DOI: 10.25205/978-5-4437-1843-9-6

PREDICTION CYTOKINE-INDUCING EPITOPES USING PROTEIN LANGUAGE MODELS

A. D. Gavrilenko, M. A. Sindeeva, S. V. Shashkova

Autonomous Non-Profit Organization "Artificial Intelligence Research Institute", Moscow

✉ Gavrilenko@airi.net

Abstract

We present a machine learning approach for classifying epitopes by their ability to induce specific cytokines (IL-2, IL-4, IL 10), using contextual embeddings from protein language models and ensemble gradient boosting classifiers. Our method achieves high predictive performance, particularly for IL-10.

In this study, we propose a data-driven method for binary classification of T cell epitopes based on their ability to induce the secretion of specific cytokines: IL-2, IL-4, or IL-10. The central idea is to leverage contextual embeddings generated by state-of-the-art protein language models (ESM Cambrian (ESM C) [1] and Ankh [2]) as features for training gradient boosting ensembles (LightGBM, CatBoost) within the LightAutoML [3] framework.

Experimentally validated human epitopes were extracted from the Immune Epitope Database (IEDB, February 2025) [4]. For each cytokine, balanced datasets were constructed using SMOTE [5] or ADASYN [6] oversampling techniques, addressing class imbalance in positive vs. negative examples.

As baselines, we implemented classifiers using handcrafted sequence-based descriptors: Amino Acid Composition (AAC), Dipeptide Composition (DPC), and Amino Acid Pair (AAP) antigenicity scores. These served as a reference to quantify the added value of deep contextual representations.

Models were evaluated using standard metrics: precision, recall, AUC, and Matthews Correlation Coefficient (MCC). For IL-10 prediction, our best model based on ESM C embeddings achieved AUC = 0.918 and MCC = 0.641 (see table), outperforming previous methods such as IL10Pred [7]. Comparable or superior results were also obtained for IL-2, while IL-4 remained more challenging due to its higher sequence diversity.

To explore this variability, we calculated sequence-level statistics across cytokine classes, including Shannon entropy, dipeptide density, and average Levenshtein distance. IL-4-inducing epitopes displayed the highest variability, while IL-10 epitopes were more conserved—explaining the observed differences in classification accuracy.

In summary, protein language models enable efficient, generalizable, and scalable prediction of cytokine-inducing epitopes directly from amino acid sequences. This approach is especially useful for in silico immunogenicity discovery and prioritization of candidate peptides for experimental validation.

Performance of ESM C model for cytokine-inducing epitope classification

Cytokine	Precision	Recall	MCC	AUC
IL-2	0.776	0.819	0.564	0.855
IL-4	0.717	0.833	0.433	0.795
IL-10	0.727	0.796	0.641	0.918

References

1. ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. 2024.

2. Elnaggar A. et al. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. 2023.

3. Vakhrushev A. et al. LightAutoML: AutoML Solution for a Large Financial Services Ecosystem // arXiv:2109.01528. 2021. URL: https://doi.org/10.48550/arXiv.2109.01528.

4. Vita R. et al. The Immune Epitope Database (IEDB): 2024 update // Nucleic Acids Research. 2025. Vol. 53, No. D1. P. 436–443. URL: https://doi.org/10.1093/nar/gkae1092.

5. Chawla N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique // J. Artif. Intell. Res. 2002. Vol. 16. P. 321–357. URL: https://doi.org/10.1613/jair.953.

6. He H., Garcia E. A. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering. 2009. Vol. 21, No. 9. P. 1263–1284.

7. Nagpal G. et al. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. Scientific Reports. 2017. Vol. 7. P. 42851. URL: https://doi.org/10.1038/srep42851.