases and thus will not be able to properly diagnose them. Here, we explore an intermediate solution: to draw the negative peptides from a database with multiple bioactivities, instead of a single bioactivity. This approach generates a distribution of negative peptides that is broad enough (by covering several distinct bioactivities) as to generalise adequately, but that is similar enough to the positive peptide distribution (by also being bioactive peptides) as to minimise confounding factors.

Sequence representation. The use of domain knowledge for building effective peptide representations can simplify the process of training ML models and bypass some of the limitations inherent to small sample sizes [27]. There exists a plethora of different representations that can be used based on peptide composition [17, 9, 11, 13, 28], evolutionary profile [29], and physico-chemical profile [18, 9, 13, 11]. The emergence of sequence language models for either DNA [30] or protein sequences [31, 32, 8, 33] has the potential of streamlining sequence representation beyond those methods. The sequence language models are DL algorithms trained to predict masked nucleotides or residues in a sequence allowing the models to capture the conditional probability that any given nucleotide or residue will occupy a certain position given the rest of the sequence. Consequently, they acquire the ability to provide high-dimensional vector representations of the sequences containing implicit structural, functional, and evolutionary properties. It is important to note that the data used during this process relies on sequence databases like UniRef [34] which does not contain explicit information about their properties, rather just nucleotide or residue distributions [30, 31, 32, 8, 33]. Although peptide sequences are less constrained than the bulk of their training data, recent studies [35, 36] have demonstrated that protein language model (PLM) representations obtained from models from the ESM (Evolutionary Scale Modelling) family [37, 31], can be used to train ML algorithms that predict peptide bioactivity with remarkable success, which led us to select them as a general representation method.

Independence of evaluation dataset. The independence between training and evaluation datasets is essential to properly evaluate any ML predictor [5, 6, 7]. Achieving this independence, however, is not straight-forward when working with biological sequences, because similar sequences tend to share structural and functional features [38]. The standard practice in protein ML development is to perform homology reduction which is a technique that clusters the sequences using algorithms like CD-HIT [39], HHblits [40] or MMseqs [41] to then preserve only the representative sequences from each cluster which is particularly useful when working with large data sets. However, when working with small datasets, the resulting loss of information hinders the models substantially [15]. Homology partitioning is an alternative approach that removes the minimal number of sequences to ensure that each cluster is independent from the others [42, 43]. Here, we propose an alternative partitioning algorithm that generalises well to short sequences and maximises the difference between training and evaluation sets.

We have combined these three approaches above (bioactive sampling of negatives, PLM representation, and homology partitioning) into AutoPeptideML, a user-friendly AutoML application that allows researchers without a strong computational background to build their own peptide bioactivity predictors easily, while supporting their reproducibility and robust evaluation. The tool is made available as an open source project under MIT licence, with three different interfaces for users with different levels of technical expertise: 1) an application programming interface (API) meant for developers to integrate the whole workflow or specific parts into their own custom pipelines; 2) a stand-alone command-line interface (CLI), for users with technical background to exert greater control over the process; and, 3) a graphical user interface (GUI) hosted in a web-server, for users without a computational background. We demonstrate that the application of the method has good performance against a series of benchmark peptide bioactivity prediction datasets, evaluating not only performance, but also the effect of different design choices on performance which can be particularly marked in smaller datasets. Thus, the software not only provides a facility for biologists to make predictions from peptide datasets, but also encourages compliance with, and appreciation of, best practices in machine learning.

2 Methods

2.1 Algorithm

The primary objective behind the design of AutoPeptideML is to provide a user-friendly tool that does not require extensive technical knowledge to use, while still remaining highly versatile. This is achieved through a pipeline that guarantees compliance with community guidelines such as DOME (Data, Optimisation, Model, and Evaluation), ensuring a robust scientific validation [5] (see Supplementary). This pipeline automates six steps: 1) dataset curation, 2) search for negative peptides, 3) dataset partitioning, 4) PLM peptide representation, 5) hyperparameter optimisation, and 6) model evaluation. The tool also allows for a final prediction step where any model previously developed with the tool can be used to predict the bioactivity of a new dataset of peptides of interest. A visual summary of this workflow can be found in Figure 1.

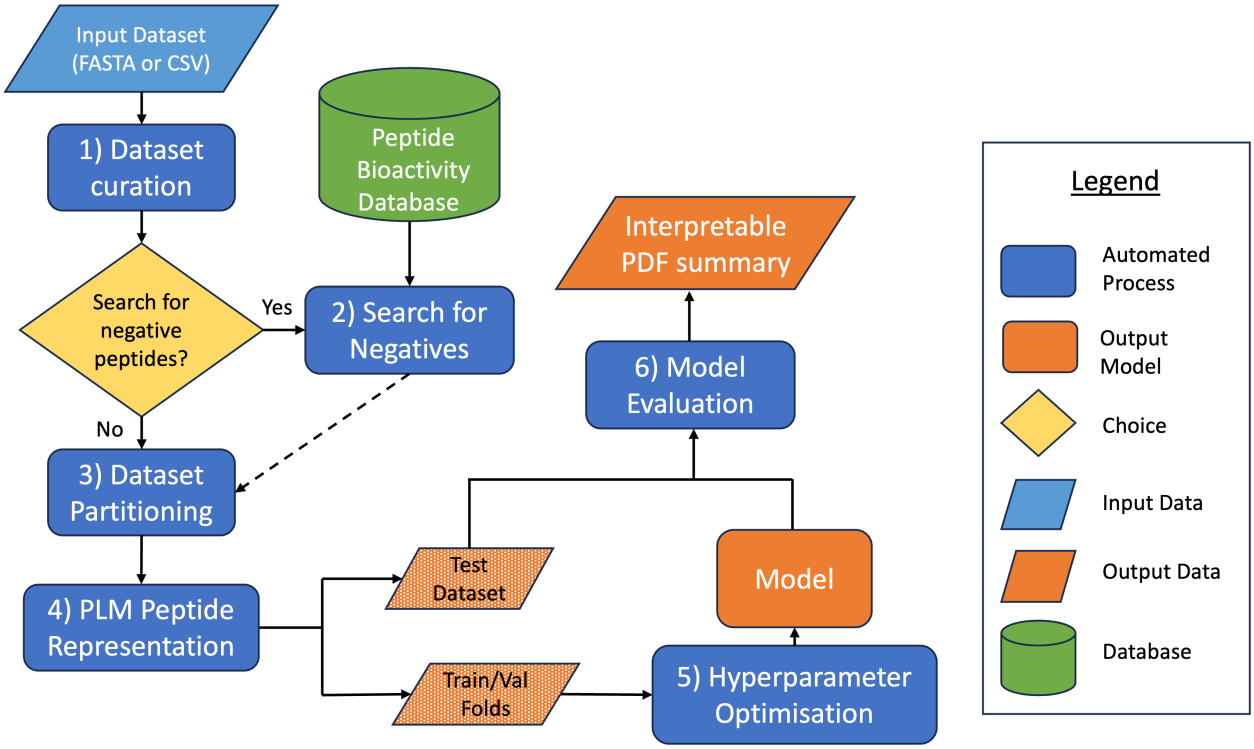

Figure 1: Visual summary of the AutoPeptideML workflow.

Inputs. AutoPeptideML only requires a dataset of peptides known to be positive for the bioactivity of interest.

Outputs. AutoPeptideML generates a directory containing the generated model, along with all the necessary information to understand its capabilities and reproduce its training. The output contains: 1) the experimental configuration used to run AutoPeptideML, 2) all the intermediary datasets generated by the program, 3) the optimal configurations found during hyperparameter optimisation, 4) the evaluation metrics for both the n-fold cross-validation and against the holdout independent evaluation set, including a set of visualisations to facilitate interpreting the results, 5) the predictive model itself, and 6) an interpretable summary including all information regarding figures and main metrics with accompanying explanations of their meaning to guide researchers not specialised in ML on how reliable the model generated is.

Step 1 - Dataset curation. Dataset curation preprocesses the data to remove all sequences that are either repeated or contain non-standard amino acids. If both positive and negative peptides are provided, the program balances the classes by oversampling the underrepresented class and continues to Step 3; if negative peptides are not provided, it executes Step 2 to build the negative set.

Step 2 - Automatic search for negative peptides. The automatic search for negative peptides queries a database of bioactive peptides. To avoid introducing false negative peptides into the negative subset, the algorithm accepts an optional input containing a list of bioactivity tags that the user considers may overlap with the bioactivity of interest

and should, therefore, be excluded. To ensure that the negative peptides are drawn from a similar distribution to the positive peptides and minimise the number of confounding factors, the system calculates the lengths of all peptides in the positive peptides and organises the lengths into a histogram with bin size of 5. Then, for each bin in the histogram, it queries the reference database for as many peptides as present in the bin, with lengths between its lower and upper bounds. If there are not enough peptides in the database, the remaining peptides are drawn from the next bin.

Regarding the reference database, the system uses a pre-prepared subset of the Peptipedia database [44], named AutoPeptideML-Peptipedia. The original Peptipedia database integrates information from 30 peptide bioactivity databases collecting almost 97,331 bioactive peptides labelled with 128 bioactivities (version `29_03_2023`). AutoPeptideML-Peptipedia is the result of removing all sequences with non-standard residues or without any known bioactivity and contains 92,092 peptides (see Supplementary).

Step 3 - Dataset Partitioning. Dataset partitioning divides a dataset into an overall training dataset, validation datasets drawn from within the training dataset using tenfold cross-validation, and an evaluation (test) set. The evaluation set is only evaluated once in the process, to assess the performance of the final choice of model [5]. By default, the proportion of training:evaluation data is 80:20. The division is performed in two subsequent steps.

The first division relies on the new homology partitioning algorithm, which will now be formally introduced. Let $G(S, E)$ be a graph where the nodes are the set of all sequences S in the dataset and the edges the subset of all pairwise alignments between those sequences with a similarity sim above a threshold λ, $E = \{(s_1, s_2)\ \forall\, s_1, s_2 \in S$ such that $sim(s_1, s_2) > \lambda\}$. Let G_{train} and G_{test} be two subgraphs of G and let $E_f = \{(s_{train}, s_{test}) \in E$ such that $s_{train} \in G_{train}$ and $s_{test} \in G_{test}\}$. The objective is to define G_{train} and G_{test} such that $E_f = \emptyset$.

GraphPart [43] is a heuristic approach that first reduces the number of edges so that for each node only the edge with maximum distance is considered, then iteratively assigns sequences to different subgraphs until finding a combination that minimises E_f, subsequently, checks whether there are nodes with other non-maximal nodes to other subgraphs, and removes the minimum possible number of sequences so that $E_f = \emptyset$. Here we propose an alternative method. It first identifies the set of all unconnected subgraphs of G, $\{U_1, U_2, \ldots, U_k\}$. Then, it iteratively assigns the smallest subgraph to G_{test} if the resulting number of samples from positive and negative labels remains below the expected values for G_{test}, until the desired size is reached. The underlying idea is that the subgraphs with fewer neighbours are more dissimilar to the main bulk of sequences in the dataset. Therefore, the evaluation set built with the most dissimilar sequences is not only independent from the train set, but it is a more close approximation to an out-of-distribution evaluation. Furthermore, by exploiting the topology of the graph rather than manipulating the number of nodes, the method does not require the removal of any sequences, minimising the loss of information.

There are three alternative methods for calculating the pairwise peptide sequence alignements: MMSeqs2 align module [45], MMSeqs2 prefilter and align module [45], and Needleman-Wunch [46, 47]. The first two options perform local alignments relying ultimately on the Smith-Waterman algorithm [48]. The difference between them is whether a prefiltering step analysing k-mer composition is performed accelerating the alignment process. The Needleman-Wunch algorithm, on the other hand, is a global alignment. Here, the default algorithm recommended is MMSeqs2 with prefiltering as it is both greatly computationally efficient and local alignments are more accurate at identifying homologous regions when the conserved regions are asymmetrical from the center of the sequence [49], which is a likely scenarios when working with peptide sequences of varying sizes. Sequence identity between a sequence pair may be estimated for the longest sequence, the shortest sequence or the alignment length. The default is to use the longest sequence, since the shortest may lead to many random matches of very short peptides to larger peptides. For longer protein sequences, however, it would be recommended to use the shortest sequence as it could capture the relationships between larger functional subdomains [50, 51].

The second division concerns the training set which will be subdivided into n cross-validation folds. Two alternatives are implemented: random stratified n-fold division (as implemented in `scikit-learn` [52]) or the implementation of stratified n-fold homology partitioning in `GraphPart` [43]. Independence between training and validation folds is not as essential as between training and evaluation sets as the main objective of the validation set is to ensure that the algorithm is learning from the training examples. Therefore, it is reasonable for the validation set to belong to the same distribution as the training set and, consequently, the default option is the stratified tenfold random division.

Step 4 - Peptide Representation. The ESM-2 model with 8 million parameters (ESM2-8M) ([31] is used to compute a representation for each residue in the peptide, using `Transformers` and `PyTorch`. The resulting peptide representation is a matrix with dimension $(n, 320)$ where n is the length of the peptide and 320 the representation size which is determined by the PLM architecture. A global representation for the whole sequence is then calculated by averaging across the individual residue representations for a final representation with dimension $(1, 320)$ [53, 54, 55].

The selection of this model is the result of an evaluation of the effect of model size on downstream performance (see Subsection 3.3). Nevertheless, users have the option to input their own peptide representations as a JSON file

that may contain handcrafted features, learnt representations obtained from a different PLM or other type of protein representation learning algorithm [55, 56, 57], or a combination thereof.

Step 5 - Hyperparameter Optimisation and Model Training. Model selection and hyperparameter optimisation (HPO) are tedious and time-consuming tasks that require extensive trial-and-error [35, 3], and if not performed appropriately, can lead to model overfitting and lack of generalisability. AutoPeptideML automates and streamlines model development, by performing a greedy Bayesian Optimisation Search, as implemented by Optuna [58]. Users of either the CLI or the API are free to define the number of models that should be included in this optimization, as well as their hyperparameter search space. AutoPeptideML supports the following algorithms: K-nearest neighbours (KNN), light gradient boosting (LightGBM), support vector machine (SVM), random forest classification (RFC), extreme gradient boosting (XGBoost), simple neural networks like the multi-layer perceptron (MLP), and 1D-convolutional neural networks (1D-CNN). Model selection and HPO are conducted simultaneously in a cross-validation regime so that the metric to optimise is the average across n folds. Thus, the system is never exposed to the evaluation set, which is kept unseen until the final model evaluation [5].

AutoPeptideML builds a soft-voting equally-weighted ensemble [59] of the n iterations of the best model or models obtained in the HPO. This allows for performing HPO individually for an arbitrary number, m, of ML algorithms, so that the final ensemble combines $n \times m$ individual models. The default configuration (see Supplementary) performs a short HPO for three simple ML algorithms, namely, K-Nearest Neighbours (KNN), Light Gradient Boosting (LightGBM), and Random Forest Classification (RFC).

Step 6 - Model Evaluation. The ensemble obtained in the previous step is evaluated against the hold-out evaluation set using a wide range of metrics that include accuracy, balanced accuracy, weighted precision, precision, F1, weighted F1, recall, weighted recall, area under the receiver-operating characteristic curve, matthew's correlation coefficient (MCC), jaccard similarity, and weighted jaccard similarity as implemented in the `scikit-learn` [52]. The plots generated include calibration curve, confusion matrix, precision-recall curve, and receiver-operating characteristic curve, as implemented in `sckit-plot` [60].

Step 7 - Prediction. AutoPeptideML can predict the bioactivity of new samples given a pre-trained model generated in Step 5. Predictions are a score within the range $[0, 1]$. This result can be interpreted as the probability of the peptide sequence having the target bioactivity, given the predictive model $(P(x \cap +|\text{MODEL}))$. This step outputs a CSV file with the peptides sorted according to their predicted bioactivity probability.

2.2 Benchmark datasets and model implementation

Benchmarks. 16 different peptide bioactivity datasets containing positive and negative samples were used to evaluate the various modules of the algorithm. These datasets were selected from a previous study, evaluating the use of PLMs for general peptide bioactivity prediction [35]. The datasets ranged in size from 200 to 20,000. In order to evaluate homology partitioning (HP) and automated selection of negatives, two new benchmark sets were constructed from these original datasets named "Autoneg" and "Autoneg+HP". In both datasets positive samples were taken from the original bioactivity sets (see Table 1), and the negative peptides were automatically selected from other bioactivity classes, as described in Step 2. In the "Autoneg" sets, training and evaluation subsets were divided by random splitting, whereas in "Autoneg+HP", they were divided using the new homology partitioning algorithm described in Step 3. Therefore, "Autoneg" and "Autoneg+HP" only differ on the partitioning method employed and were composed of the same sets of peptides. In the original benchmarking datasets (see Table 1), the two anticancer and the two antimalarial sets respectively shared the same positive peptides and only differed on how the negative classes were defined. Therefore 16 "Autoneg" and "Autoneg+HP" datasets were constructed, and compared to 18 models in the original benchmarking dataset.

AutoPeptideML configuration. Dataset partitioning used MMSeqs2 [45] with prefilter for calculating all pairwise sequence alignments in the train/evaluation division and random partitioning for creating 10 cross-validation folds. The maximum sequence similarity threshold allowed is 30%. Automatic Hyperparameter Optimisation was performed separately for KNN, LightGBM and RFC and the final ensemble contained 10 instances (one per cross-validation fold) of each of these models for a total of 30 models. Final model predictions were the average of all 30 individual predictions.

Statistical Analysis. All statistical analyses were performed using the corresponding functions within `Pandas` [61], as detailed in the accompanying code.

UniDL4BioPep-A. We wished to compare our findings with a state-of-the-art framework, UniDL4BioPep [35], which trained a convolutional neural network on ESM2-8M representations for peptide bioactivity prediction. Analysis of the source code in the UniDL4BioPep [35] project GitHub Repository [`https://github.com/dzjxzyd/UniDL4BioPep`]

indicated that the hold-out test set is used during model training to determine the best checkpoint for evaluation. However, DOME guidelines recommend avoiding this practice [5], as it compromises the independence of the hold-out evaluation dataset, which can lead to the overestimation of model performance. Accordingly, an amended version of the UniDL4BioPep model, which we named UniDL4BioPep-A, was used. This version instead used the 10-fold cross-validation folds to determine the best possible checkpoint, maintaining the hold-out set unseen until final model evaluation. This implementation is written in PyTorch and was adapted from code provided by the UniDL4BioPep authors in a secondary project Github Repository `https://github.com/David-Dingle/UniDL4BioPep_ASL_PyTorch`.

2.3 Analysis of sequence homology between train and evaluation sets

The dependence between training and evaluation sets in both the original datasets and the new ones generated throughout the study was evaluated by first calculating pairwise alignments between the sequences in either set using MMSeqs2 (Version: `14-7e284`) [41, 45] with both prefilter and align modules. The prefiltering step is conducted with high sensitivity (`-s 7.5`). The align module uses the following parameters: `-alignment-mode 3 -e inf -seq-id-mode 2` (corresponding to using the length of the longest sequence as the denominator for the sequence identity calculation). Then, the proportion of the proteins within the training dataset with at least one homologous protein in the evaluation set is calculated. Here, we considered that two proteins are homologous if they have a sequence identity above 30%.

3 Results and discussion

AutoPeptideML is the first end-to-end AutoML solution proposed for building peptide bioactivity ML predictors. It combines three features: automatic selection of negatives from a background dataset of bioactive peptides to increase specificity, homology partitioning of datasets to evaluate generalisability, and implementation of a protein language model representation to streamline representation. One property of datasets of bioactive peptides is that they are often quite small. Thus, computational methods that embed knowledge from outside the training dataset, that are relevant to training, can serve to strongly streamline the efficiency of model estimation, reducing stochastic elements of parameter choice. We leveraged the ESM-8M PLM [31] in AutoPeptideML to generate a representation of each peptide as a vector of length 320, which is used as the set of features for a downstream model.

3.1 Analysis of new benchmarks

Choice of benchmarks. Benchmark datasets of bioactive peptides along with their negative comparison datasets that had previously been used in training models were taken from a collection assembled by [35]. These datasets have been used to benchmark handcrafted ML models specific to each of those peptide bioactivities, allowing us to compare the features and performance of the datasets and models to an automated approach. They represent a wide variety of bioactivity prediction tasks. Table 1 describes the source of each benchmark dataset, the negative class definition, the strategy used for building the evaluation set, and the reference of the best specific model reported on that dataset.

Choice of negative peptides. Choosing negative peptides for ML is frequently difficult as typically the literature is heavily biassed towards reporting positive peptide bioactivities and there are much smaller datasets of negatives. Selecting random peptide sequences (either made up, or drawn from fragments of large protein databases) may confound the specific bioactivity with other more general features (including solubility, extracellular milieu and other factors), thus creating a predictor that is not informative regarding the specificity of the activity of the peptide. Accordingly, we chose to draw peptides from other classes of bioactive peptides, so that the predictor should have greater specificity in predicting activity for the particular defined class, over and above other classes of bioactive peptide. Positive peptides were extracted from each of the benchmark datasets. Negative peptides for each dataset were automatically drawn by AutoPeptideML from a curated version of the Peptipedia database of bioactive peptides (see Methods - Algorithm Step 2), excluding similar bioactivity classes to the dataset under test (see Supplementary). To consider in isolation the effect of automated negative choice, we constructed a dataset (termed "Autoneg") of these negatives and positives without considering homology relationships, by splitting data into training and evaluation subsets randomly.

Correcting for sequence homology between training and evaluation sets. It is interesting to note that among the 18 datasets, only 8 consider the homology relationships between training and evaluation sets, of which four of them [18] perform an approximation to homology maximisation between the training and evaluation sets, by clustering the peptides according to their feature similarities and assigning 80% of the samples in each cluster to training and 20% to testing, and the remaining randomly assign peptides to either train or evaluation. The problem with homology maximisation is that it deliberately increases the non-independence of training and evaluation sets, thus impeding the

Table 1: Original benchmark datasets.

Dataset	Negative Class	Partition Strategy	Model
Antibacterial [18]	Random non-antimicrobial peptides from UniProt	Approximation to homology maximisation	[18]
ACE inhibitor [16]	Random fragments of non-antihypertensive proteins in SwissProt	Homology reduction with CD-HIT at 90%	[16]
Anticancer 1 [4]	Antimicrobial peptides	Random	[28]
Anticancer 2 [4]	Random fragments of proteins in SwissProt	Random	[4]
Antifungal [18]	Random non-antimicrobial peptides from UniProt	Approximation to homology maximisation	[18]
Antimalarial 1 [12, 62]	Random non-antiparasitical peptides from UniProt	Random	[12]
Antimalarial 2 [12, 4]	Random fragments of proteins in SwissProt	Random	[12]
Antimicrobial [18]	Random non-antimicrobial peptides from UniProt	Approximation to homology maximisation	[18]
Antioxidant [17]	Mix of experimentally validated negatives and random peptides from UniProt with same length distribution as positives	Homology reduction with Needleman-Wunch pairwise alignments at 90%	[17]
Antiparasitic [63]	Random non-antiparasitical peptides from Uniprot	Homology reduction with CD-HIT at 90% for positives and 60% for negatives	[63]
Antiviral [18]	Random non-antimicrobial peptides from UniProt	Approximation to homology maximisation	[18]
Brain-blood barrier [15]	Random non-brain-blood barrier crossing peptides from Uniprot	Homology reduction with CD-HIT at 90%	[15]
DPPIV inhibitor [10]	Mix of random peptides and peptides known to belong to phage virion proteins, be tumour T-cell antigens, and be perceived by humans as bitter	Random	[64]
Anti MRSA [13]	Random peptides from UniProt	Homology reduction with CD-HIT at 80%	[13]
Neuropeptides [9]	Random fragments of proteins in SwissProt	Homology reduction with CD-HIT-2D at 90%	[14]
Quorum sensing [65]	Random peptides from UniProt	Random	[66]
Toxic [29]	Random non-toxic peptides from Swiss-Prot	Random	[29]
Tumor T-cell antigens [11]	T-cell antigens not associated to any disease	Random	[11]

proper evaluation of generalisability or the diagnosis of over-fitting. Four of the studies (Table 1) perform homology reduction at a very high similarity threshold, which has the advantage of not reducing the size of the dataset for training, but the disadvantage of not significantly increasing the independence of the evaluation set. To consider the effects of the homology partitioning we constructed a second set of datasets (termed Autoneg+HP), where homology partitioning (see Methods - Algorithm Step 3) strongly minimised similarity between training and evaluation datasets.

Table 4 indicates that for 13 of the 18 datasets, at least 10% of the peptides in the training set share homology with at least one sequence in the evaluation dataset, which compromises their independence. Looking at the partitioning strategies used to generate the datasets, ACE inhibitor, Antioxidant, Antiparasitic, Anti-MRSA, and Neuropeptides all have been constructed with homology reduction strategies with high threshold 80-90%. Of the five datasets, there are two with only 1% of their training sequences being homologous to at least one sequence in the evaluation set, thus showing how homology reduction strategies, though insufficient for fully ensuring the independence, can, nonetheless, reduce the magnitude of the problem. The datasets built with the approximation to homology maximisation (Antibacterial, Antifungal, Antimicrobial, and Antiviral) all show relatively high proportion of homologous sequences in training (26-47%) as expected.

The middle column (Table 4) shows that this effect is similar when the negatives are automatically chosen from other bioactivity classes. However, when homology partitioning is applied, the dependence between training and evaluation sets is largely eliminated, with no peptide pairs displaying a similarity above 30%.

Application. We applied the AutoPeptideML automated analysis to the benchmark datasets, using the default settings for negative selection, homology partitioning, hyperparameter optimisation and protein language model representation. As the automatic search for negatives is a random process, and this stochasticity can influence model performance, every experiment was performed in triplicate. Figure 2, shows that the original datasets in general appear to outperform

Table 2: Proportion of sequences in training set with at least one homolog in the evaluation set. We consider a sequence to be homolog to another when their sequence identity is superior to 30%.

Dataset	Original	Autoneg	Autoneg+HP
Antibacterial	0.36	0.64	0.0
ACE inhibitor	0.01	0.03	0.0
Anticancer 1	0.59	0.50	0.0
Anticancer 2	0.30		
Antifungal	0.34	0.58	0.0
Antimalarial 1	0.41	0.15	0.0
Antimalarial 2	0.10		
Antimicrobial	0.47	0.63	0.0
Antioxidant	0.01	0.0	0.0
Antiparasitic	0.11	0.32	0.0
Antiviral	0.26	0.44	0.0
Blood-brain barrier	0.04	0.04	0.0
DPPIV inhibitor	0.05	0.04	0.0
Anti-MRSA	0.15	0.34	0.0
Neuropeptide	0.24	0.24	0.0
Quorum sensing	0.18	0.09	0.0
Toxicity	0.56	0.56	0.0
Tumour T-cell antigens	0.00	0.03	0.0

the automatically assigned peptides in terms of predictive performance. Thus, the definition used for the negative class has a significant impact on the perceived performance of the models. A closer inspection of the results highlights this observation. Particularly interesting are the pairs of datasets Anticancer 1 and 2, and Antimalarial 1 and 2 which have been created with the same set of positive peptides and differ on the definition used in the original benchmark [4] to create the negative set. Anticancer 1 defines its negatives as peptides with antimicrobial activity, which is a bioactivity that overlaps heavily with the positive class [67], thus explaining the comparatively low performance achieved when compared with Anticancer 2, which draws its negative set from a collection of random peptides. Similarly, Antimalarial 1 draws its negative class from a distribution of peptides from UniProt which is a narrower distribution than the one used in Antimalarial 2, random protein fragments [12, 62]. These results indicate that broad negative class definitions lead to better perceived model performance, even though this may not reflect their real world-application.

The definition used in this study to generate the new datasets is a middle point between Anticancer 1 and 2. This is reflected in the drop in apparent performance observed in most datasets with the exception of Anticancer 1 (which has a narrower definition) and Tumour T-cell antigens (which also use a narrower negative class definition, namely, T-cell antigens not known to be associated with any disease). The Antiviral dataset is the only case where performance increased after narrowing the negative class definition and the reason behind this discrepancy remains unclear.

This hypothesis is further supported by the results of Table 2 showcasing the independence between training and evaluation sequences. The dataset with the new bioactive negative peptides for dataset Anticancer 1 has 50% of its training sequences with at least one homolog in the evaluation set, this value is an intermediate point between the original Anticancer 1 and Anticancer 2 (59% and 30%) which is a difference that reflects the further difference in apparent performance in Figure 2.

Overall, this result shows that using bioactive peptides as a negative class generally leads to more challenging classification problems than distinguishing between a certain bioactivity and random peptides. Generally speaking, the definition of negative class depends on the problem the researcher wants to tackle specifically, however, for the purposes of an automated tool, the proposed definition balances specificity with usability and interpretability.

The principle of the new homology partitioning algorithm is to exploit the topology of the similarity graph to effectively generate out-of-distribution evaluation sets. Homology partitioning thus manifests as a drop in perceived model performance. This indicates that alternative methods tend to overestimate model performance by not properly diagnosing model overfitting. However, a thorough comparison among alternative homology reduction and partitioning methods remains an open research question.

3.2 Effect of HPO and model ensemble on performance

The current approaches for peptide bioactivity predictors can be divided into 2 categories: dataset-specific models [3] and general model development frameworks [35]. The former, as the name suggests, refers to those models that

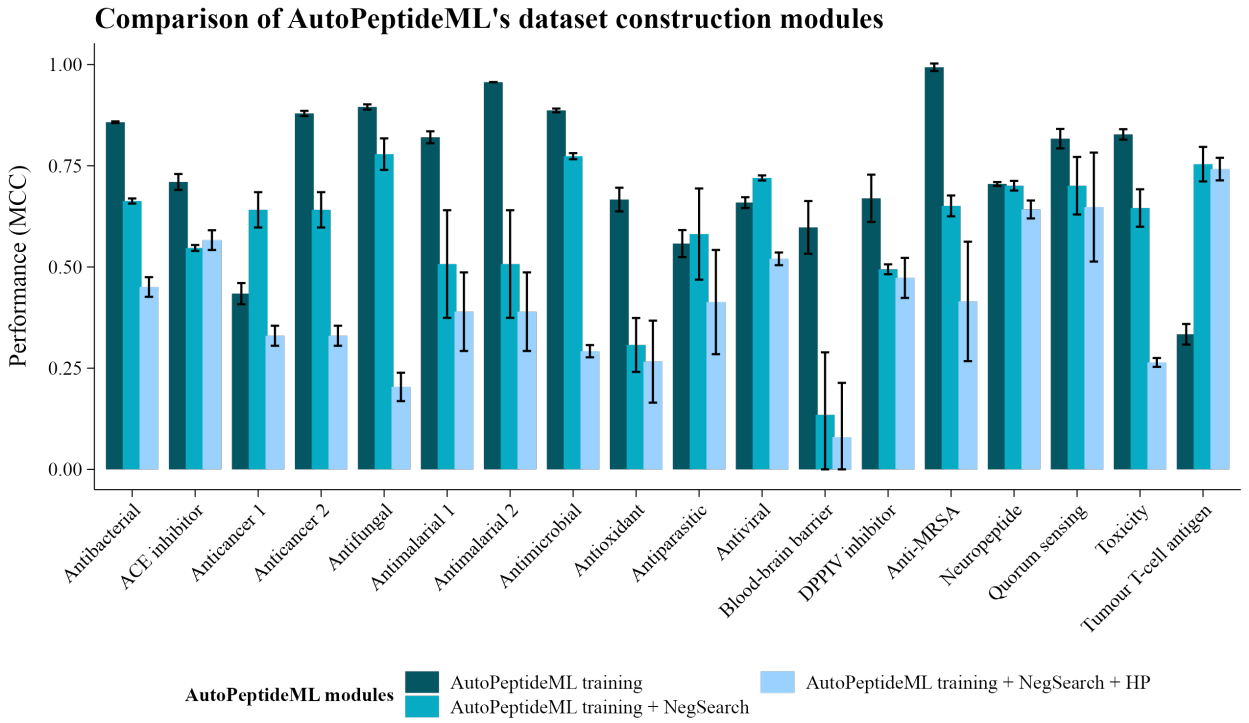

Figure 2: Evaluation of AutoPeptideML's dataset construction modules. Error bars reflect the standard deviation across three replicates. NegSearch: AutoPeptideML's negative search module (Algorithm - Step 2); HP: AutoPeptideML's homology partitioning algorithm (Algorithm - Step 3).

have been designed for a specific dataset while the latter, comprises general approaches that can be applied to any given dataset. AutoPeptideML would belong to the second category. UniDL4BioPep [35] is another algorithm that can be regarded as a general-purpose framework in this domain. Therefore, it was used as a baseline against which AutoPeptideML can be compared.

An inherent problem when comparing algorithms is that the data used to train and evaluate the models have to be the same. To do this a custom implementation of UniDL4BioPep was prepared, named UniDL4BioPep-A (see Methods) which selects the best checkpoint out of a 10-fold cross-validation. The training module of AutoPeptideML was evaluated against UniDL4BioPep-A in both the old and the new sets of benchmarks. The new benchmarks were again partitioned using both the random and homology partitioning strategies, to determine whether using an ensemble of simple ML models (AutoPeptideML) was a better strategy for avoiding overfitting in small datasets than the more complex CNN model (UniDL4BioPep-A). The self-reported values for the handcrafted models referenced in Table 1 are included with the evaluation in the original set of benchmarks to contextualise the contributions of both general purpose frameworks.

Figure 3.A shows that when applied to a literature derived benchmark set of positive and negative datasets, the two general purpose PLM-enabled bioactivity predictors have a performance comparable with the literature self-reported performance of models handcrafted for each specific dataset. Secondly, the two general purpose predictors, one of which is based on a convolutional neural network (UniDL4BioPep-A), and the other on an ensemble of three simpler ML models (AutoPeptideML), both have comparable performance.

When compared with Fig 3.A, Figure 3.B shows that when automatic negative selection from a bioactive dataset are introduced, both ML and DL general purpose models show an equivalent drop in apparent performance, reflecting the more challenging task of predicting specificity versus other bioactives. This performance drops further in both models when homology partitioning is introduced, illustrating the effect of the literature benchmark negatives' lack of independence on inflating the performance results. Remarkably, there is no significant evidence of greater overfitting on the part of the DL model, despite the small dataset sizes, this might be due to the relatively small number of parameters within the 1D-CNN. Both approaches are better than their counterparts only in 8 out 16 datasets, showing how similar their performance is. Overall, these results allow us to conclude that using an ensemble of ML models as a downstream

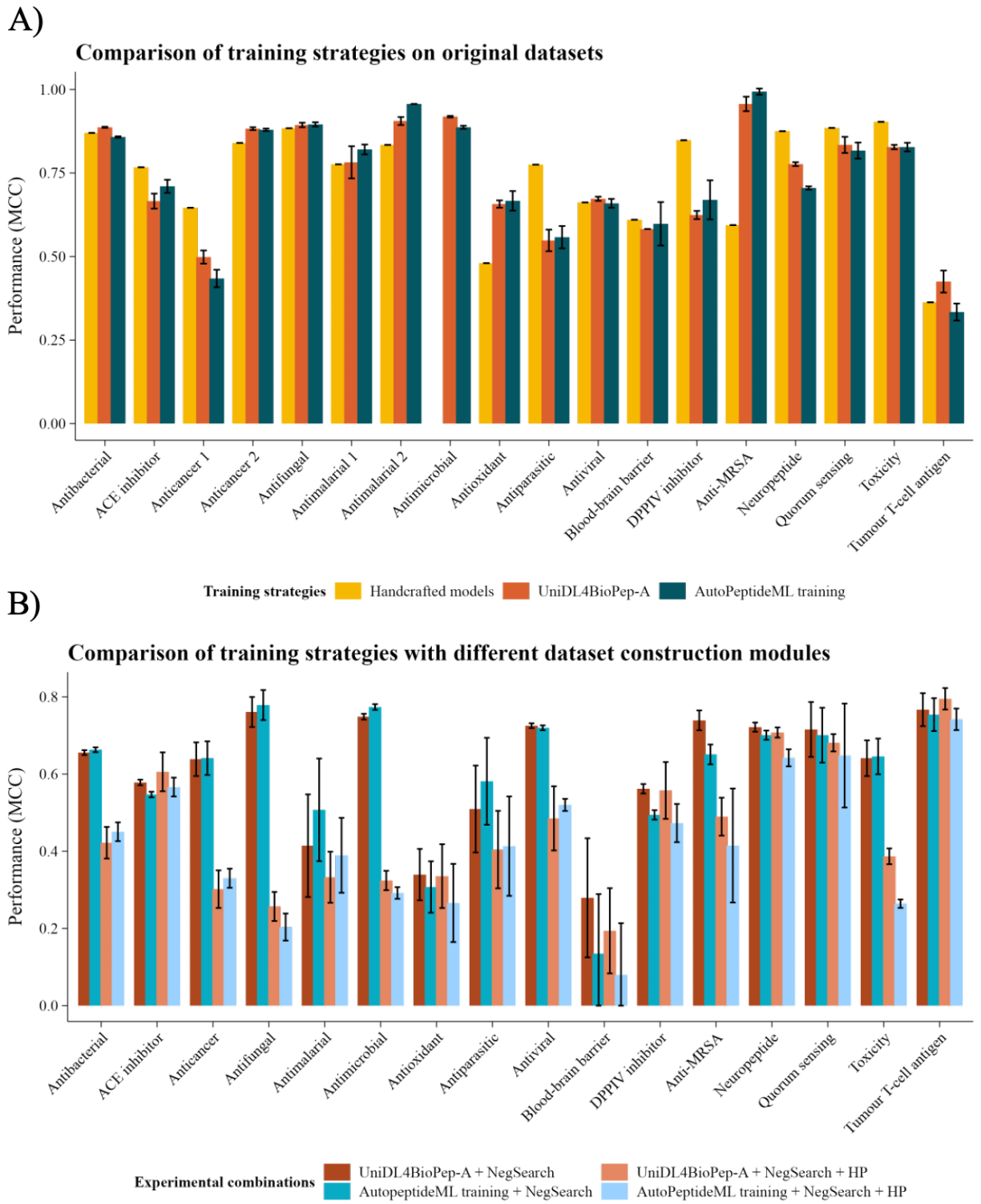

Figure 3: A: Comparison of training strategies on original datasets. B: Comparison of training strategies with different dataset construction modules. Error bars reflect the standard deviation across three replicates. NegSearch: AutoPeptideML's negative search module (Algorithm - Step 2); HP: AutoPeptideML's homology partitioning algorithm (Algorithm - Step 3).

model to predict peptide bioactivity is not only more computationally efficient than using a DL model, but that it also does not compromise model performance.

3.3 Effect of ESM-2 model size on downstream performance

Recent studies focused on language model development have reported that there exists a direct correlation between model size and downstream performance on natural language processing tasks [68, 69]. The same phenomenon has been observed for PLMs for different protein representation related tasks [31]. While PLMs have been utilised in predicting peptide bioactivity [35, 36], the relationship between model size and downstream performance has not been explored.

We evaluated four different PLM models from the ESM family with increasing size: ESM2-8M (8M parameters), ESM2-35M (35M parameters), ESM2-150M (150M parameters), ESM2-650M (650M parameters). We also evaluated ESM1b-650M (650M parameters), from a previous version of ESM. We performed the comparisons with auto-negative and homology partitioning strategies in place.

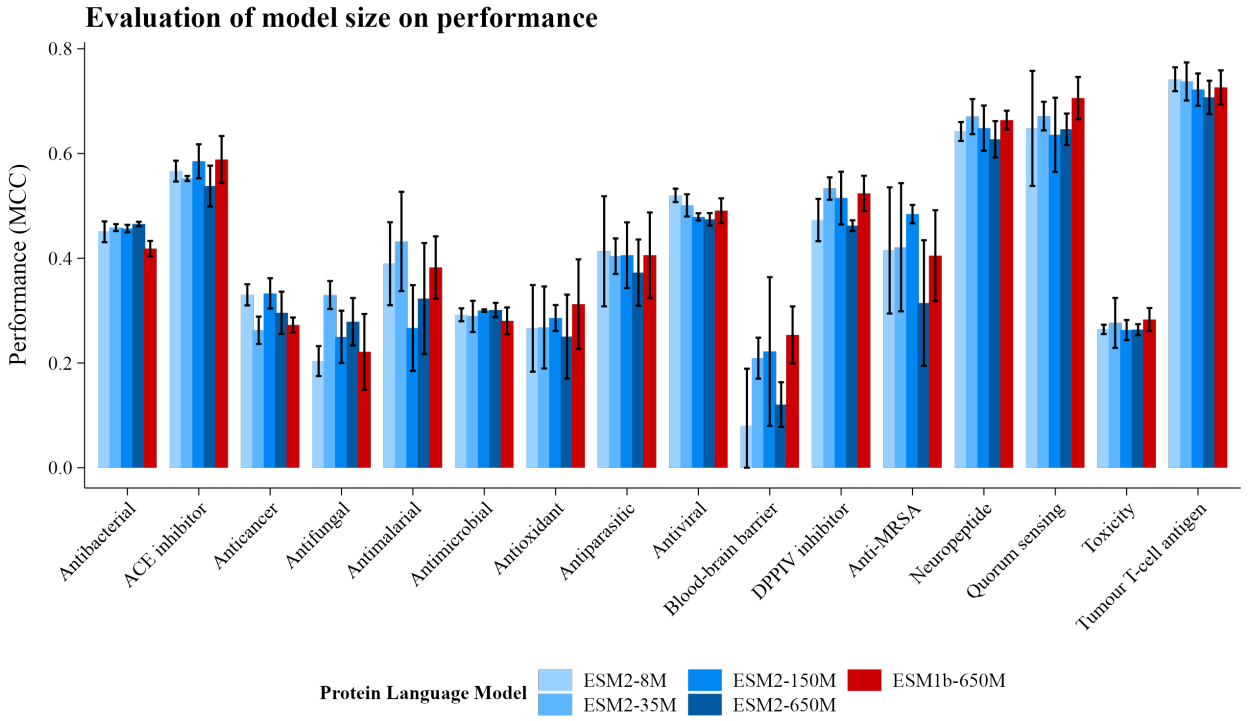

Figure 4: Model size effect on performance. Error bars reflect the standard deviation across three replicates.

Figure 4 shows there is no significant difference between models across all datasets and no correlation between model size and performance. Consequently, the smallest model is a valid default representation method. This result does not align with prior literature that describes improved downstream performance of protein predictors as model size increases [31, 32].There are two factors, however, that may explain this apparent incongruity. First, peptide sequences are less constrained than that of proteins and, therefore, the evolutionary rules available to the larger models may not be as relevant for informing downstream models. Secondly, the shorter length of peptide sequences limits the context available for the attention mechanism within the PLM model, thus, reducing the importance of model size.

4 Conclusion

AutoPeptideML is a computational tool that allows non-specialised users to develop ML models for the prediction of peptide bioactivity and facilitates compliance with ML community guidelines for the life-sciences. It is able to handle several key steps in the peptide bioactivity predictor development life-cycle including: 1) automatic search for negative samples, 2) dataset partitioning using a novel algorithm for homology partitioning, 3) hyperparameter search optimisation, 4) evaluation of soft-voting ensemble with best simple ML models generated, and 5) prediction

of new samples. Further, the output is generated in the form of a PDF summary easily interpretable by researchers not specialised in ML; alongside a directory that ensures reproducibility by containing all necessary information for re-using and re-training the models.

The definition of the negative class used for building peptide bioactivity datasets has a significant impact on the model performance and has to be controlled in order to properly understand the predictions of the models being built. The partitioning strategy employed to define training and evaluation sets impacts model generalisability and thus has an impact on apparent performance of models. Maximising dissimilarity between training and evaluation creates more challenging evaluation sets that test model generalisation to out-of-distribution tasks, closely reflecting future real-world performance.

The training module in AutoPeptideML reaches state-of-the-art performance when compared both to an alternative general-purpose-framework and dataset-specific, handcrafted models across a set of 18 different datasets. Furthermore, there is no significant difference between using an ensemble of simple ML algorithms (AutoPeptideML) and more complex DL algorithms (UniDL4BioPep-A), even though the former is more computationally efficient.

Finally, PLM model size does not affect significantly downstream performance on the peptide bioactivity datasets evaluated in this study. We hypothesise that it can be due to two reasons: first biological, second computational. The biological hypothesis is that peptide sequences evolve with fewer constraints compared to protein sequences. Hence the rules governing complex contextual situations in larger models may not be relevant for informing downstream models. The computational cause is that the shorter length of peptide sequences limits the context for the PLM attention mechanism thus reducing its relevance. Further research is needed to confirm the role of these two factors in the phenomenon observed.

The foundational principles underlying the issues described and solutions implemented throughout this study are relevant for the application of trustworthy ML predictors for any other biosequence (e.g., DNA, RNA, proteins, peptides, DNA methylation, etc.) and their automation facilitates the rigorous evaluation of new models by researchers not specialised in ML.

Acknowledgments

The authors would like to thank Silvia González López for designing the AutoPeptideML logo. The authors would also like to thank Marcos Martínez Galindo for his technical assistance in setting up the AutoPeptideML webserver. RFD would also like to thank Science Foundation Ireland for the support received through the Centre for Research Training in Genomics Data Science.

Funding

CA was supported by Enterprise Ireland and received funding from the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 847402. RCP received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 778247 (IDPFun).

References

[1] Patel, V. & Shah, M. Artificial intelligence and machine learning in drug discovery and development. *Intelligent Medicine* **2**, 134–140 (2022).

[2] Jones, D. T. & Thornton, J. M. The impact of alphafold2 one year on. *Nature methods* **19**, 15–20 (2022).

[3] Attique, M., Farooq, M. S., Khelifi, A. & Abid, A. Prediction of therapeutic peptides using machine learning: computational models, datasets, and feature encodings. *IEEE Access* **8**, 148570–148594 (2020).

[4] Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N. & Raghava, G. P. Anticp 2.0: an updated model for predicting anticancer peptides. *Briefings in bioinformatics* **22**, bbaa153 (2021).

[5] Walsh, I. *et al.* Dome: recommendations for supervised machine learning validation in biology. *Nature methods* **18**, 1122–1127 (2021).

[6] Jones, D. T. Setting the standards for machine learning in biology. *Nature Reviews Molecular Cell Biology* **20**, 659–660 (2019).

[7] Walsh, I., Pollastri, G. & Tosatto, S. C. Correct machine learning on protein sequences: a peer-reviewing perspective. *Briefings in bioinformatics* **17**, 831–840 (2016).

[8] Rao, R. *et al.* Evaluating protein transfer learning with tape. *Advances in neural information processing systems* **32** (2019).

[9] Bin, Y. *et al.* Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features. *Journal of proteome research* **19**, 3732–3740 (2020).

[10] Charoenkwan, P., Kanthawong, S., Nantasenamat, C., Hasan, M. M. & Shoombuatong, W. idppiv-scm: a sequence-based predictor for identifying and analyzing dipeptidyl peptidase iv (dpp-iv) inhibitory peptides using a scoring card method. *Journal of proteome research* **19**, 4125–4136 (2020).

[11] Charoenkwan, P., Nantasenamat, C., Hasan, M. M. & Shoombuatong, W. ittca-hybrid: Improved and robust identification of tumor t cell antigens by utilizing hybrid feature representation. *Analytical biochemistry* **599**, 113747 (2020).

[12] Charoenkwan, P. *et al.* iamap-scm: A novel computational tool for large-scale identification of antimalarial peptides using estimated propensity scores of dipeptides. *ACS omega* **7**, 41082–41095 (2022).

[13] Charoenkwan, P. *et al.* Scmrsa: a new approach for identifying and analyzing anti-mrsa peptides using estimated propensity scores of dipeptides. *ACS omega* **7**, 32653–32664 (2022).

[14] Chen, S., Li, Q., Zhao, J., Bin, Y. & Zheng, C. Neuropred-clq: incorporating deep temporal convolutional networks and multi-head attention mechanism to predict neuropeptides. *Briefings in Bioinformatics* **23** (2022).

[15] Dai, R. *et al.* Bbppred: sequence-based prediction of blood-brain barrier peptides with feature representation learning and logistic regression. *Journal of Chemical Information and Modeling* **61**, 525–534 (2021).

[16] Manavalan, B., Basith, S., Shin, T. H., Wei, L. & Lee, G. mahtpred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* **35**, 2757–2765 (2019).

[17] Olsen, T. H. *et al.* Anoxpepred: using deep learning for the prediction of antioxidative properties of peptides. *Scientific Reports* **10**, 21471 (2020).

[18] 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. *Journal of Chemical Information and Modeling* **61**, 3141–3157 (2021).

[19] Wang, L. *et al.* Therapeutic peptides: Current applications and future directions. *Signal Transduction and Targeted Therapy* **7**, 48 (2022).

[20] Zhao, L. *et al.* Potential of food-derived bioactive peptides in alleviation and prevention of alzheimer's disease. *Food & Function* **13**, 10851–10869 (2022).

[21] Fernandez-Tome, S. *et al.* Role of food proteins and bioactive peptides in inflammatory bowel disease. *Trends in Food Science & Technology* **88**, 194–206 (2019).

[22] He, X., Zhao, K. & Chu, X. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems* **212**, 106622 (2021).

[23] Yao, Q. *et al.* Taking human out of learning applications: A survey on automated machine learning. *arXiv preprint arXiv:1810.13306* (2018).

[24] Giguere, S. *et al.* Machine learning assisted design of highly active peptides for drug discovery. *PLoS computational biology* **11**, e1004074 (2015).

[25] Kaur, D., Uslu, S., Rittichier, K. J. & Durresi, A. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)* **55**, 1–38 (2022).

[26] Pang, Y., Yao, L., Xu, J., Wang, Z. & Lee, T.-Y. Integrating transformer and imbalanced multi-label learning to identify antimicrobial peptides and their functional activities. *Bioinformatics* **38**, 5368–5374 (2022).

[27] Harper, D. R. *et al.* Representations and strategies for transferable machine learning improve model performance in chemical discovery. *The Journal of Chemical Physics* **156** (2022).

[28] Charoenkwan, P. *et al.* Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. *Scientific reports* **11**, 3017 (2021).

[29] Wei, L., Ye, X., Xue, Y., Sakurai, T. & Wei, L. Atse: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Briefings in Bioinformatics* **22**, bbab041 (2021).

[30] Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

[31] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

[32] Elnaggar, A. *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **44**, 7112–7127 (2021).

[33] Unsal, S. *et al.* Learning functional properties of proteins with language models. *Nature Machine Intelligence* **4**, 227–245 (2022).

[34] Suzek, B. E. *et al.* Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

[35] Du, Z., Ding, X., Xu, Y. & Li, Y. Unidl4biopep: a universal deep learning architecture for binary classification in peptide bioactivity. *Briefings in Bioinformatics* bbad135 (2023).

[36] Dee, W. Lmpred: Predicting antimicrobial peptides using pre-trained language models and deep learning. *Bioinformatics Advances* **2**, vbac021 (2022).

[37] Rao, R. M., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. *bioRxiv* (2020). URL `https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1`.

[38] Rost, B. Twilight zone of protein sequence alignments. *Protein engineering* **12**, 85–94 (1999).

[39] Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

[40] Remmert, M., Biegert, A., Hauser, A. & Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods* **9**, 173–175 (2012).

[41] Hauser, M., Steinegger, M. & Söding, J. Mmseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323–1330 (2016).

[42] Almagro Armenteros, J. J. *et al.* Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology* **37**, 420–423 (2019).

[43] Teufel, F. *et al.* Graphpart: homology partitioning for biological sequence analysis. *NAR genomics and bioinformatics* **5**, lqad088 (2023).

[44] Quiroz, C. *et al.* Peptipedia: a user-friendly web application and a comprehensive database for peptide research supported by machine learning approach. *Database* **2021** (2021).

[45] Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026–1028 (2017).

[46] Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **48**, 443–453 (1970).

[47] Rice, P., Longden, I. & Bleasby, A. Emboss: the european molecular biology open software suite. *Trends in genetics* **16**, 276–277 (2000).

[48] Farrar, M. Striped smith–waterman speeds database searches six times over other simd implementations. *Bioinformatics* **23**, 156–161 (2007).

[49] Polyanovsky, V. O., Roytberg, M. A. & Tumanyan, V. G. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. *Algorithms for molecular biology* **6**, 1–12 (2011).

[50] Russell, R. B. & Barton, G. J. Structural features can be unconserved in proteins with similar folds: An analysis of side-chain to side-chain contacts secondary structure and accessibility. *Journal of molecular biology* **244**, 332–350 (1994).

[51] May, A. C. Percent sequence identity: the need to be explicit. *Structure* **12**, 737–738 (2004).

[52] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[53] Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer go annotations beyond homology. *Scientific reports* **11**, 1–14 (2021).

[54] Schütze, K., Heinzinger, M., Steinegger, M. & Rost, B. Nearest neighbor search on embeddings rapidly identifies distant protein relations. *bioRxiv* 2022–09 (2022).

[55] Zhang, N. *et al.* Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147* (2022).

[56] Zhou, H.-Y., Fu, Y., Zhang, Z., Bian, C. & Yu, Y. Protein representation learning via knowledge enhanced primary structure modeling. *bioRxiv* 2023–01 (2023).

[57] Zhang, Z. *et al.* A systematic study of joint representation learning on protein sequences and structures (2023). `2303.06275`.

[58] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631 (2019).

[59] Kumari, S., Kumar, D. & Mittal, M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering* **2**, 40–46 (2021).

[60] Nakano, R. *et al.* scikit-plot: v0.3.5 (2018). URL `https://doi.org/10.5281/zenodo.1245853`.

[61] McKinney, W. *et al.* pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing* **14**, 1–9 (2011).

[62] Xiao, X., Shao, Y.-T., Cheng, X. & Stamatovic, B. iamp-ca2l: a new cnn-bilstm-svm classifier based on cellular automata image for identifying antimicrobial peptides and their functional types. *Briefings in bioinformatics* **22**, bbab209 (2021).

[63] Zhang, W. *et al.* Predapp: predicting anti-parasitic peptides with undersampling and ensemble approaches. *Interdisciplinary Sciences: Computational Life Sciences* 1–11 (2022).

[64] Charoenkwan, P. *et al.* Stackdppiv: A novel computational approach for accurate prediction of dipeptidyl peptidase iv (dpp-iv) inhibitory peptides. *Methods* **204**, 189–198 (2022).

[65] Rajput, A., Gupta, A. K. & Kumar, M. Prediction and analysis of quorum sensing peptides based on sequence features. *PLoS One* **10**, e0120066 (2015).

[66] Wei, L. *et al.* Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Briefings in Bioinformatics* **21**, 106–119 (2020).

[67] Tornesello, A. L., Borrelli, A., Buonaguro, L., Buonaguro, F. M. & Tornesello, M. L. Antimicrobial peptides as anticancer agents: functional properties and biological activities. *Molecules* **25**, 2850 (2020).

[68] Taylor, R. *et al.* Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085* (2022).

[69] Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

A Peptipedia length distribution

Figure 5 describes the distribution of the lengths of the peptides comprising the subset of Peptipedia generated in this study and named "AutoPeptideML-Peptipedia".

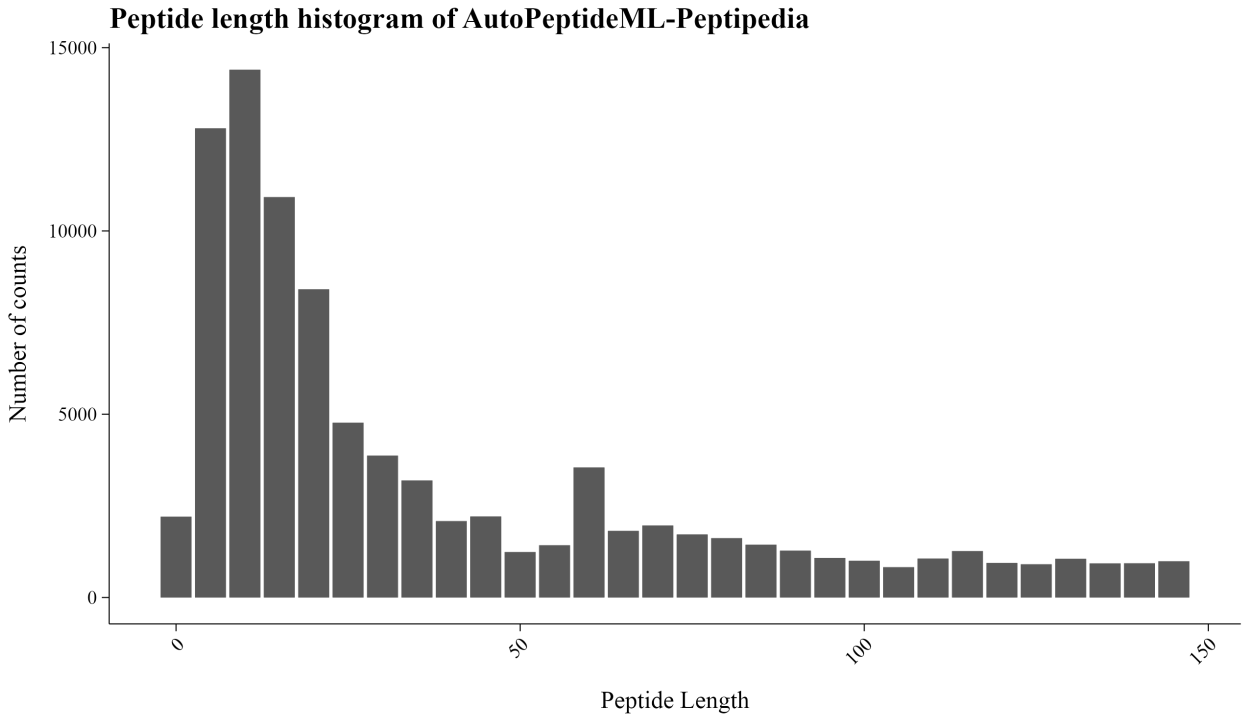

Figure 5: Histogram of peptide lengths with bin size 5.

B Default hyperparameter search space

Table 3: Default hyperparameter search space for the ensemble used throughout the paper.

Model	**Trials**	**Hyperparameter search space**			
		Name	**Type**	**Range**	**Log-scale**
KNN	10	K Weights	integer categorical	1-30 uniform or distance	No No
RFC	10	Max depth Number of estimators	integer integet	2-20 10-100	No No
LightGBM	10	Max depth Number of leaves Learning rate	integer integer float	1-30 5-50 10e-3 - 0.3	Yes Yes Yes

C Recommendations for using AutoPeptideML and reporting its results

This section explores how the structure of the outputs from AutoPeptideML facilitates compliance with DOME guidelines [5], nevertheless, it is important to note that no system can fully avoid its misuse or abuse and the ultimate responsibility of following proper guidelines and accurately reporting the results lies in the final users.

- **Data:** The algorithm ensures independence between the optimisation (training) and evaluation (test) sets. The hyperparameter optimisation and model selection, which can be considered as meta-optimisation strategies,

relies on n-fold cross-validation and maintains the independence of the evaluation set. Further, the constraints upon the algorithm in the web-server application impedes malpractices like the manual curation of parameters to meta-optimise the results in the independent test sets.

The datasets generated during the automatic search for negative samples, the train/test partitions, and the n train/validation folds are included in the ZIP-compressed output file, thus making their release and sharing easy. The automatic search for negatives is also compliant with the recommendation that the distribution of the data is representative of the domain in which the model is going to be applied. The use of random seeds for any stochastic process improves the reproducibility when the same exact datasets are used, thus guaranteeing that the different runs will produce similar results.

- **Optimisation:** Metrics for each fold in cross-validation are provided alongside the final evaluation metrics of the model so that train versus test error can be calculated as a measure of possible under- or over-fitting. The hyper-parameter configurations of the final models are included in the output file and are therefore easy to share.
- **Model:** PLMs are not directly explainable and it follows that models built on top of their representations are thus not explainable.
- **Evaluation:** Models are evaluated with a wide array of metrics and a PDF summary, meant for researchers that are not familiar with ML concepts, of the main model performance plots and evaluation metrics is provided with a guide on how to interpret them depending on different application contexts meant for researchers that are not familiar with ML concepts. Most common problems when analysing evaluation metrics arise when working with imbalanced evaluation datasets, the automatic dataset construction module bypasses this problem by generating balanced datasets.

D Search for Negative Peptides

Table 4: Overlapping classes excluded from the negative set for each of the benchmark datasets.

Dataset	**Overlapping bioactivities**
Antibacterial	`Antibacterial/antibiotic`
ACE inhibitor	`Blood_pressure,Blood_processes,Vasodilator,Vascular`
Anticancer	`Anicancer,Cytotoxic,Antitumour`
Antifungal	`Antifungal`
Antimalarial	`Antimalarial/antiplasmodial`
Antimicrobial	`Antimicrobial`
Antioxidant	`Antioxidant`
Antiparasitic	`Antiparasitic`
Antiviral	`Antiviral`
Blood-brain barrier	`Neuropeptide,Blood-brain_barrier_crossing`
DPPIV inhibitor	`Diabetic`
Anti-MRSA	`Antibacterial/antibiotic`
Neuropeptide	`Neuropeptide`
Quorum sensing	`Quorum_sensing`
Toxicity	`Cytotoxic,Neurotoxin,Toxic,Toxins`
Tumour T-cell antigens	`Immunological_activity`