BaggingCPP: An Inductive PU-Learning Framework for Discovering Cell-Penetrating Peptides

Omri Yakir
Blavatnik School of Computer Science and AI
Tel Aviv University, 69978, Tel Aviv, Israel
omriyakir@mail.tau.ac.il

Chieh Chang
Department of Oral Biology, Goldschleger School of Dental Medicine
Gray Faculty of Medical and Health Sciences
Tel Aviv University, 69978, Tel Aviv, Israel
chiehc@mail.tau.ac.il

Maayan Gal
Department of Oral Biology, Goldschleger School of Dental Medicine
Gray Faculty of Medical and Health Sciences
Tel Aviv University, 69978, Tel Aviv, Israel
mayyanga@tauex.tau.ac.il

Jérôme Tubiana
Blavatnik School of Computer Science and AI
Tel Aviv University, 69978, Tel Aviv, Israel
jeromet@mail.tau.ac.il

Abstract

Cell Penetrating Peptides (CPPs) are a promising approach for intracellular delivery of diverse molecular cargos. Although hundreds of CPPs have been previously characterized, most are cationic peptides with limited penetration efficiency or poor pharmaceutical properties; new high-throughput discovery approaches are thus needed. Here we introduce **BaggingCPP**, a machine learning-based CPP discovery framework that integrates inductive Positive-Unlabeled (PU) learning, protein language models and parameter-efficient fine-tuning algorithms. Unlike prior works, we do not use an artificial negative set that leads to distribution shift but instead use PU learning to train and infer on the same dataset - a large corpora of naturally expressed peptides such as hormones, neuropeptides, and small proteins. BaggingCPP reaches a cross-validated **AUC-ROC of 0.984** on our dataset and matches the performance of the state-of-the-art GraphCPP when both methods are trained *and* inferred on the public CPP1708 benchmark. We used BaggingCPP to identify several candidate CPPs with low similarity to any known CPP and experimentally validated two. BaggingCPP thus represents a data-driven, biologically grounded route to expand the chemical diversity of known CPPs.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

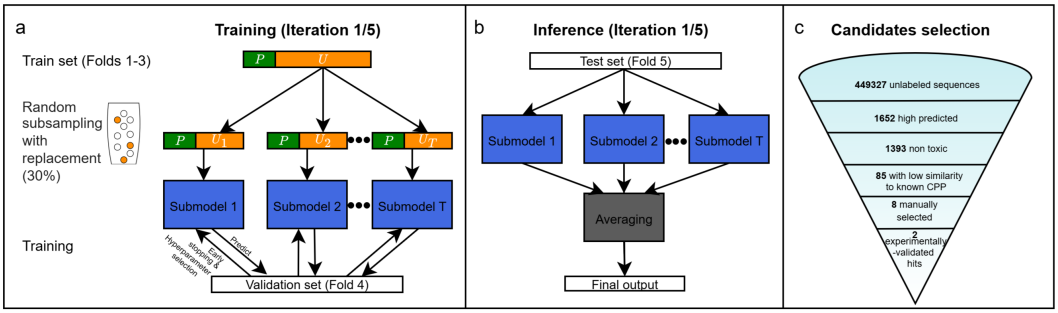

Figure 1: The inductive PU learning setting: (a) training with a bagging ensemble, (b) ensemble inference, (c) candidate selection.

1 Introduction

Cell-penetrating peptides (CPPs) are short sequences that can transport macromolecules across cellular membranes. Since the discovery of the HIV-1 Tat fragment in 1988 [1], CPPs have been widely explored as delivery tools in biomedicine [2], being tested for cancer therapy and diagnostics [3]. Yet their clinical translation remains limited by issues of stability, selectivity, and endosomal escape [4]. Recently, natural protein fragments have inspired novel CPPs, such as a defensin-derived peptide from *Medicago truncatula* that efficiently penetrates mammalian cells and delivers protein cargo with low cytotoxicity [5]. At the same time, deep learning for peptide science has matured into an established field: predictive and generative models have been applied to antimicrobial, anticancer, and cell-penetrating peptides, demonstrating that AI can efficiently explore sequence space and propose candidates with tailored properties [6, 7]. Machine learning has also enabled the design of multifunctional therapeutic peptides, such as those engineered for sustained ocular drug delivery [8]. Together, these advances highlight both the promise and the limitations of current approaches, motivating the development of **BaggingCPP**, an inductive PU-learning framework designed to overcome label bias and expand CPP diversity beyond the classical cationic-rich motifs.

Positive–Unlabeled Setting: Validated CPPs are biased toward easily detectable cationic motifs, i.e. unrandom sample of the true distribution—this is the **PGPU** assumption in PU learning [9]. As shown in Appendix Figure C.1, many CPPs cluster near the Tat fragment first described in 1988 [1]. PGPU implies that despite label bias, score rankings remain informative, making it possible to prioritize candidates. Since explicit negatives are unavailable and estimates over the unlabeled pool are noisy, we stabilize predictions with **bagging ensembles** [10], where each bootstrap resample offers a complementary view of the unlabeled set (Figure 1a), reducing variance while preserving PGPU-based ranking structure.

Methodological Innovations: BaggingCPP integrates **protein language modelling** for sequence embeddings, **Low-Rank Adaptation (LoRA)** for efficient fine-tuning, and **Light-Attention** pooling to capture long-range interactions. These components are combined within a **bagging ensemble**, an approach particularly effective in PU learning. A submodel architecture is illustrated in Figure 2.

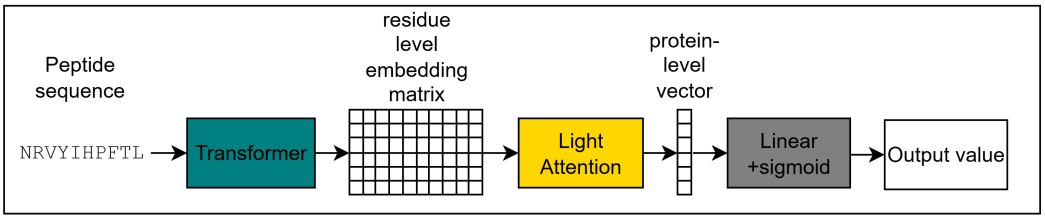

Figure 2: Architecture of the BaggingCPP submodels.

2 Methods

We implemented our framework using **PyTorch** [11] and the **Transformers** library [12], with parameter-efficient fine-tuning provided by the **PEFT/LoRA** package [13]. Supporting bioinformatics operations such as sequence parsing and structure handling were carried out with **Biopython** [14] and **Biotite** [15]. The complete workflow integrates positive–unlabeled ensembling, representation learning via Light Attention, and candidate filtering, as detailed in the subsections below. For comparison, we additionally implemented a convolutional neural network ensemble baseline, described in Appendix A.1. Training hyperparameters and optimization settings are provided in Appendix A.2.

2.1 Dataset

To maximize the discovery of novel CPPs while maintaining biological relevance, we assembled a large and diverse peptide dataset enriched for sequences with a higher *a priori* likelihood of cell penetration. Known CPPs were drawn from **CPPsite2.0** [16] (around 1200 experimentally validated sequences). For the unlabeled pool, we combined **Hmrbase2** [17] (around 2500 peptide hormones), **NeuroPep1** [18] (around 1000 neuropeptides), and **BIOPEP_UWMix** [19] (around 3500 bioactive peptides) as biologically enriched sources, and **SmProt2** [20] (around 450,000 small proteins identified from ribosome profiling) as a representative background of naturally expressed peptides. This design increases the odds that the unlabeled pool contains true but as-yet undiscovered CPPs while still reflecting the general peptide distribution, enabling our PU-learning framework to operate in a realistic discovery setting. The resulting dataset is highly imbalanced, with a positive-to-unlabeled ratio of 0.0026:1. We adopted **five-fold cross-validation** to ensure that every positive contributed to both training and testing while mitigating overfitting. To prevent data leakage, we clustered sequences with MMseqs2 [21] at 50% identity and 40% coverage and then distributed entire clusters across folds such that positives were balanced as evenly as possible.

2.2 Positive–unlabeled ensemble training protocol

Following Mordelet *et al.*[10], we trained $K{=}50$ sub-models per ensemble. For each base learner we combined the full positive set P with a disjoint 30% sample of the unlabeled pool U (treated as negatives) (Figure 1a). A base score is thus defined as $g_i(x) = \Pr\big[s{=}1 \mid x;\, P \cup U_i\big], \qquad i \in \{1, \ldots, K\}$. Where $s{=}1$ denotes a labeled (positive) instance. Final predictions are obtained by averaging the g_i, as illustrated in Figure 1b.

2.3 Model architecture

Each base learner consists of an `ESM2_t6_8M_UR50D` transformer backbone [22], fine-tuned with **Low-Rank Adaptation (LoRA)** [13] applied to the key and value projections of every layer (rank 10, $\alpha{=}8$, dropout 0.3). This reduces trainable parameters by $\sim$95% while retaining expressivity. Per-residue embeddings are passed to a **Light Attention** head [23] with a convolutional kernel of size 7, which aggregates them into a 128-dimensional sequence representation via attention-weighted pooling. The resulting vector feeds into the classification layer. An overview of the architecture is shown in Figure 2, where the transformer block denotes the LoRA-adapted ESM2 backbone.

2.4 Candidate Selection

We applied a multi-stage filtering and clustering pipeline to reduce the large unlabeled pool into a small set of experimentally testable candidates. Briefly, we combined predictions from our ensembles with additional toxicity, novelty, and developability constraints. This process reduced $\sim$450,000 unlabeled peptides to a final ranked set of 64 cluster representatives, from which 8 were selected for wet-lab validation. Full details of the candidate selection procedure are provided in Appendix A.3.

3 Results

AUC-ROC serves as our primary evaluation metric, given its threshold-free nature and direct probabilistic interpretation in the positive–unlabeled setting (see Appendix A.4)

3.1 Architecture-Related Results

Benchmarking Against Existing Methods: On the CPP1708 benchmark (Table 1), our sequence-only model achieves state-of-the-art AUC-ROC. Training on our corpus yields 0.686, while retraining on CPP1708 (∗) boosts performance to 0.848, matching `GraphCPP` [24] (0.846). This gap reflects distribution shift between datasets. Notably, a single `ESM2+LoRA+LA*` model slightly outperforms its ensemble, suggesting CPP1708 is less contaminated and more like a standard benchmark where ensembling adds little. Thus, BaggingCPP reaches GraphCPP-level accuracy without 3D structures, enabling scalable high-throughput screening.

Table 1: Comparison on the CPP1708 dataset. (*) indicates models trained on CPP1708.

Method	F1	ACC	MCC	AUC
CellPPD [25]	0.000	0.590	0.000	–
C2Pred [26]	0.328	0.571	0.054	0.650
MLCPP2.0 [27]	0.386	0.667	0.300	0.686
GraphCPP [24]	0.752	**0.795**	**0.579**	0.846
ESM2+LoRA+LA	0.535	0.646	0.316	0.686
ESM2+LoRA+LA (ensemble)	0.605	0.650	0.302	0.696
ESM2+LoRA+LA*	0.758	0.768	0.536	**0.848**
ESM2+LoRA+LA (ensemble)*	**0.776**	0.759	0.530	0.846

3.2 Impact of Training–Inference Distribution Alignment

Reliable CPP discovery requires training under deployment-like conditions; we therefore adopt an *inductive* PU learning setting, training and inferring on the same biologically enriched peptide corpus. To highlight the benefits of this approach, we compare it against a *non-inductive* regime, in which the model is trained on the public CPP1708 benchmark but evaluated on our curated corpus.

Quantitative Performance: Aligning train and inference distributions yields substantial gains: our model reaches 0.984 AUC-ROC in the inductive PU setting versus 0.846 when trained on CPP1708 and tested on our corpus (Table 2). Appendix C.2 further shows t-SNE embeddings from a single ESM2+LoRA+LA model, where positives and unlabeled peptides are clearly separated, illustrating the ensemble's discriminative capability.

Table 2: Evaluation metrics for different architectures over our dataset. Total training time is for 5 folds; params are per fold. (*) ensemble trained on CPP1708.

Base model	Fine-tune	Head	Type	Trainable params	Train time	AUC-ROC
CNN	–	MLP	Single	291,841	∼3 min	0.926
CNN	–	MLP	Ensemble	14,592,050	∼2 h	0.970
ESM2	Frozen	LA	Single	286,977	∼1.67 h	0.976
ESM2	Frozen	LA	Ensemble	14,348,850	∼33 h	0.980
ESM2	LoRA	MLP	Single	179,841	∼1.67 h	0.968
ESM2	LoRA	MLP	Ensemble	8,992,050	∼41 h	0.981
ESM2	LoRA	LA	Single	363,777	∼2 h	0.976
ESM2	**LoRA**	**LA**	**Ensemble**	**18,188,850**	**∼41 h**	**0.984**
ESM2	LoRA	LA	Ensemble*	18,188,850	∼1.5 h	0.846

Enrichment Analysis: Figure 3 compares score distributions across unlabeled sources. The CPP1708-trained model (a) fails to distinguish **SmProt** (red), the broad background of natural peptides, from enriched groups such as neuropeptides, hormones, and bioactive peptides. By contrast, the inductive PU-trained model (b) shifts SmProt scores sharply lower relative to these groups, indicating clear separation. This distinction is essential: it ensures that high-scoring candidates reflect genuine biological signal rather than dataset-specific artifacts.

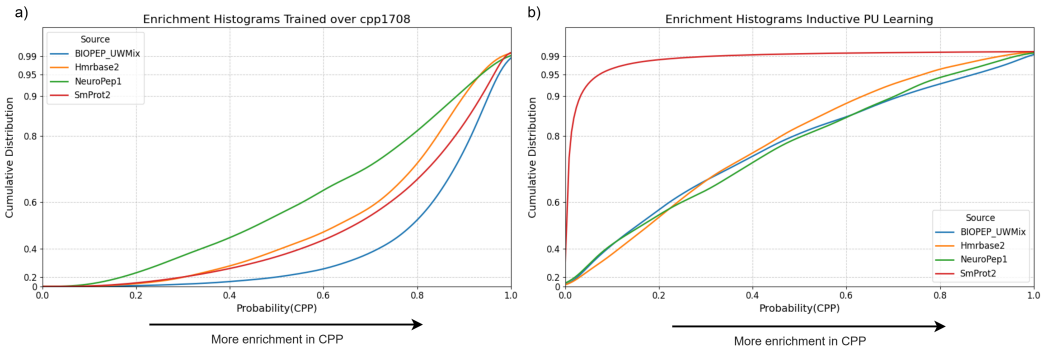

Figure 3: Enrichment histograms comparison.

3.3 Experimental validation

We experimentally tested eight model-selected candidates alongside two positive and two negative controls using a standardized FITC-uptake assay in HeLa cells (detailed in Appendix A.5). Both classes of controls behaved as expected: the known CPPs showed robust intracellular fluorescence, while the negative peptides remained at background levels. Among the eight candidates, two peptides demonstrated clear and reproducible intracellular uptake. Notably, one corresponds to a fragment of Cortistatin (fig 4,absent from our training corpora) and the other to Angiotensin I, a hormone precursor not previously reported to act as a CPP. These results highlight BaggingCPP's ability to generalize beyond peptide families represented in the data and to uncover biologically novel uptake events. Representative images and peptide-level descriptions are provided in Appendix A.5.

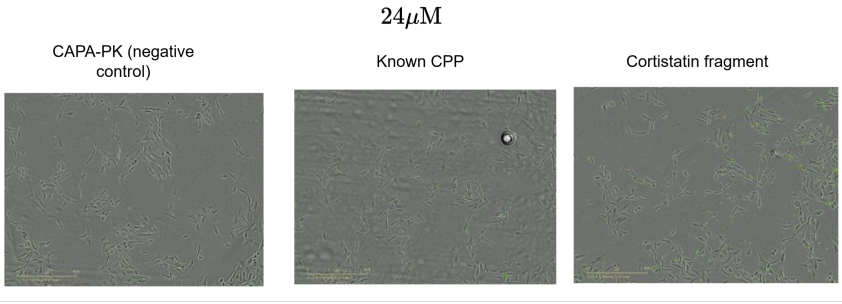

Figure 4: Representative fluorescence microscopy image of HeLa cells incubated with the Cortistatin-derived peptide (PCKNFFWKTFSSCK) at 24 µM. Clear intracellular fluorescence indicates robust uptake in this single-experiment condition.

4 Discussion

Most computational CPP studies stop short of experimental testing. BaggingCPP closes this gap through wet-lab validation of computationally predicted CPPs, achieving high predictive accuracy (AUC-ROC $= 0.984$) via inductive PU learning paired with ensembling. Training and inference on biologically enriched distributions mitigated the distribution shift that limited earlier methods. Although many models are trained, the framework remains lightweight: a five-fold ensemble trained in $\sim$41 hours with only $\sim$18M parameters per fold (Table 2), enabling rapid screening of hundreds of thousands of peptides. For comparison, we estimate that retraining the GraphCPP architecture with a similar bagging protocol our on dataset would take about 20 days.

Two predicted CPPs were validated in the first experimental round, though more follow-up assays will be necessary to evaluate their merits as delivery agents. Scaling to additional peptide sources and larger assays will accelerate discovery and elucidate the therapeutic potential of novel CPPs. In the broader landscape, our discriminative PU-learning approach is complementary to recent generative peptide LMs such as PepMLM [28], which could be coupled with BaggingCPP for integrated design–screening pipelines.

References

[1] Alan D Frankel and Carl O Pabo. Cellular uptake of the tat protein from human immunodeficiency virus. *Cell*, 55(6):1189–1193, 1988.

[2] Alessandro Gori, Giulia Lodigiani, Stella G Colombarolli, Greta Bergamaschi, and Alberto Vitali. Cell penetrating peptides: classification, mechanisms, methods of study, and applications. *ChemMedChem*, 18(17):e202300236, 2023.

[3] Prem Prakash Tripathi, Hamed Arami, Ivneet Banga, Jalaj Gupta, and Sonu Gandhi. Cell penetrating peptides in preclinical and clinical cancer diagnosis and therapy. *Oncotarget*, 9 (98):37252, 2018.

[4] Gyu Chan Kim, Dae Hee Cheon, and Yan Lee. Challenge to overcome current limitations of cell-penetrating peptides. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1869(4):140604, 2021.

[5] Lucia Adriana Lifshits, Yoav Breuer, Marina Sova, Sumit Gupta, Dar Kadosh, Evgeny Weinberg, Zvi Hayouka, Daniel Z Bar, and Maayan Gal. Nature-inspired peptide of mtdef4 c-terminus tail enables protein delivery in mammalian cells. *Scientific Reports*, 14(1):4604, 2024.

[6] Fangping Wan, Daphne Kontogiorgos-Heintz, and Cesar de la Fuente-Nunez. Deep generative models for peptide design. *Digital Discovery*, 1(3):195–208, 2022.

[7] Garyk Brixi, Tianzheng Ye, Lauren Hong, Tian Wang, Connor Monticello, Natalia Lopez-Barbosa, Sophia Vincoff, Vivian Yudistyra, Lin Zhao, Elena Haarer, et al. Salt&peppr is an interface-predicting language model for designing peptide-guided protein degraders. *Communications Biology*, 6(1):1081, 2023.

[8] Henry T Hsueh, Renee Ti Chou, Usha Rai, Wathsala Liyanage, Yoo Chun Kim, Matthew B Appell, Jahnavi Pejavar, Kirby T Leo, Charlotte Davison, Patricia Kolodziejski, et al. Machine learning-driven multifunctional peptide engineering for sustained ocular drug delivery. *Nature communications*, 14(1):2509, 2023.

[9] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine learning*, 109(4):719–760, 2020.

[10] Fantine Mordelet and J-P Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[14] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

[15] Patrick Kunzmann and Kay Hamacher. Biotite: a unifying open source computational biology framework in python. *BMC bioinformatics*, 19(1):346, 2018.

[16] Kimia Kardani and Azam Bolhassani. Cppsite 2.0: an available database of experimentally validated cell-penetrating peptides predicting their secondary and tertiary structures. *Journal of molecular biology*, 433(11):166703, 2021.

[17] Dashleen Kaur, Akanksha Arora, Sumeet Patiyal, and Gajendra Pal Singh Raghava. Hmrbase2: a comprehensive database of hormones and their receptors. *Hormones*, 22(3):359–366, 2023.

[18] Yan Wang, Mingxia Wang, Sanwen Yin, Richard Jang, Jian Wang, Zhidong Xue, and Tao Xu. Neuropep: a comprehensive resource of neuropeptides. *Database*, 2015:bav038, 2015.

[19] Piotr Minkiewicz, Anna Iwaniak, and Małgorzata Darewicz. Biopep-uwm database of bioactive peptides: Current opportunities. *International journal of molecular sciences*, 20(23):5978, 2019.

[20] Yanyan Li, Honghong Zhou, Xiaomin Chen, Yu Zheng, Quan Kang, Di Hao, Lili Zhang, Tingrui Song, Huaxia Luo, Yajing Hao, et al. Smprot: a reliable repository with comprehensive annotation of small proteins identified from ribosome profiling. *Genomics, proteomics & bioinformatics*, 19(4):602–610, 2021.

[21] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[22] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[23] Hannes Stärk, Christian Dallago, Michael Heinzinger, and Burkhard Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 2021.

[24] Attila Imre, Balázs Balogh, and István Mándity. Graphcpp: The new state-of-the-art method for cell-penetrating peptide prediction via graph neural networks. *British Journal of Pharmacology*, 182(3):495–509, 2025.

[25] Ankur Gautam, Kumardeep Chaudhary, Rahul Kumar, Arun Sharma, Pallavi Kapoor, Atul Tyagi, Open Source Drug Discovery Consortium Info@ osdd. net, and Gajendra PS Raghava. In silico approaches for designing highly effective cell penetrating peptides. *Journal of translational medicine*, 11(1):74, 2013.

[26] Hua Tang, Zhen-Dong Su, Huan-Huan Wei, Wei Chen, and Hao Lin. Prediction of cell-penetrating peptides with feature selection techniques. *Biochemical and biophysical research communications*, 477(1):150–154, 2016.

[27] Balachandran Manavalan and Mahesh Chandra Patra. Mlcpp 2.0: an updated cell-penetrating peptides and their uptake efficiency predictor. *Journal of Molecular Biology*, 434(11):167604, 2022.

[28] Tianlai Chen, Madeleine Dumas, Rio Watson, Sophia Vincoff, Christina Peng, Lin Zhao, Lauren Hong, Sarah Pertsemlidis, Mayumi Shaepers-Cheu, Tian Zi Wang, et al. Pepmlm: target sequence-conditioned generation of therapeutic peptide binders via span masked language modeling. *ArXiv*, pages arXiv–2310, 2024.

[29] Federica Barbieri, Adriana Bajetto, Alessandra Pattarozzi, Monica Gatti, Roberto Würth, Stefano Thellung, Alessandro Corsaro, Valentina Villa, Mario Nizzari, and Tullio Florio. Peptide receptor targeting in cancer: the somatostatin paradigm. *International journal of peptides*, 2013(1):926295, 2013.

A Appendix

A.1 Convolutional neural network ensemble baseline

As a lightweight baseline we designed a 1-D CNN comprising two convolutional blocks (128 filters, kernel size 5, zero-padding to maintain length) followed by two fully connected layers and a sigmoid output. Input peptides were right-padded one-hot tensors of shape 20×50. The same **PU ensemble strategy** (Section 2.3) was applied, i.e. 50 independently trained CNNs each receiving the entire positive set and 30% of U.

A.2 Training details

All models were optimised with AdamW (learning rate 10^{-3}, weight decay 10^{-2}) and trained with binary cross-entropy loss using balanced class weights for up to 200 epochs, using mini-batches of size 256. Early stopping monitored the AUC-ROC on the held-out fold with a patience of five epochs.

A.3 Candidate Selection

We began with a pool of 447,327 unlabeled peptide sequences. The selection procedure followed a multi-stage filtering and clustering pipeline (Figure 1):

1. **Model-based filtering:** Sequences were retained if both our `ESM2+LoRA+LA` ensemble and the convolutional ensemble assigned probabilities above 0.5, leaving 1,650 candidates.
2. **Toxicity filtering:** A custom toxicity predictor (see Toxicity model paragraph next) was applied, and only peptides with predicted toxicity below 0.2 were kept, resulting in 1,407 sequences.
3. **Novelty and developability constraints:** To prioritize non-redundant, short, and biophysically favorable peptides, we required sequences to satisfy:
   - $9 <$ length < 20,
   - isoelectric point < 10.5,
   - net charge < 4,
   - sequence identity < 0.6 to any known CPP (with an additional boolean constraint: the closest peptide must differ by at least 5 residues),
   - instability index < 50,
   - hydrophobicity < 1.0.

   This filtering left 81 sequences.
4. **Clustering:** Pairwise sequence identities were computed using Biotite's optimal alignment with the standard protein substitution matrix. An agglomerative clustering algorithm (average linkage, distance cutoff 0.35) grouped peptides at 65% identity. Within each cluster, the sequence with the highest average prediction score (mean of `ESM2+LoRA+LA` ensemble and convolutional ensemble) was chosen as representative, resulting in 64 clusters.

The final ranked list of 64 cluster representatives was used for downstream manual curation, from which 8 peptides were selected for wet-lab validation, as can be seen in Table 3.

Toxicity model: To remove harmful peptides, we trained a classifier on 1,826 known toxic and 7,490 non-toxic peptides. We used peptide features such as length, charge, hydrophobicity, instability index, and amino acid frequencies. To prevent overfitting, we grouped similar sequences by identity and split data by groups for training and testing. A Histogram Gradient Boosting model reached an $F_{\max}$ of 0.483 on the test set. For candidate filtering, we set a threshold of 0.2, chosen to achieve at least 75% recall on toxic peptides. This ensured that most toxic peptides were removed while still keeping many safe ones for further screening.

Table 3: Representative peptide candidates after filtering and clustering. The top 8 entries are the selected candidates; the last 4 rows are controls (two known CPPs and two negative checks).

Sequence	ESM2+ LoRA+LA ensemble	CNN ensemble	Description	Source	Toxicity	Cluster size	Mean prob.
PCKNFFWKTFSSCK	0.977	0.899	Cortistatin	Hmrbase2	0.051	8	0.938
NRVYIHPFTL	0.907	0.704	Angiotensin I	Hmrbase2	0.034	3	0.805
DNIQGITKPAIR	0.885	0.939	Immunomodulating peptide	BIOPEP_UWMix	0.021	1	0.912
YGGFLRKYPK	0.822	0.999	Opioid peptide alpha-neoendorphin	BIOPEP_UWMix	0.005	1	0.910
APEKWAAFHGSW	0.870	0.641	MIP III	Hmrbase2	0.005	1	0.755
GQTTVTKIDEDY	0.771	0.694	Antifungal peptide	BIOPEP_UWMix	0.029	1	0.732
KCIPRKDKGCI	0.604	0.713	–	SmProt2	0.037	1	0.658
GYRKPPFNGSIF	0.372	0.720	SIFamide	Hmrbase2	0.014	–	–
SGGGEGSGMWFGPRL	0.036	0.307	CAPA-Pyrokinin (CAPA-PK; FXPRL-amide)	Hmrbase2	0.044	–	–
VYRKPPFNGSIF	0.004	0.048	SIFamide	Hmrbase2	0.007	–	–
MIIYRDLISKK	0.999	1.000	–	CPPsite2.0	0.037	–	–
NYRWRCKNQN	0.998	1.000	–	CPPsite2.0	0.098	–	–

A.4 Evaluation Protocol

We report **AUC-ROC** exclusively, as it is threshold-free and invariant to the (unknown) positive-to-negative ratio, facilitating fair comparison across datasets with different class priors. Importantly, AUC-ROC has a direct probabilistic interpretation: it represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen unlabeled instance. This makes it more interpretable in the positive–unlabeled setting compared to PR-AUC, where recall and precision are confounded by the unknown number of true negatives and cannot be meaningfully interpreted.

A.5 Experimental Validation over the candidates

Experiments protocol: Peptides bearing an N-terminal fluorescein (FITC) label were synthesized by GLchina and supplied as lyophilized powders. Each peptide was first dissolved to 50 mM in DMSO, then diluted in HeLa cell culture medium to final concentrations of 24 µM and 12 µM. HeLa cells were incubated with the peptide solutions for 1 hour at 37 °C, washed twice with PBS to remove excess peptide, and following fixation with 4 percent PFA were imaged using an Incucyte live-cell analysis system. Fluorescence signal intensity within the cell monolayer was quantified directly in the Incucyte software to assess and compare the cell-penetrating efficiency of each peptide. In parallel, the number of cells in each field was determined using the Incucyte cell-count algorithm. Fluorescence values were then normalized to the corresponding cell counts to assess and compare the cell-penetrating efficiency of each peptide.

Experiments results: As a sanity check, both control classes behaved as expected: the positive controls showed strong intracellular fluorescence at the tested concentrations, whereas the negative checks remained at background levels.

Out of the eight peptide candidates, **two** exhibited clear and reproducible uptake (Table 4):

- **PCKNFFWKTFSSCK:** Its C-terminal segment *NFFWKTFSSCK* corresponds to the C-terminus of *Cortistatin*, a neuropeptide related to Somatostatin that binds all human SSTR subtypes. While Somatostatin itself is not considered a canonical CPP, several Somatostatin analogs are known to enter cells via SSTR-mediated endocytosis [29]. Notably, Cortistatin-family sequences were *absent* from our training corpora, supporting BaggingCPP's ability to generalize beyond families represented in the data.
- **NRVYIHPFTL (Angiotensin I):** A decapeptide hormone precursor. To the best of our knowledge, CPP-like uptake of Angiotensin I has not been previously reported, marking it as a novel discovery.

Representative microscopy images of these experiments are shown in Figure 5.

It should be noted that among the candidates that did not display uptake in this assay, some may still act as CPPs in other cellular contexts. In particular, they could rely on receptor-mediated inter-

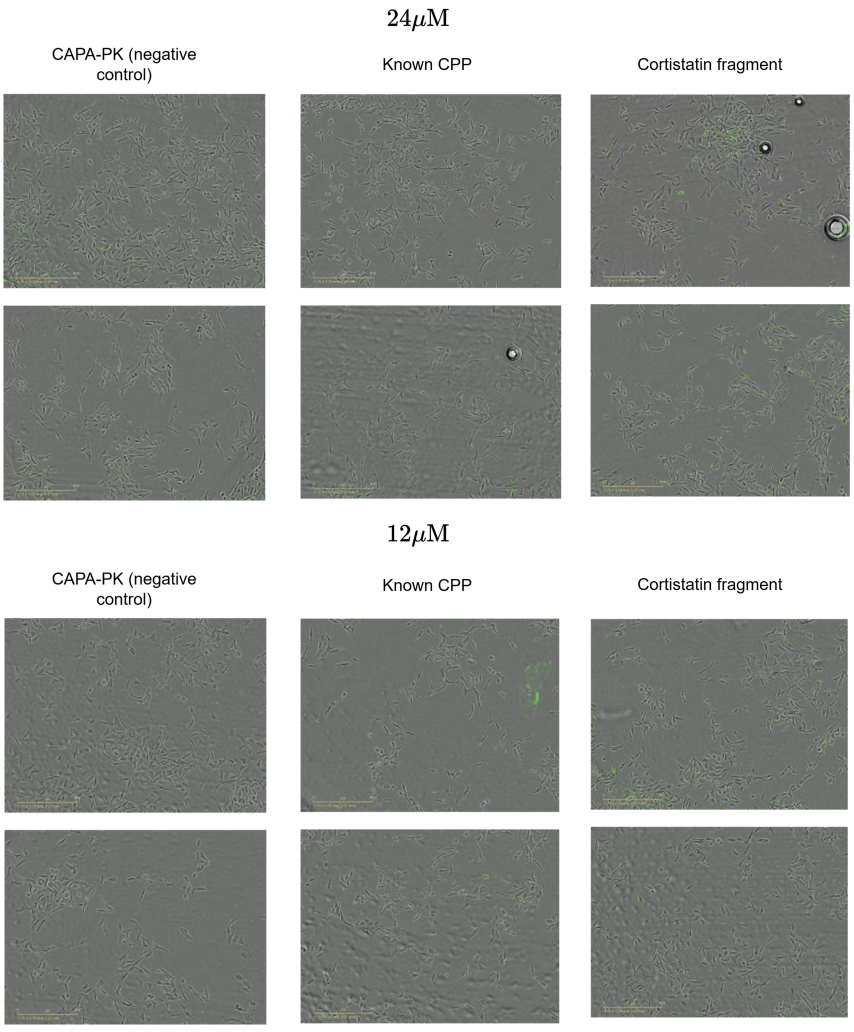

Figure 5: Representative fluorescence microscopy images of HeLa cells incubated with peptides at 24 µM and 12 µM. Left: negative control (CAPA-PK). Middle: positive control (known CPP). Right: Cortistatin fragment, showing strong intracellular fluorescence.

nalization pathways that are absent or underrepresented in HeLa cells, and might therefore require validation in additional cell lines.

Table 4: Summary of candidate peptides with model score (ESM2+LoRA+LA) and experimental fluorescence values (green integrated intensity per image normalized to confluency, mean±SD).

Name	Sequence (N-terminus FITC)	Score	Fluor. 12 µM	Fluor. 24 µM
Cortistatin fragment	FITC-PCKNFFWKTFSSCK	0.977	$108{,}944.5 \pm 46{,}227.1$	$367{,}007.8 \pm 139{,}605.2$
Angiotensin I	FITC-NRVYIHPFTL	0.907	$13{,}871.9 \pm 10{,}093.7$	$15{,}835.6 \pm 9{,}062.5$
Immunomodulating pep.	FITC-DNIQGITKPAIR	0.885	0.0 ± 0.0	454.3 ± 759.9
Alpha-neoendorphin	FITC-YGGFLRKYPK	0.822	$6{,}265.9 \pm 10{,}951.8$	$5{,}570.5 \pm 7{,}128.3$
MIP III	FITC-APEKWAAFHGSW	0.870	$3{,}006.8 \pm 6{,}103.7$	$3{,}678.7 \pm 10{,}053.3$
Antifungal peptide	FITC-GQTTVTKIDEDY	0.771	96.6 ± 198.6	117.6 ± 332.5
–	FITC-KCIPRKDKGCI	0.604	$4{,}825.6 \pm 8{,}208.2$	$3{,}555.4 \pm 4{,}186.2$
SIFamide	FITC-GYRKPPFNGSIF	0.372	$5{,}116.5 \pm 8{,}378.7$	$4{,}201.6 \pm 8{,}677.8$
CAPA-PK (control)	FITC-SGGGEGSGMWFGPRL	0.036	$2{,}860.3 \pm 1{,}712.9$	$8{,}670.7 \pm 16{,}585.8$
SIFamide (control)	FITC-VYRKPPFNGSIF	0.004	$8{,}290.0 \pm 16{,}562.4$	$7{,}576.2 \pm 5{,}616.5$
Known CPP (control)	FITC-MIIYRDLISKK	0.999	$25{,}232.8 \pm 25{,}941.3$	$45{,}197.3 \pm 22{,}710.1$
Known CPP (control)	FITC-NYRWRCKNQN	0.998	$25{,}517.2 \pm 22{,}998.3$	$34{,}303.7 \pm 19{,}091.4$

B Additional Figures

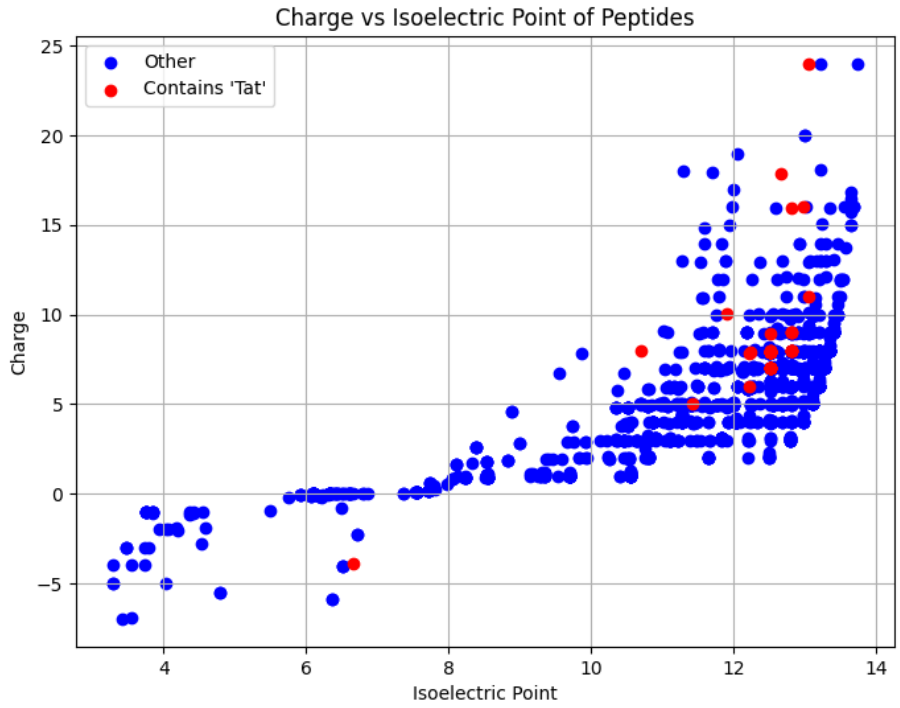

Figure C.1: Scatter plot of peptide charge versus isoelectric point for the positive set. Sequences containing the canonical HIV-1 *Tat* fragment (YGRKKRRQRRR) [1] are highlighted in red, while all other positives are shown in blue. Most points cluster in the highly cationic region, and might reflect a bias in experimentally validated CPPs toward Tat-like motifs.

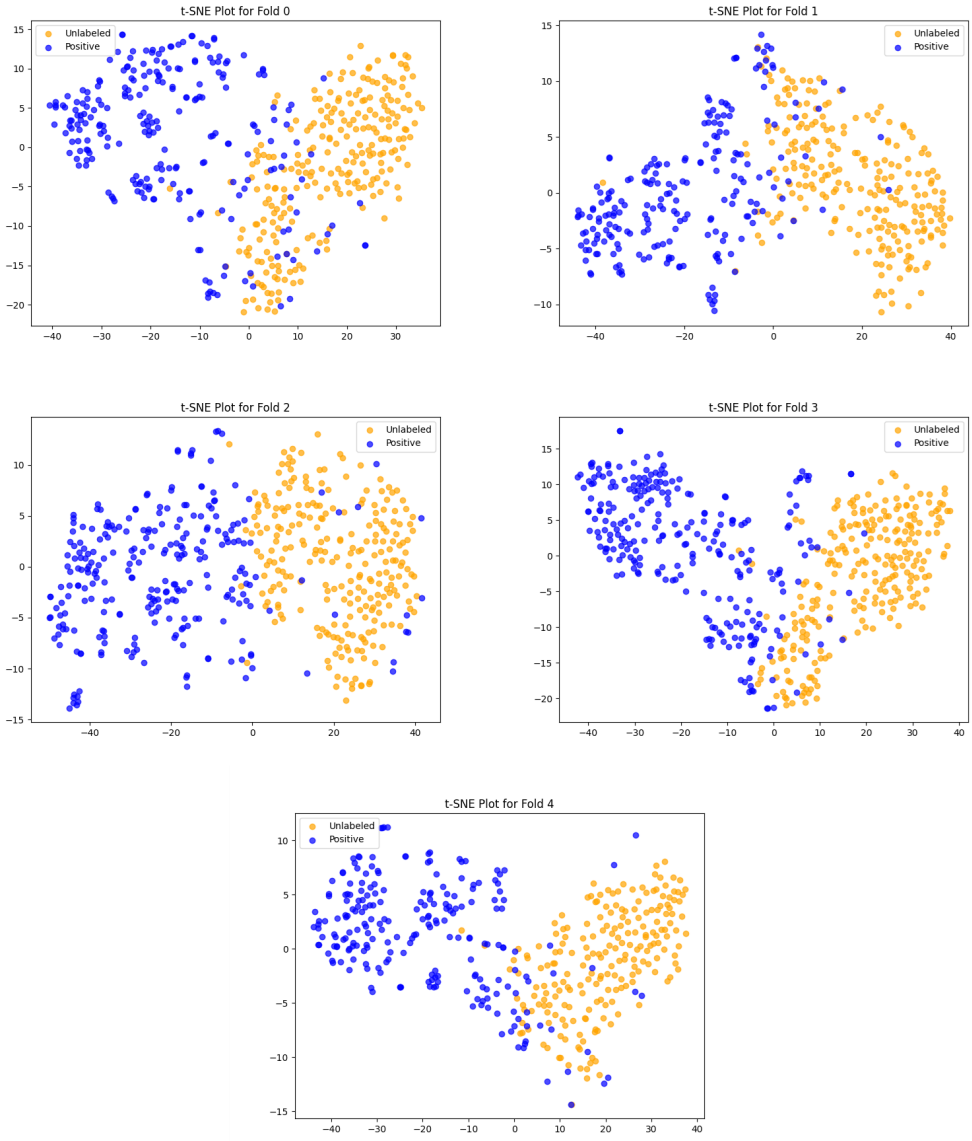

Figure C.2: Embeddings of ESM2+LoRA+LA 5 folds over a 1:1 positives to negatives, subset of our dataset were all of the positives were taken and the negatives were randomly selected.