Universidad La Salle

Prediction of peptide MHC binding and presentation using transformers and transfer learning in cancer immunology context

MSc. Vicente Machaca Arceda

2023

Content

Content

Immunotherapy for Cancer

Personalized Vaccines

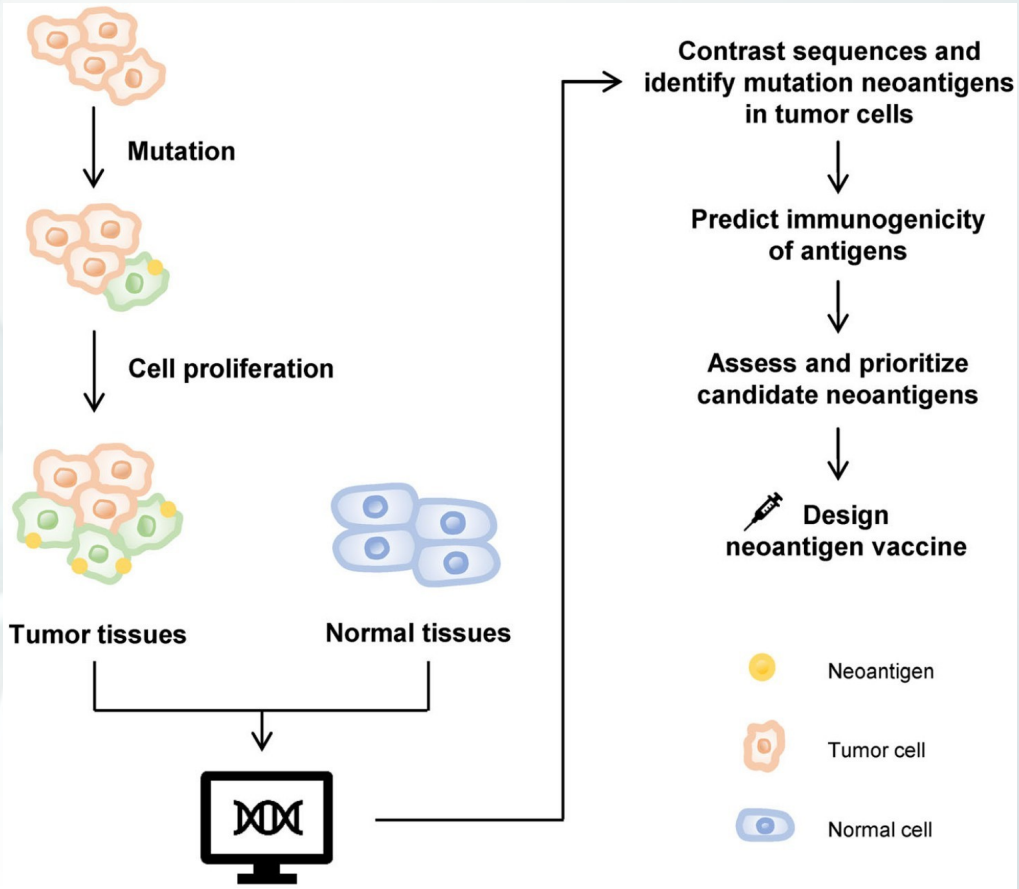

Figure: Personalized vaccines process for Cancer [1].

pMHC binding and presentation prediction

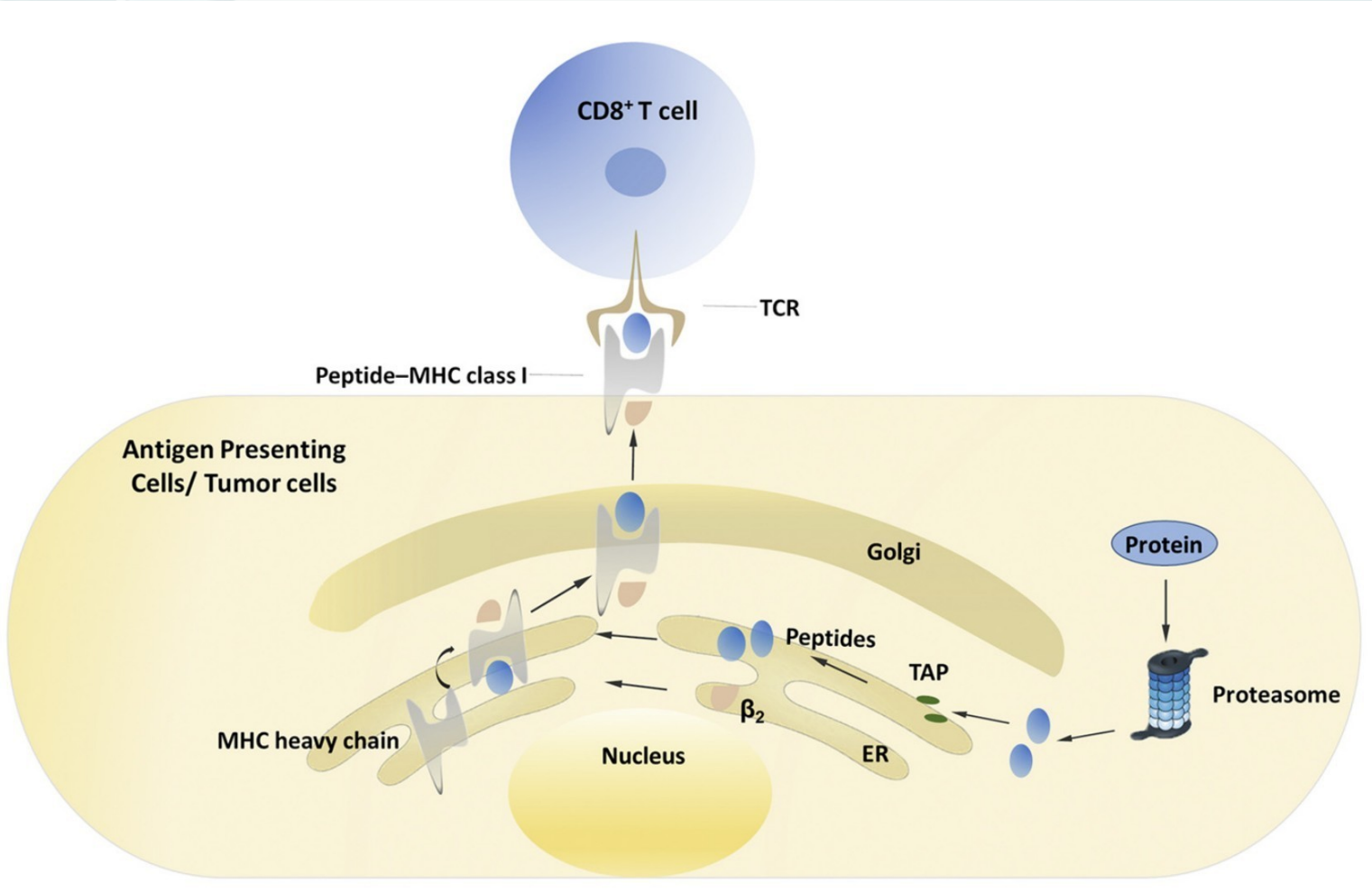

Figure: pMHC presentation process in MHC class I [2].

Content

Problem

Less than 5% of detected neoantigens (peptides binded to MHC) succeed in activating the immune system [3]. Moreover, recent proposals only achieve 0.6 precision and 0.4 recall [4].

This is a **binary classification problem**. A peptide could be represented like: $p = \{A, ..., Q\}$ and a MHC like: $q = \{A, N, ..., Q, E\}$. Finally, we need to know the probability of affinity between p and q (pMHC)

Problem

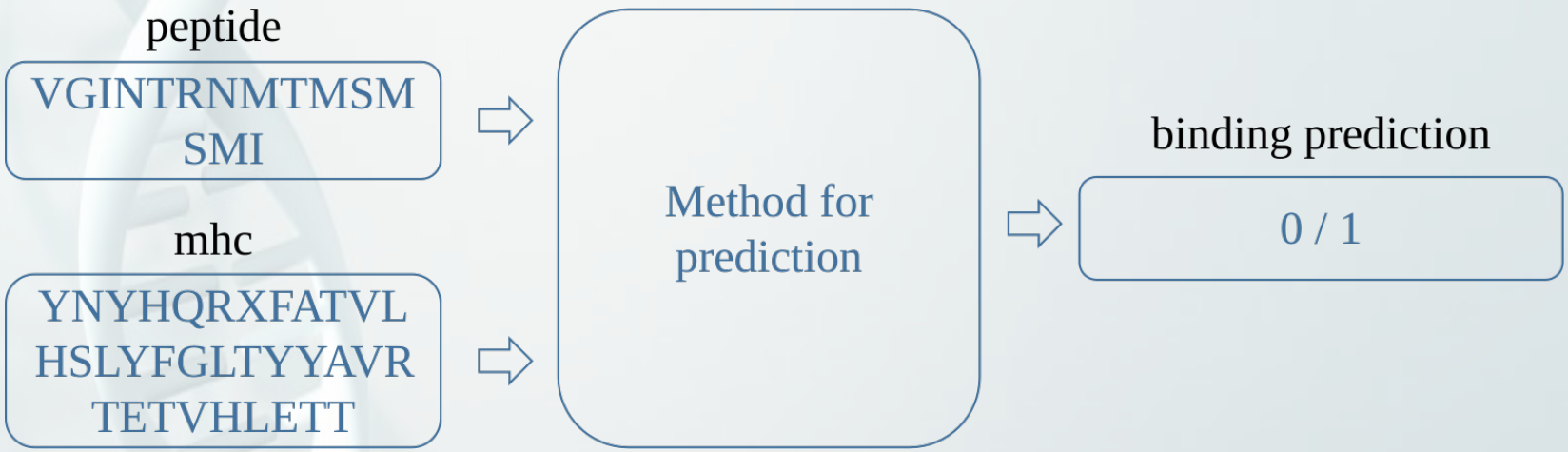

Figure: pMHC binding prediction problem.

Related Works
Transformers

Table: Recent works based on transformers and transfer learning.

Year	Ref.	Name	Method
2022	[5]	**HLAB**	Uses protBert model incascade with a RNN with attention
2022	[6]	MHCRoBERTa	Five encoders with 12 multiple-head self-attention pre-trainned with self-supervision
2022	[7]	**TransPHLA**	Based on four modules: an embedding block, an encoder block (multiple self-attention), a feature optimization block (FC layer), and a projection block (FC layer used to predict)
2021	[8]	BERTMHC	Uses TAPE model followed by a linear layer.
2021	[9]	ImmunoBERT	The same as BERTMHC focused on MHC-class I

Proposal

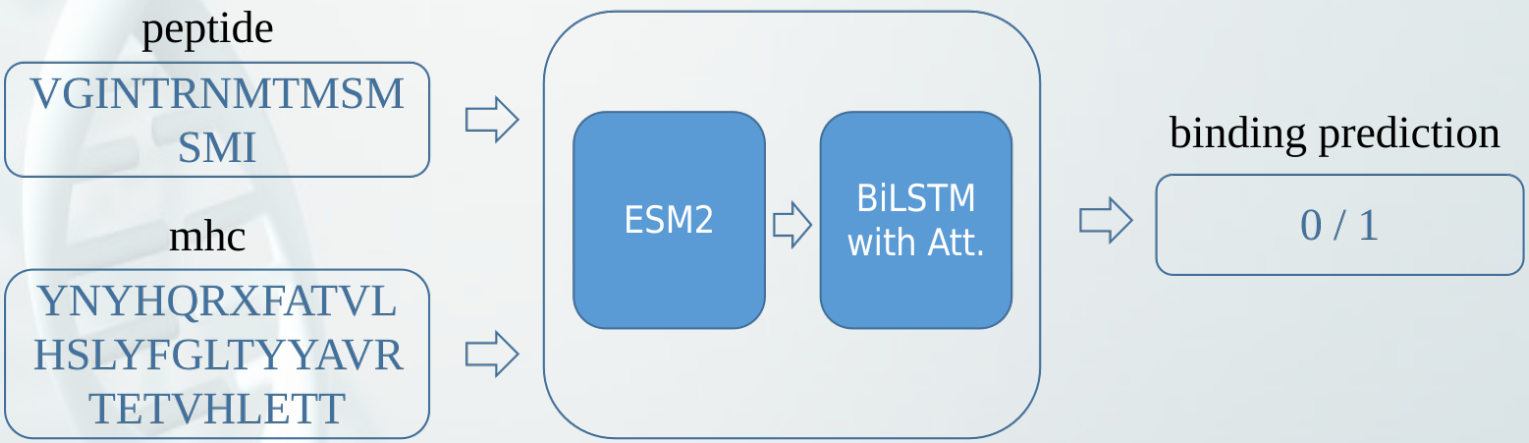

Figure: Proposal for pMHC binding and presentation prediction.

Content

Databases

We used the dataset from NetMHCIIpan3.2 [10].

Table: Samples used in training, evaluation and testing.

	Samples
Train	107424
Validation	13428
Testing	13429

Models

Instead of ESM2 [11] model, we used TAPE [12] because it is smaller and easier to train. Moreover, the Bi-LSTM with attention layer is based on HLAB [5].

Table: Models used in experiments.

	Description
BERTMHC-LINEAR	BERT architecture followed by a linear layer
BERTMHC-RNN	BERT architecture followed by a BiLSTM layer and then a Linear layer
BERTMHC-RNN-ATT	BERT architecture followed by a BiLSTM layer with attention and then a Linear layer

Content

Training

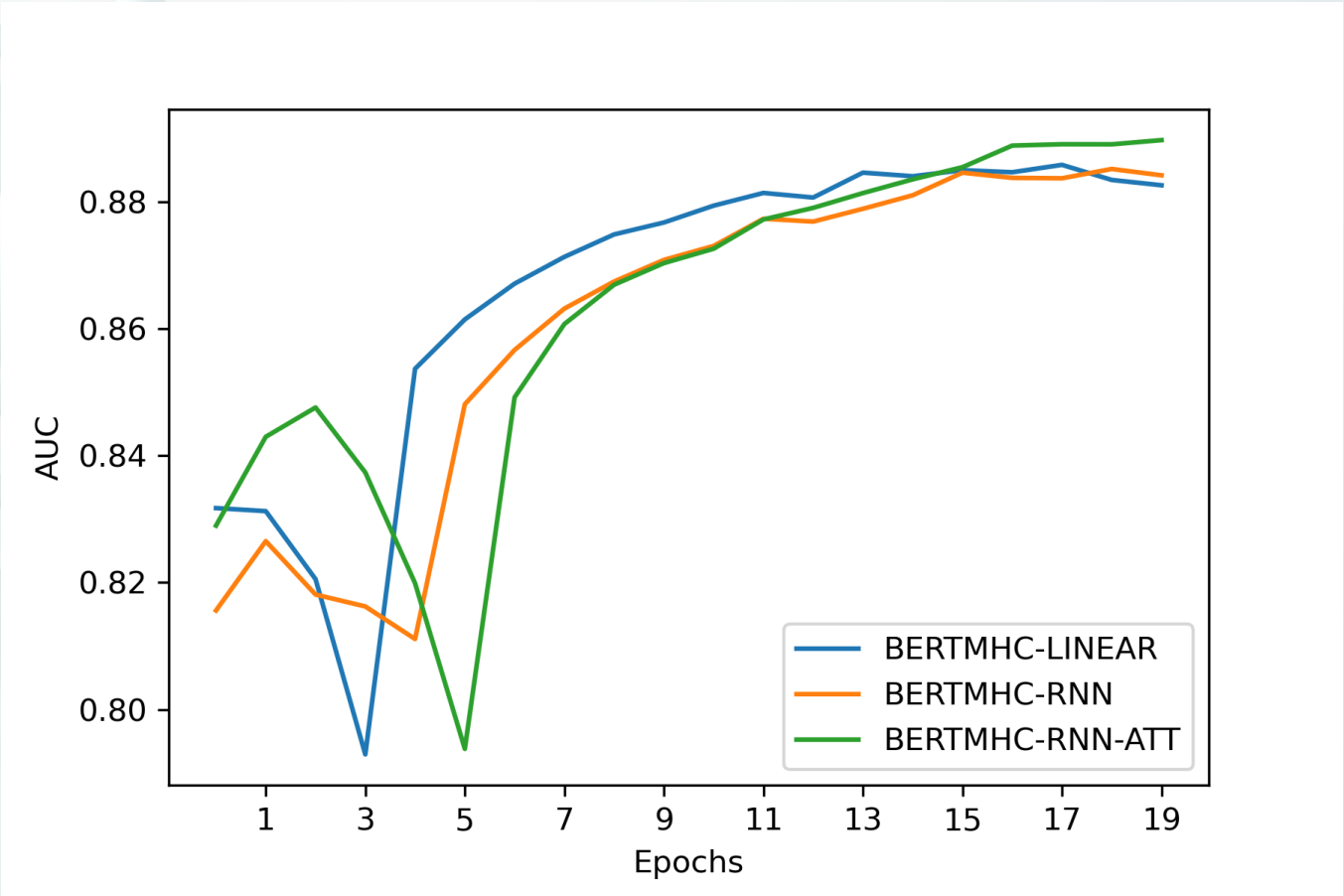

Figure: AUC per epoch of models.

Comparison

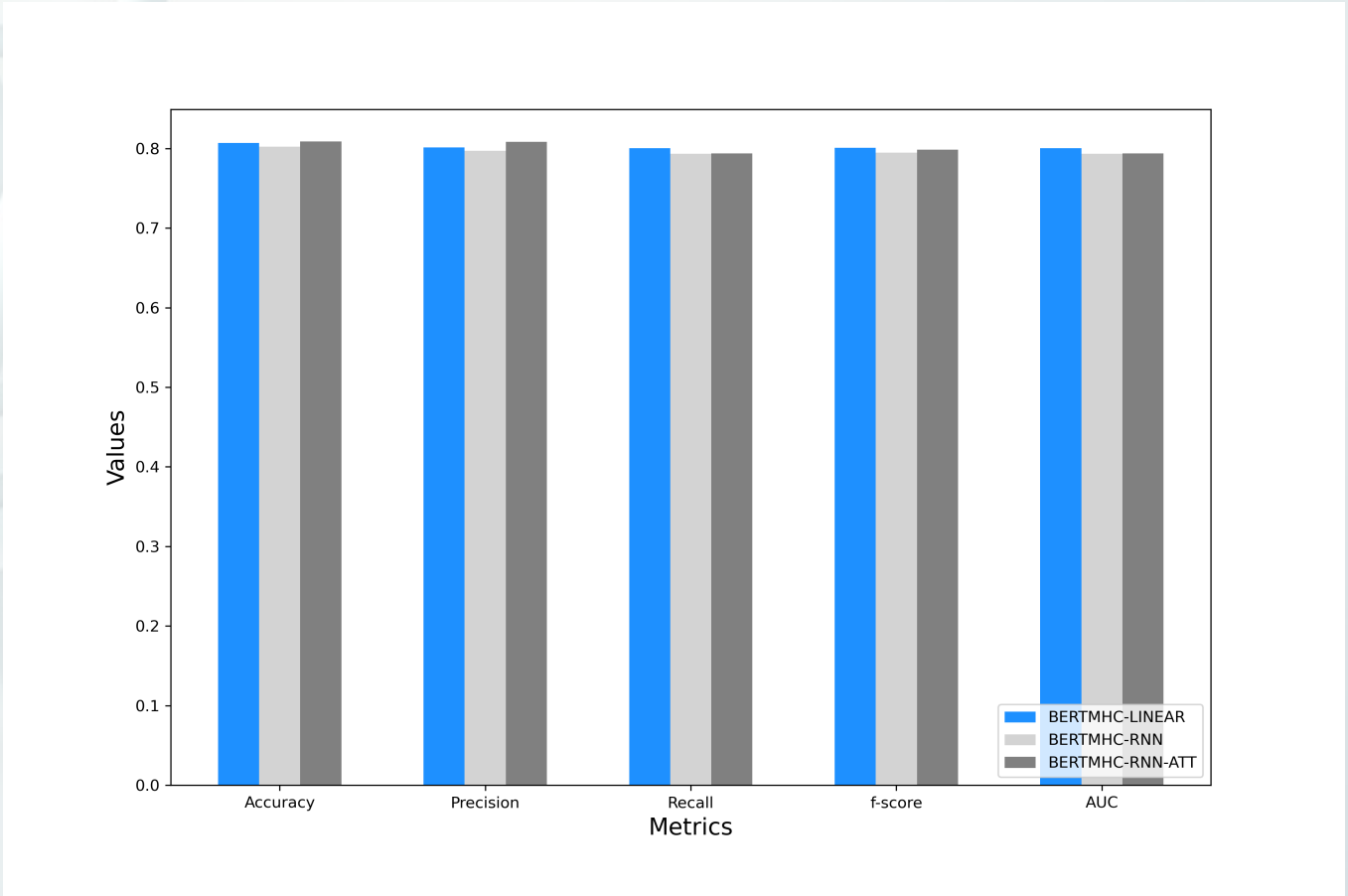

Figure: Metrics comparison.

Comparison

Table: Metrics comparison of BERTMHC-LINEAR, BERTMHC-RNN and BERTMHC-RNN-ATT

Model	Acc	Precision	Recall	Fscore	AUC
LINEAR	0.8070	0.8012	**0.8005**	**0.8009**	**0.8005**
RNN	0.8023	0.7972	0.7932	0.7949	0.7932
RNN-ATT	**0.8086**	**0.8082**	0.7937	0.7985	0.7937

Comparison

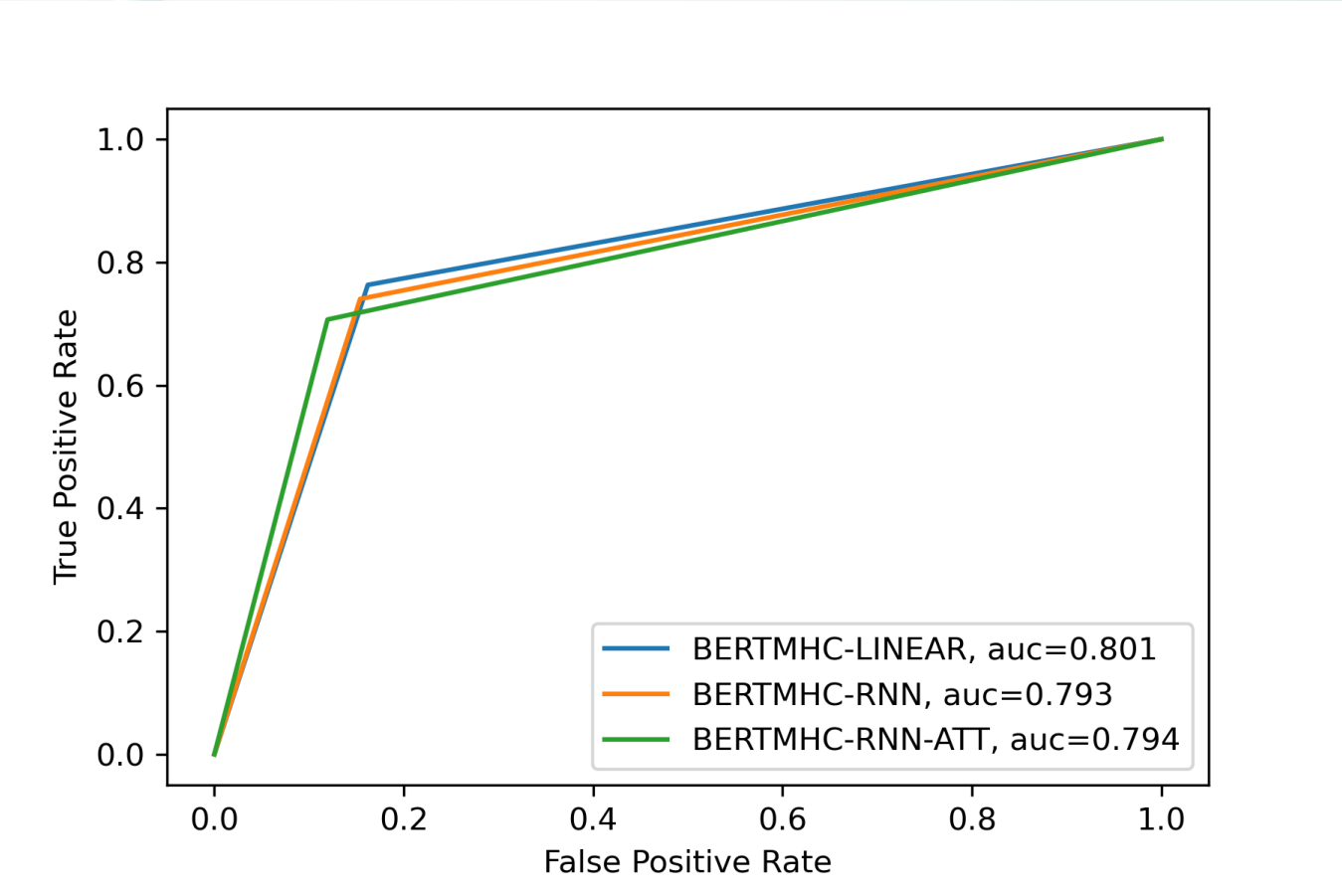

Figure: ROC curve.

Conclusions

In this preliminary results, we evaluated a BERT architecture (transformer) with transfer learning from TAPE. We choose TAPE because, it is smaller and easier to train. In future experiments, we will evaluate ESM2.

According to experiments, BERTMHC-LINEAR and BERTMHC-RNN-ATT got better results in netMHCIIpan3.2 dataset. This happens, because we evaluated these models in a small dataset. In future experiments, we will evaluated these models in a larger dataset.

References I

[1] M. Peng, Y. Mo, Y. Wang, P. Wu, Y. Zhang, F. Xiong, C. Guo, X. Wu, Y. Li, X. Li *et al.*, "Neoantigen vaccine: an emerging tumor immunotherapy," *Molecular cancer*, vol. 18, no. 1, pp. 1–14, 2019.

[2] X. Zhang, Y. Qi, Q. Zhang, and W. Liu, "Application of mass spectrometry-based mhc immunopeptidome profiling in neoantigen identification for tumor immunotherapy," *Biomedicine & Pharmacotherapy*, vol. 120, p. 109542, 2019.

[3] L. Mattos, M. Vazquez, F. Finotello, R. Lepore, E. Porta, J. Hundal, P. Amengual-Rigo, C. Ng, A. Valencia, J. Carrillo *et al.*, "Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the esmo precision medicine working group," *Annals of oncology*, vol. 31, no. 8, pp. 978–990, 2020.

References II

[4] N. A. Mill, C. Bogaert, W. van Criekinge, and B. Fant, “neoms: Attention-based prediction of mhc-i epitope presentation,” *bioRxiv*, 2022.

[5] Y. Zhang, G. Zhu, K. Li, F. Li, L. Huang, M. Duan, and F. Zhou, “Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction,” *Briefings in Bioinformatics*, 2022.

[6] F. Wang, H. Wang, L. Wang, H. Lu, S. Qiu, T. Zang, X. Zhang, and Y. Hu, “Mhcroberta: pan-specific peptide–mhc class i binding prediction through transfer learning with label-agnostic protein sequences,” *Briefings in Bioinformatics*, vol. 23, no. 3, p. bbab595, 2022.

References III

[7] Y. Chu, Y. Zhang, Q. Wang, L. Zhang, X. Wang, Y. Wang, D. R. Salahub, Q. Xu, J. Wang, X. Jiang *et al.*, "A transformer-based model to predict peptide–hla class i binding and optimize mutated peptides for vaccine design," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 300–311, 2022.

[8] J. Cheng, K. Bendjama, K. Rittner, and B. Malone, "Bertmhc: improved mhc–peptide class ii interaction prediction with transformer and multiple instance learning," *Bioinformatics*, vol. 37, no. 22, pp. 4172–4179, 2021.

[9] H.-C. Gasser, G. Bedran, B. Ren, D. Goodlett, J. Alfaro, and A. Rajan, "Interpreting bert architecture predictions for peptide presentation by mhc class i proteins," *arXiv preprint arXiv:2111.07137*, 2021.

References IV

[10] K. K. Jensen, M. Andreatta, P. Marcatili, S. Buus, J. A. Greenbaum, Z. Yan, A. Sette, B. Peters, and M. Nielsen, "Improved methods for predicting peptide binding affinity to mhc class ii molecules," *Immunology*, vol. 154, no. 3, pp. 394–406, 2018.

[11] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[12] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, and Y. Song, "Evaluating protein transfer learning with tape," *Advances in neural information processing systems*, vol. 32, 2019.