



Teaching AI the Language of Life

*Ravi^{1,2}, Anu Sharma¹, Ashish Gupta^{1,2}, Lal Dhari Patel^{1,2} and Prabhakar Verma^{1,2}

¹ICAR- Indian Agricultural Statistics Research Institute, New Delhi-110012, India

²The Graduate School, ICAR-Indian Agricultural Research Institute,
New Delhi, 110012, India

*Corresponding Author's email: ravitendu7390@gmail.com

Imagine screening thousands of tiny protein fragments for potential antimicrobial, blood-pressure-lowering or antioxidant activity without testing every candidate in the laboratory. Artificial intelligence can help researchers shortlist promising peptides from their amino-acid sequences. One example is UniDL4BioPep (Du et al., 2023), a deep-learning model that uses a protein language model to represent peptide sequences. Here, its published results are compared with results from conventional machine-learning classifiers.

10	3	320	8 M
benchmark	conventional rivals: RF,	numbers describe	parameters in the
bioactivity datasets	SVM, XGBoost	each peptide	ESM-2 model used

Small chains, big benefits

Bioactive peptides (BPs) are short strings of amino acids, typically 2–20 long. They are often released from food proteins during digestion, enzymatic hydrolysis or fermentation, and can act as antimicrobial, antioxidant, anti-inflammatory or antihypertensive agents (Du and Li, 2022). No wonder functional foods, nutraceuticals, pharmaceuticals and cosmeceuticals all want them: BP research has grown about threefold in the past decade, and thousands of peptides now sit in public databases.

From test tubes to code

The traditional route (protein pretreatment, hydrolysis, fractionation and purification, then in vitro or in vivo testing) is costly, slow and dependent on advanced instruments and skilled staff (Ulug et al., 2021). Computational QSAR modelling helped: collect peptide data, convert sequences into numerical features, then build a model. But picking the right representation, features, algorithm and settings is itself laborious. Artificial intelligence (AI) offers a way to learn complex, non-linear patterns directly from data.

Reading proteins like sentences

Protein sequences resemble human language: amino-acid “letters” form “words” and “sentences”. Pretrained protein language models such as ESM (Luo et al., 2024), UniRep and ProtTrans (Fenoy et al., 2022) have learned from billions of sequences and can describe a protein from its sequence alone. UniDL4BioPep (Du et al., 2023) uses the smallest ESM-2 model, esm2_t6_8M_UR50D (6 layers, 8 million parameters), to turn every peptide into a fixed list of 320 numbers, called an embedding. Because the length never changes, one convolutional neural network (CNN) can scan these numbers for local patterns and answer a simple question: active or not? The same architecture works for any bioactivity, hence “universal”.

The showdown

To explore how conventional machine-learning methods compare with UniDL4BioPep (Du et al., 2023), Random Forest (RF), Support Vector Machine (SVM) and XGBoost were evaluated on ten peptide bioactivity datasets. Their performance was assessed using accuracy, balanced accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC) and area under the receiver operating characteristic curve (AUC). The UniDL4BioPep results were taken from Du et al. (2023), while the conventional-classifier results were generated in the present analysis.

The verdict: UniDL4BioPep (Du et al., 2023) showed strong performance across several peptide bioactivity prediction tasks, although the best-performing model varied by dataset. For bitter peptides, UniDL4BioPep achieved 93.8% accuracy and an MCC of 0.875, compared with 81% accuracy and an MCC of 0.62 for XGBoost in the present analysis. The difference was also evident for tumour T-cell antigen prediction: UniDL4BioPep (Du et al., 2023) achieved an AUC of 0.788 and an MCC of 0.457, whereas the conventional classifiers recorded AUC values of 0.51–0.63 and MCC values of 0.03–0.15.

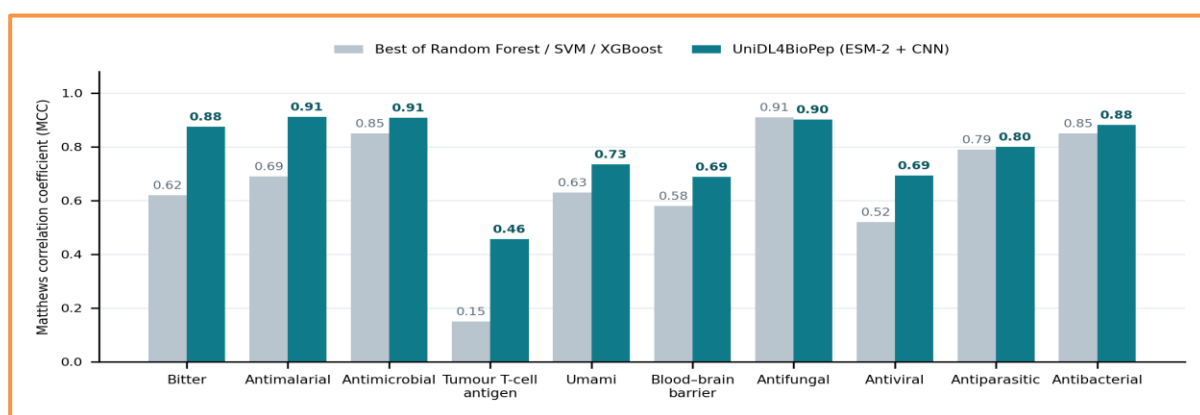


Figure 1. Comparison of Matthews correlation coefficient (MCC) values for UniDL4BioPep and conventional machine-learning classifiers (Random Forest, Support Vector Machine and XGBoost) across ten peptide bioactivity datasets. UniDL4BioPep values were obtained from Du et al. (2023), whereas conventional machine-learning results were generated in the present analysis.

A lesson in imbalance: The antimalarial training dataset contained 111 positive and 1,708 negative examples. In the present analysis, Random Forest achieved 97% accuracy but a balanced accuracy of 0.75, while SVM achieved a balanced accuracy of 0.50. These results illustrate why accuracy alone can be misleading when one class greatly outnumbers the other. UniDL4BioPep achieved an MCC of 0.815 on the main antimalarial dataset, while its focal-loss variant achieved an MCC of 0.793. On the alternative dataset, the corresponding MCC values were 0.912 and 0.957, respectively (Du et al., 2023). These results highlight the importance of distinguishing between datasets and model configurations when interpreting performance.

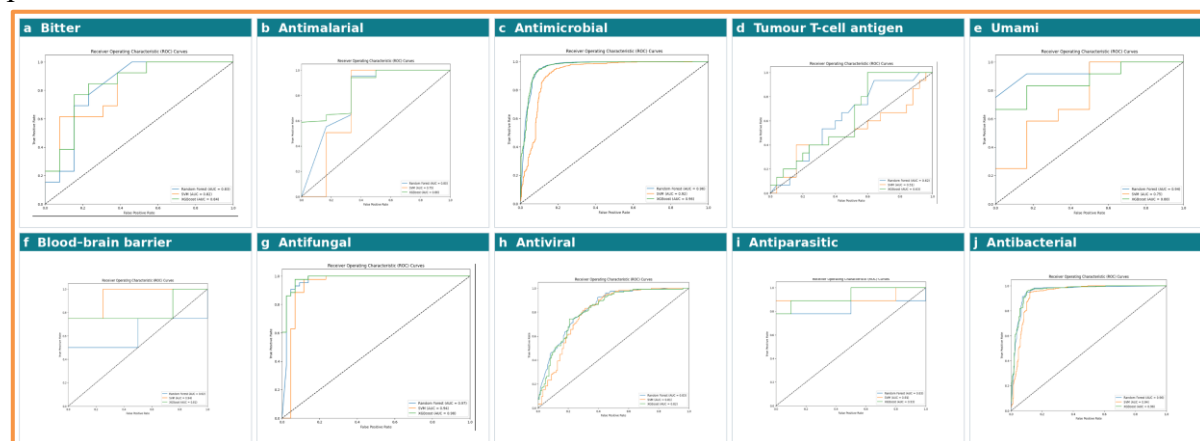


Figure 2. Receiver operating characteristic (ROC) curves for Random Forest, Support Vector Machine and XGBoost across ten peptide bioactivity datasets, generated from the present analysis.

Why it matters

AI-based peptide prediction can help researchers shortlist promising candidates before laboratory testing, potentially reducing the time and cost of experimental screening. UniDL4BioPep (Du et al., 2023) illustrates how protein language models can support the prediction of diverse peptide bioactivities. However, conventional machine-learning classifiers may perform competitively on particular tasks, and performance depends on the dataset and evaluation procedure. Computational predictions should therefore be treated as a way to prioritise candidates, not as proof of biological activity; experimental validation remains essential.

References

1. Du, Z., Ding, X., Xu, Y., and Li, Y. (2023). UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. *Briefings in Bioinformatics*, 24(3), bbad135.
2. Du, Z., and Li, Y. (2022). Review and perspective on bioactive peptides: A roadmap for research, development, and future opportunities. *Journal of Agriculture and Food Research*, 9, 100353.
3. Fenoy, E., Edera, A. A., & Stegmayer, G. (2022). Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. *Briefings in Bioinformatics*, 23(4), bbac232.
4. Luo, Z., Wang, R., Sun, Y., Liu, J., Chen, Z., & Zhang, Y. J. (2024). Interpretable feature extraction and dimensionality reduction in ESM2 for protein localization prediction. *Briefings in Bioinformatics*, 25(2), bbad534.
5. Ulug, S. K., Jahandideh, F., and Wu, J. (2021). Novel technologies for the production of bioactive peptides. *Trends in Food Science and Technology*, 108, 27-39.