Proceedings of the 36th European Peptide Symposium
Michal Lebl (Editor)
European Peptide Society, 2022 *https://doi.org/10.17952/36EPS.2022.207*

Overcoming the Challenges in Machine Learning-Guided Antimicrobial Peptide Design

Fabien Plisson

CINVESTAV Unidad Irapuato, Department of Biotechnology and Biochemistry. Km 9.6 Libramiento Norte Carretera Irapuato-León, C.P. 36824 Irapuato, Guanajuato, Mexico

Introduction

Antimicrobial peptides (AMPs) are rich and structurally polypeptide sequences of 12-50 residues that can kill pathogens by either disrupting their membranes or interacting with their intracellular targets [1,2]. Their direct antibacterial activities and the lack of bacterial resistance have stimulated their therapeutic avenues against antibiotic-resistant infections [3]. Major limitations preventing AMPs from translating into clinics are their low metabolic stability, poor oral bioavailability and high toxicity. Reducing hurdles to clinical trials without compromising the therapeutic promises of peptide candidates becomes an essential step in peptide-based drug design. Artificial intelligence (AI) algorithms intertwine predictive and generative models to design optimal AMP sequences rationally [4,5]. Here, we present some of the current challenges to develop robust and fair models for the discovery and design of safe AMPs.

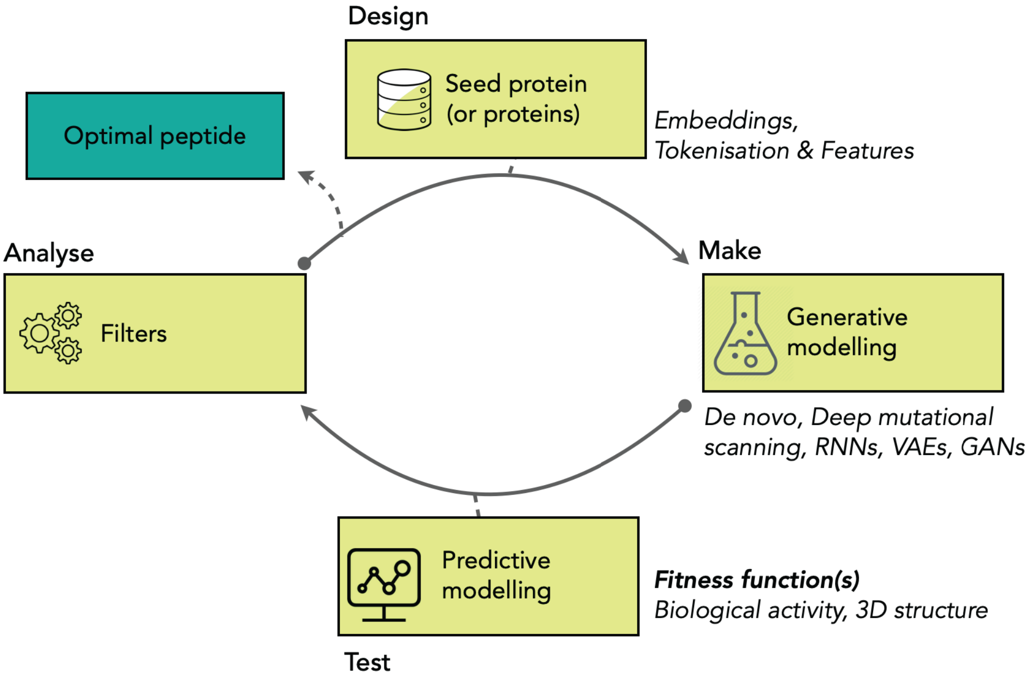
Design
Seed protein (or proteins)
Embeddings, Tokenisation & Features
Optimal peptide
Analyse
Filters
Make
Generative modelling
De novo, Deep mutational scanning, RNNs, VAEs, GANs
Predictive modelling
Fitness function(s)
Biological activity, 3D structure
Test

Fig. 1. Streamlining the DMTA cycle for computational peptide design - Machine learning (ML) and deep learning (DL) models are cost-effective and time-saving strategies used to predict biological activities from primary sequences or generate novel sequences.

Results and Discussion

Use the right descriptors. Traditional machine learning (ML) algorithms require features/variables (*e.g.*, amino acid composition, *k*-mers, global physicochemical properties) to illustrate the diversity of peptide sequences and their biological functions. The right "encoding" strategy is key to maximize the performance of ML models [6]. Dimensionality reduction techniques allow the rapid visualization of one or more peptide space(s). In Figure 2, we mapped 56 physicochemical descriptors used to predict the hemolytic activity of antimicrobial peptides and reduced to a bidimensional t-distributed Stochastic Neighbouring Embedding (t-SNE) manifold [7]. On the left side of the figure, we displayed the training set Hemo-PI-1 (N=1,104) and on the right, the testing set Antimicrobial Peptide Database (APD, N=3,081). Our features informed the model to distinguish between hemolytic (purple) and non-hemolytic sequences (orange). Nearly 70% of APD were predicted as hemolytic peptides.

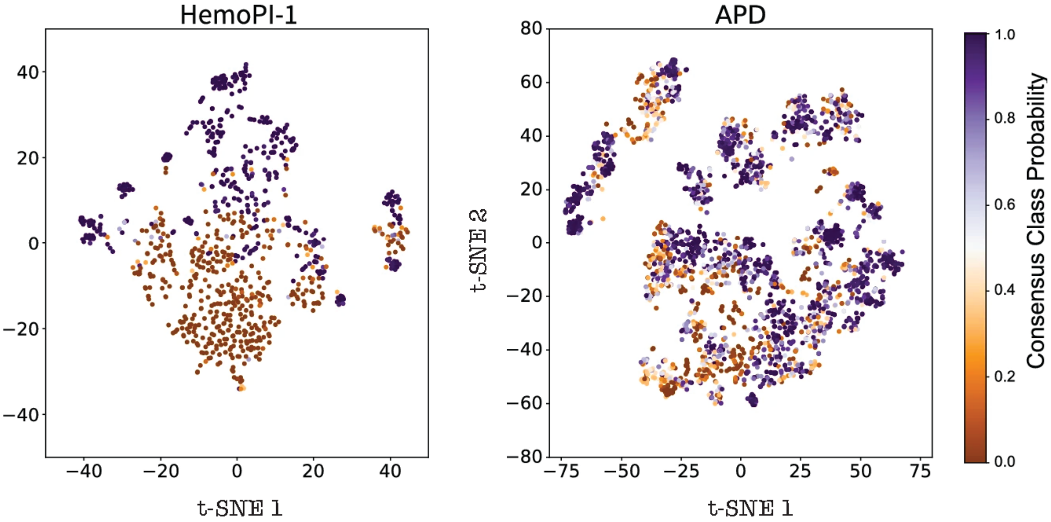

Fig. 2. T-SNE projections showing the distributions of HemoPI-1 training model (left) and APD testing dataset (right) with their respective consensus class probabilities, extracted from [7].

Mitigate imbalance and algorithmic bias. AI-powered peptide models are predominantly built from primary sequences. Sequence homology and structural information are often undermined. Taxonomic bias in AMP predictive models was recently brought to our attention [8] whereas the AMP structural folds remain mostly unknown – only 2.5% have been characterized experimentally (*i.e.*, X-ray or nuclear magnetic resonance) [9]. It is therefore crucial to identify the most reliable, fast, and efficient peptide structure prediction method to estimate the peptide structural landscape. Here, we explored the structural landscape of 5,840 sequences from GRAMPA [10], the Giant Repository of AntiMicrobial Peptide Archive using the Peptide Secondary Structure Prediction method PEP2D [11] for its ability to measure the 3 secondary states (H:Helix, E:Strand or Sheet and C:Coil) over a medium-large dataset of sequences. In Figure 3, we mapped the structural landscape of GRAMPA and its subsets using the 3 secondary states in terms of percentage in a ternary plot. Each face (or axis) of the ternary plot indicates the content of a peptide sequence in a single secondary state. The global AMP structure is represented with a dot at the intersection of the three axes. In Figure 3 (left), we summarized GRAMPA structural landscape where the dataset appeared as predominantly made of alpha-helical structures – highest density between 40 and 80% in helical content. On the right side of the figure, we mapped two GRAMPA subsets with reported antimicrobial activity against Gram-negative bacteria *Escherichia coli*. The top ternary plot shows that the majority of 4,540 sequences are likely to adopt an alpha-helical structure and occupies the same structural landscape than the entire GRAMPA dataset. The bottom ternary plot illustrates the structural landscape of 3,367 sequences used by Dean and co-workers to train their model against *E. coli*.[12]. In their study, the authors voluntarily excluded cysteine- and proline-rich sequences from the original dataset, which led to a training set entirely made of alpha-helical and coiled structures.

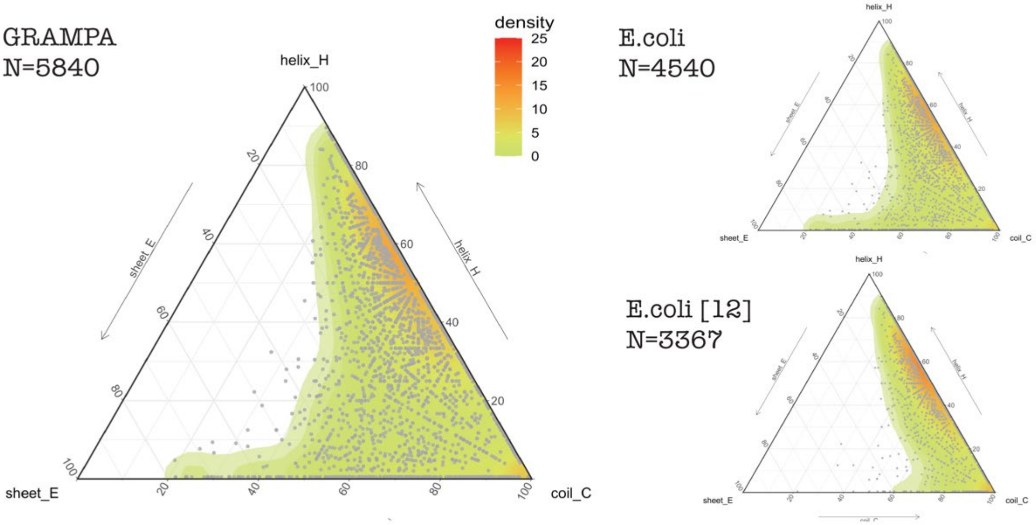

Fig. 3. Ternary plots showing the structural landscapes of three AMP datasets, i.e., GRAMPA, GRAMPA E.coli subset, E.coli subset[12]. Each landscape is mapped according to the calculated contents (%)) in the three possible secondary structures, namely, helices (H), strands or sheets (E), or coils (C) for each peptide sequence.

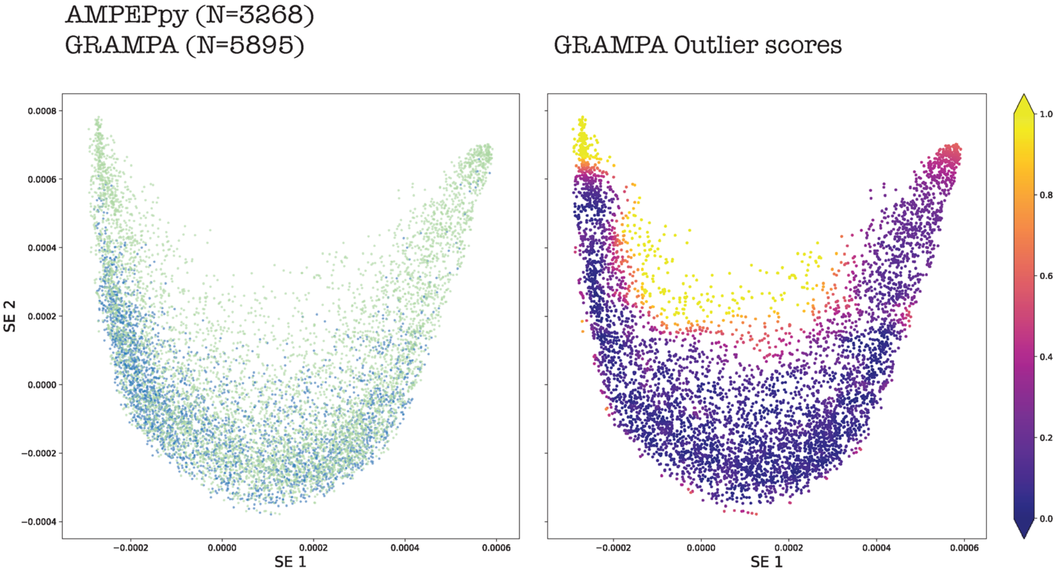

Fig. 4. Bidimensional spectral embeddings (SE) of two datasets; the positive training set AMPEPpy (N=3,268, blue) and the testing set GRAMPA (N=5,895, green). On the right side, Local Outlier Factor method best detected GRAMPA outliers shown in yellow (outlier score gradient).

Consider limits of model predictability. Limitations in peptide predictive and generative modeling lie in the diversity of peptide sequences and biological information. Detecting the boundaries of the applicability domain (sequence space where predictions are considered reliable) is key to build robust AI models. It implies identifying the peptide sequences "within-distribution" (inliers) and "out-of-distribution" (outliers). We co-ranked several dimensionality reduction techniques including principal component analysis, t-SNE and spectral embedding. In Figure 4 (left), we described two datasets – positive training set AMPEPpy (N=3,268, blue) [13] and GRAMPA (N=5,895, green) after reducing 83.9 millions of pairwise Smith-Watermann distances into bidimensional spectral embeddings (SE). We implemented multivariate outlier detection methods with Python module PyOD [14] to the reduced peptide sequence space, and we associated an outlier score (gradient) to each sequence. In Figure 4 (right), we displayed the results after applying the Local Outlier Factor method that best detected GRAMPA outliers shown in yellow.

Acknowledgments

We thank the Mexican research council *Consejo Nacional de Ciencia y Tecnología* (CONACyT). F.P. was supported by a Cátedras CONACyT fellowship (2017-2022).

References

1. Hancock, R. & Sahl, H.G. *Nature Biotech*. **24**, 1551-1557 (2006), https://doi.org/10.1038/nbt1267
2. Nguyen, L.T., Haney, E.F., & Vogel, H.J. *Trends in Biotechnology* **29** (9), 464-472 (2011), https://doi.org/10.1016/j.tibtech.2011.05.001
3. Haney, E.F., Straus, S.K., & Hancock, R. *Frontiers in chemistry* **7**, 43 (2019), https://doi.org/10.3389/fchem.2019.00043
4. Fjell, C., Hiss, J., Hancock, R., & Schneider, G. *Nature Reviews Drug Discovery* **11**, 37-51 (2012), https://doi.org/10.1038/nrd3591
5. Melo, M.C.R., Maasch, J.R.M.A., & de la Fuente-Nunez, C. *Communications Biology* **4**, 1050 (2021), https://doi.org/10.1038/s42003-021-02586-0
6. Erjavac, I., Kalafatovic, D., & Mauša, G. *Artificial Intelligence in Life Sciences* **2**, 100034 (2022), https://doi.org/10.1016/j.ailsci.2022.100034
7. Plisson, F., Ramírez-Sánchez, O., & Martínez-Hernández, C. *Scientific Reports* **10**, 16581 (2020), https://doi.org/10.1038/s41598-020-73644-6
8. Rázai, Z., Kiss, J., & Nagy, N.A. *Scientific Reports* **11**, 17924 (2021), https://doi.org/10.1038/s41598-021-97415-z
9. Van Oort, C.M., Ferrell, J.B., Remington, J.M., Wshah, S., & Li, J. *Journal of Chemical Information and Modelling* **61**(5), 2198-2207 (2021), https://doi.org/10.1021/acs.jcim.0c01441
10. Witten, J. & Witten, Z. *bioRxiv* 692681 (2019), https://doi.org/10.1101/692681
11. Singh, H., Singh, S., & Raghava G.P.S. *bioRxiv* 558791 (2019), https://doi.org/10.1101/558791
12. Dean, S.N., Alvarez, J.A.E., Zabetakis, D., Walper, S.A., & Malanoski, A.P. *Frontiers in Microbiology* **12**, 725727 (2021), https://doi.org/10.3389/fmicb.2021.725727
13. Lawrence, T.J., Carper, D.L., Spangler, M.K., Carrell, A.A., Rush, T.A., Minter, S., Weston, D.J., & Labbé, J.L. *Bioinformatics (Oxford, England)* **37**(14), 2058-2060 (2021), https://doi.org/10.1093/bioinformatics/btaa917
14. Zhao, Y., Nasrullah, Z., & Li, Z. *Journal of Machine Learning Research* **20**(96), 1-7 (2019), https://jmlr.org/papers/v20/19-011.html