Language modeling of peptide-HLA interactions achieves state-of-the-art performance on prediction of peptide presentation by HLA Class II

Abstract 5364

Daniel Sprague, Joshua Klein, Italo Faria do Valle, Olivia Petrillo, Matthew Davis, Monica Lane, Karin Jooss, Ankur Dhanik

Background

- Epitope prediction for Gritstone's personalized cancer vaccines (PCV) has relied upon HLA Class I epitope presentation prediction using our EDGE^{TM} platform.
- CD4+ T cells likely augment CD8+ T cell responses, and non-specific CD4 epitopes are currently deployed in the Gritstone PCV.
- Combination of tumor neoantigen-specific Class II and Class I epitopes may augment PCV immunogenicity and efficacy.
- Current models for epitope prediction perform better for Class I than Class II – a superior model for the latter is needed.
- Here, we introduce a new addition to EDGE: a state-of-the-art model for the prediction of the presentation of peptides by HLA Class II.

Approach

- EDGE^{TM} leverages the pretrained protein sequence embeddings from the "Evolutionary Scale Model" (ESM2) language model.
- A learned genotype network ("LGN") is used to aggregate embeddings from all alleles prior to prediction, rather than after.
- Immunoaffinity purified mass spectrometry data was used to refine EDGE's sequence predictions prior to training the LGN.

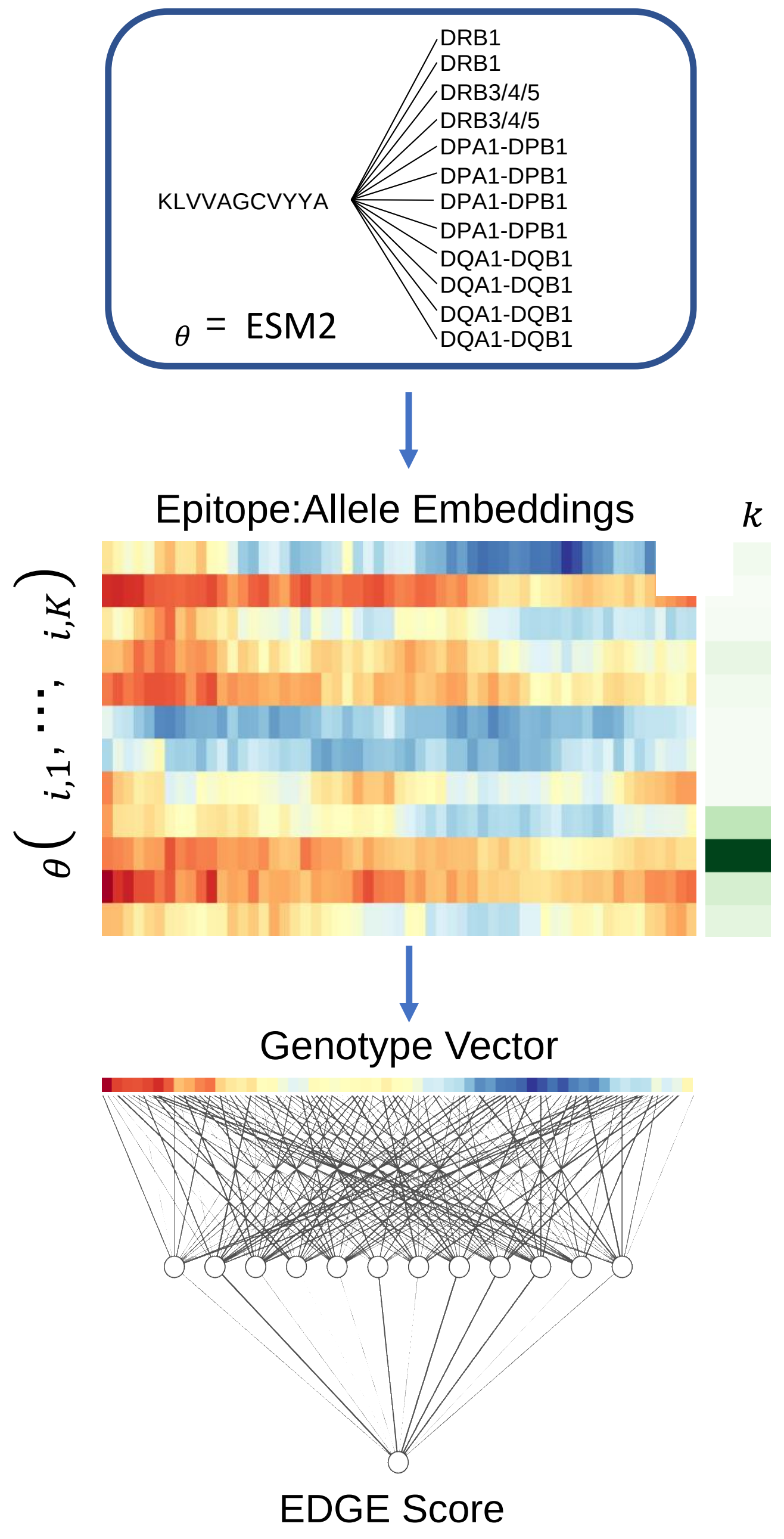
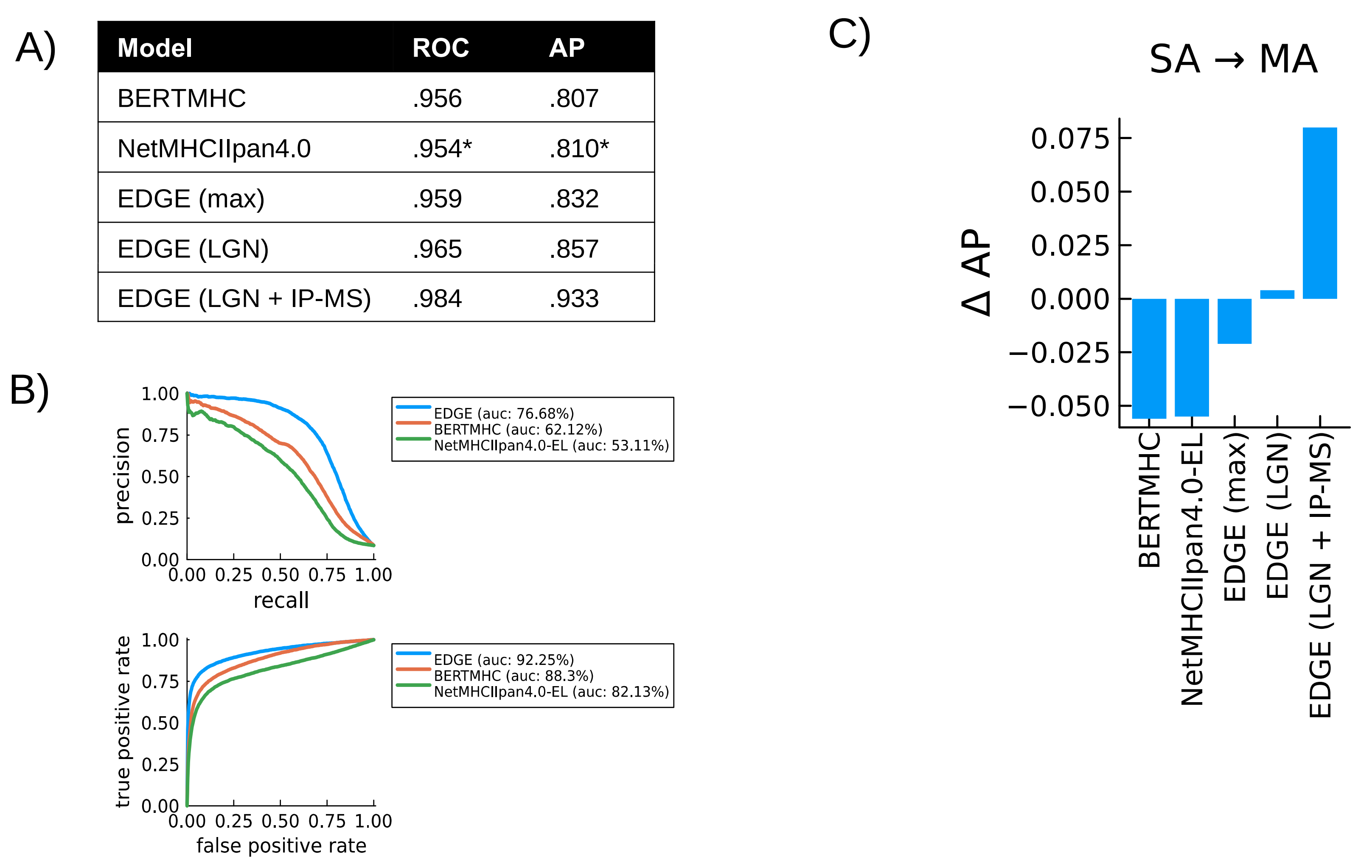


Figure 1. EDGE architecture for prediction of class II epitope presentation probabilities over full HLA class II genotypes.

Results

- EDGE trained and evaluated on the class II Reynisson et al. dataset substantially improves ROC-AUC and Average Precision/PPV (AP) over existing methods [3,4].



Model	ROC	AP
BERTMHC	.956	.807
NetMHCIIpan4.0	.954*	.810*
EDGE (max)	.959	.832
EDGE (LGN)	.965	.857
EDGE (LGN + IP-MS)	.984	.933

Figure 2. A) EDGE achieves superior performance to prior models and improves its performance on reference multi-allelic (MA) and single-allelic (SA) data. B) Substantial increase in AP on an independent test set of presentation data. C) LGN causes EDGE to substantially outperform other models on MA data.

- The LGN allows all alleles in a genotype to contribute to the gradient during backpropagation, as opposed to the max operator.
- Inclusion of the LGN in EDGE improves predictive power on multi-allelic data over selecting the highest scoring allele (max).
- EDGE does not simply learn the mean of all the alleles present in a sample's genotype but rather selectively identifies alleles that are likely to present.

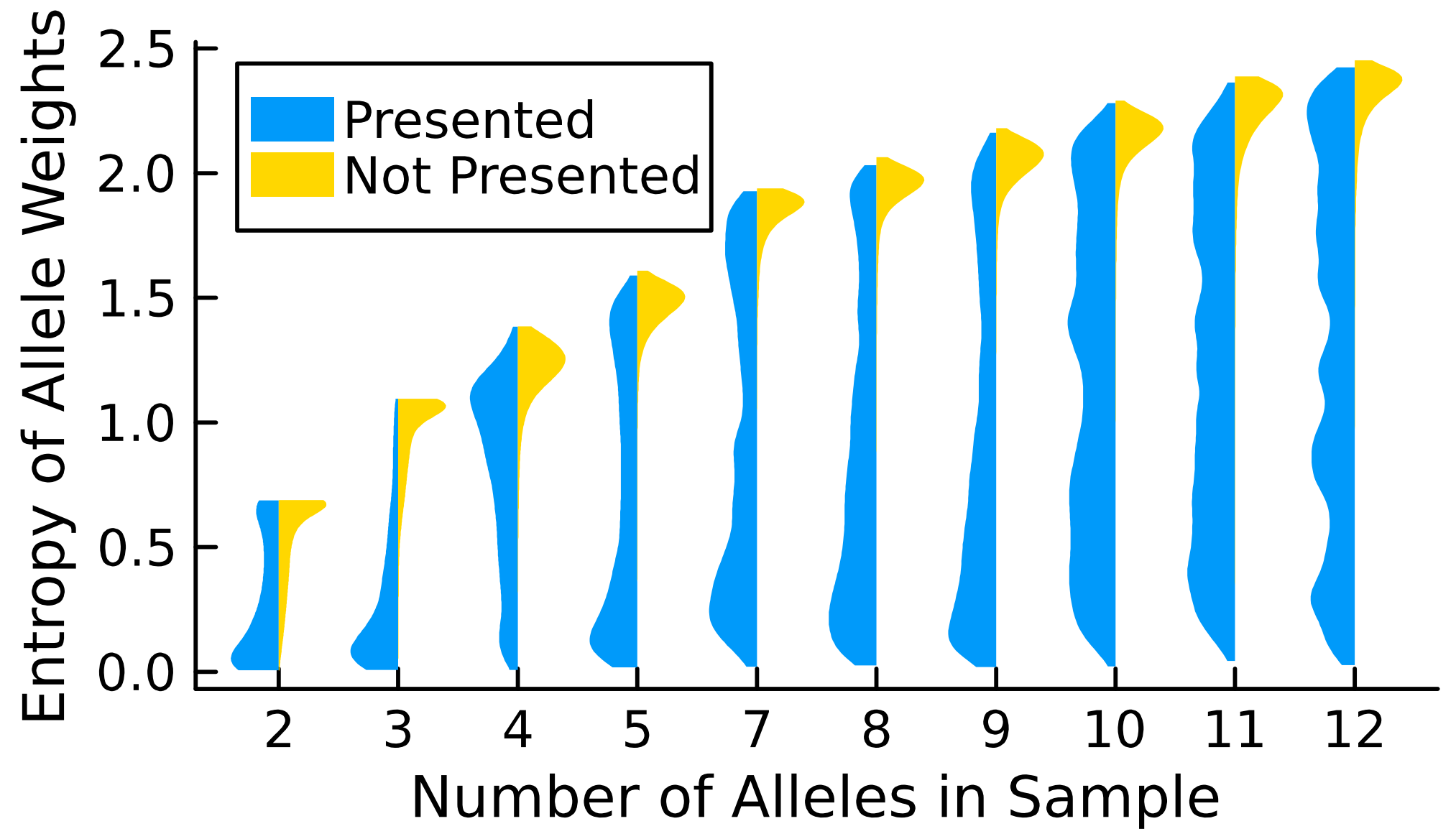


Figure 3. The LGN of EDGE places increased weight on individual alleles when an epitope is presented. Decrease in entropy in true positive samples indicates less uniform weights across alleles.

Results

- Publicly available data were collected from two personalized neoantigen vaccine studies with class II restricted epitopes and ELISPOT response data.
- EDGE score (logit) was significantly predictive of immunogenic response, particularly for scores greater than zero.

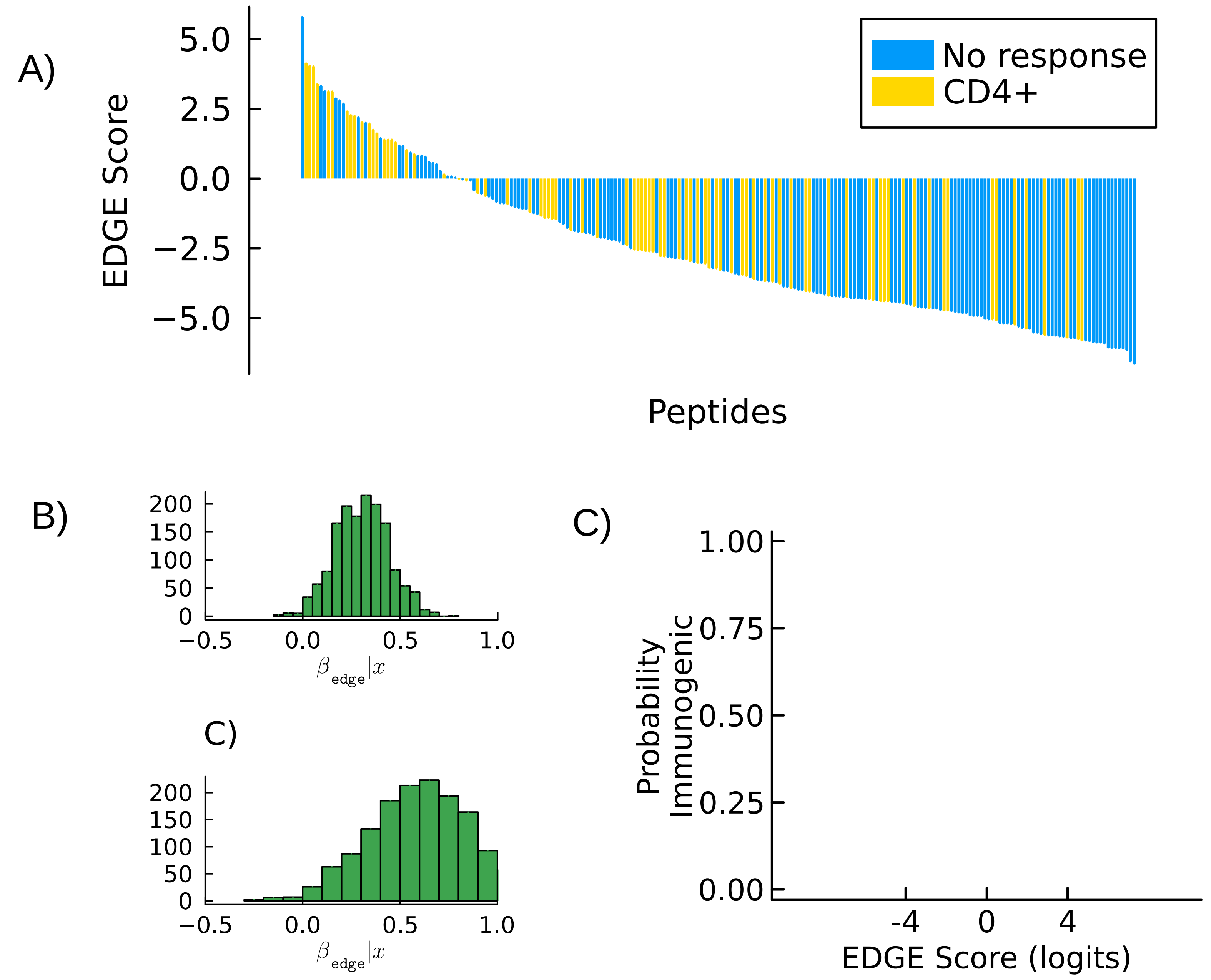


Figure 4. A) Peptides ranked by EDGE score and colored by their positive or negative CD4+ response. B) Posterior distribution of logistic coefficient demonstrates that EDGE is predictive of immunogenicity in personalized mRNA vaccines (upper), and this association is particularly strong for high scoring peptides (lower). C) Posterior predictive distribution indicates relatively low scoring peptides have immunogenic potential.

Conclusions

- The new addition to Gritstone's EDGE^{TM} platform significantly improves performance over prior HLA class II peptide presentation models.
- On independent data, prior methods would select Class II epitopes with Avg. PPV of 62%, whereas our new EDGE model would have Avg. PPV of 77%.
- In a PCV context, despite EDGE not being trained to predict immunogenicity (training in progress currently), EDGE predicts immunogenic class II epitopes with an Avg. PPV of 47%.
- This advance is likely to lead to superior PCV antigenic composition over current approaches.

All authors are employees of Gritstone bio

[1] Bulik-Sullivan, B. *et al.* Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nature Biotechnology 2018* (2018) doi:10.1038/nbt.4313

[2] Palmer, C. D. *et al.* Individualized, heterologous chimpanzee adenovirus and self-amplifying mRNA neoantigen vaccine for advanced metastatic solid tumors: phase 1 trial interim results. *Nat Med* **28**, 1619–1629 (2022)

[3] Cheng, J., Bendjama, K., Rittner, K. & Malone, B. BERTMHC: improved MHC–peptide class II interaction prediction with transformer and multiple instance learning. *Bioinformatics* **37**, 4172–4179 (2021)

[4] Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **48**, W449–W454 (2021)