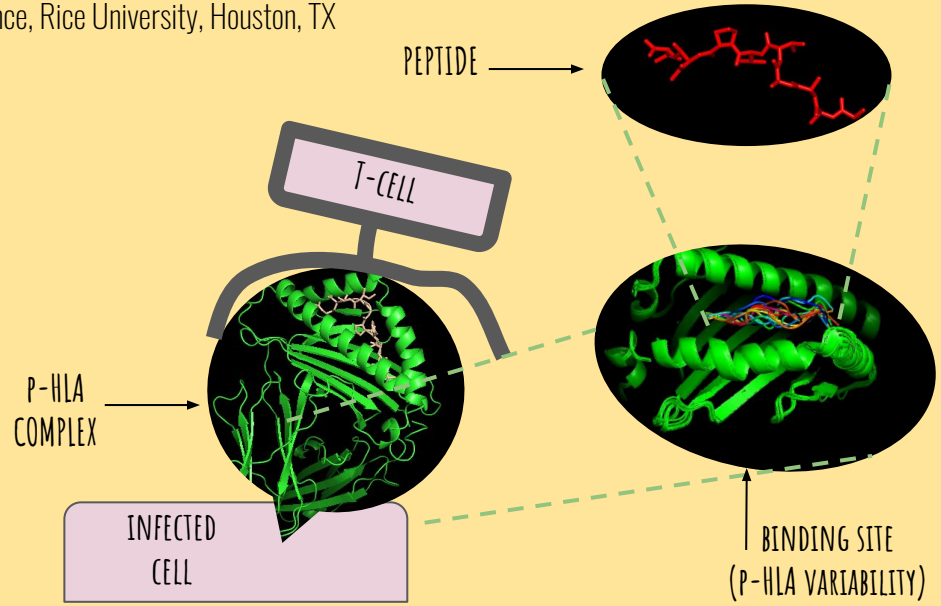
Combining protein structure and sequence data to predict peptide-HLA binding affinity

Conev A (1), Rigo M (1), Antunes DA (1), Fasoulis R (1), Hall-Swan S (1), Kavraki LE (1)

(1) Department of Computer Science, Rice University, Houston, TX

pHLA complex and its role in human immune response

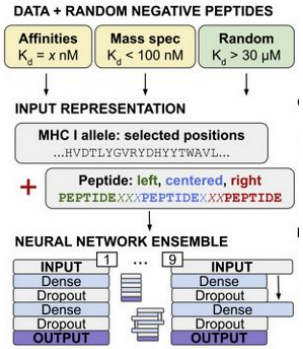
- ★ HLA class I is responsible for displaying small peptides for recognition by $CD8^+$ T-cell lymphocytes *inducing immune response*.
- ★ This pathway can be exploited to engineer a response (peptide / T-cell vaccines); *predicting peptide-HLA binding* is essential in designing the vaccines.
- ★ *High allele variability* makes this task challenging.
- ★ By providing reliable pHLA binding predictions *computational methods can speed up* the process.



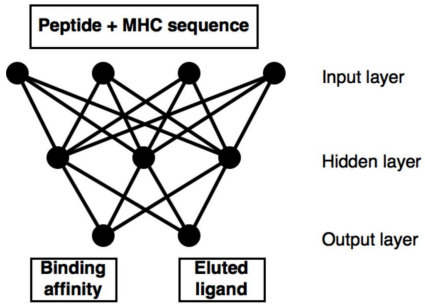
PROBLEM : identify a peptide that binds to the HLA allele and induces an immune response

pHLA binding prediction using sequence based methods

- ★ Neural network ensembles trained on the sequence information about peptide and HLA; Methods: MHCFlurry [1], NetMHC [2], NetMHCpan [3].
- ★ Abundant sequence data available.



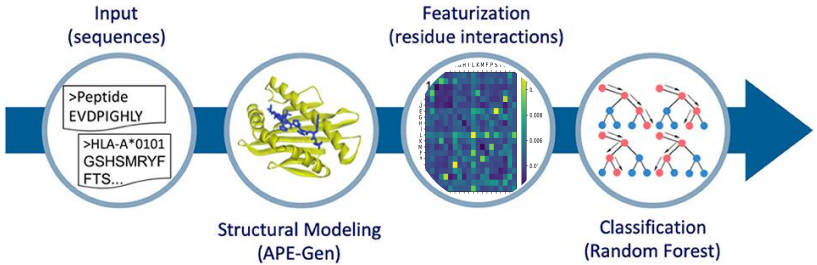
[O'Donnell et al, 2020]



[Jurtz et al, 2017]

pHLA binding prediction using structure based methods

- ★ Modelled structures (Ape-Gen[4]) -> structural features (AA distance matrix).
- ★ Random forest trained on the structural features [5].
- ★ Predicts binder/non-binder labels, not the affinity.



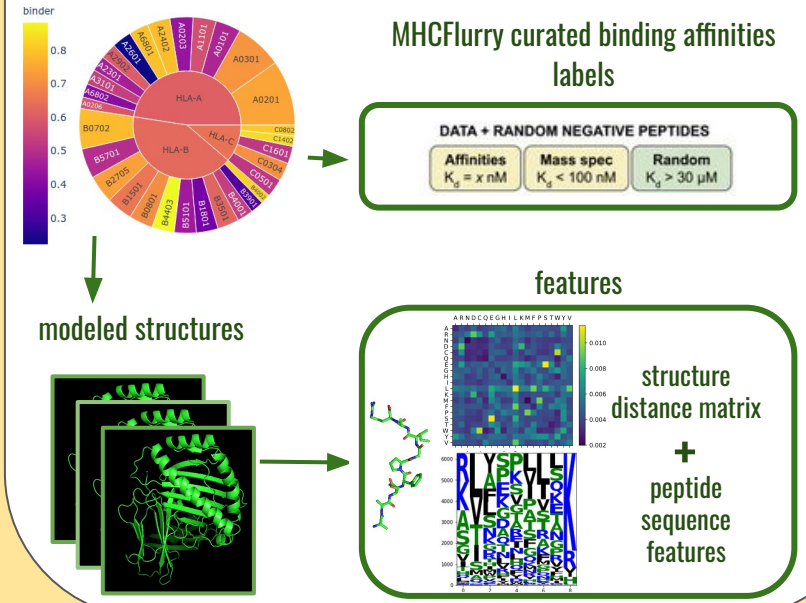
[Abella et al, 2020]

Combining available data for more accurate predictions of binding affinity

Dataset of modeled structures combined with affinity values

★ 82,000 pHLA modeled structures across 30 HLA alleles.

★ Mapped to binding affinity values from curated MHCFlurry dataset.



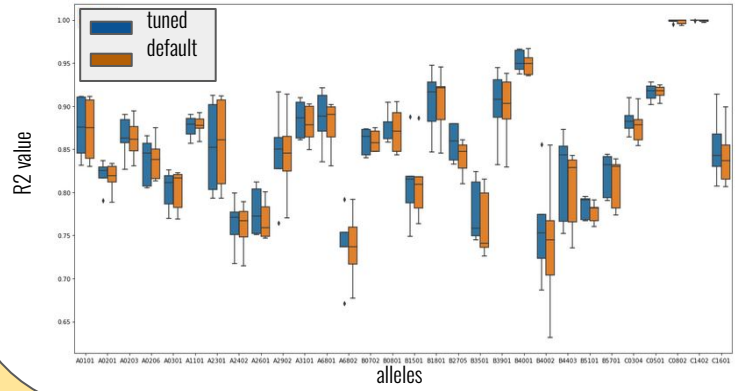
Random Forest regressor models

★ 30 per-allele random forest regressor models trained.

★ Test set: 20% of balanced data.

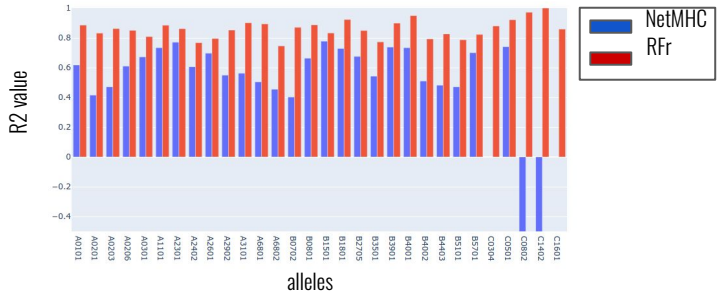
★ Parameter tuning: 5-fold cross validation.

★ Parameters tuned: *number of trees, maximum tree depth, minimum samples per leaf, minimum samples per split, bootstrapping.*

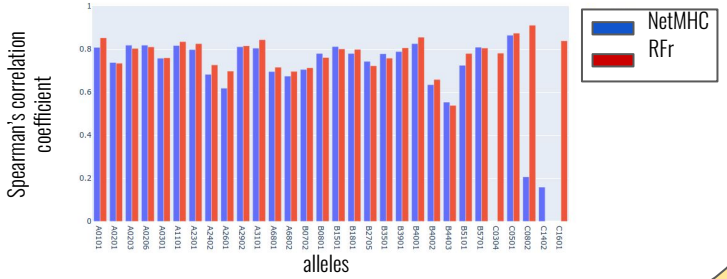


Results
-comparison with NetMHC4.0-

★ Capturing the variability of the data: R2 scores.



★ Correctly ranking the peptides: Spearman's correlation coefficient.



Conclusions

★ Sequence-based methods dominate the field of pHLA predictions; but still produce a lot of false-positive binders.

★ Structural data gives important insight into pHLA binding.

★ Our method deals with the lack of available structural by modeling the structures.

★ Our method shows that incorporating structural features can enhance the power of models predicting pHLA binding affinity.

Acknowledgments: *This work was funded by the National Science Foundation (NSF) (award number 2033262) and Rice University funds.*

RICE

Contact: ac121@rice.edu

References:

[1] O'Donnell TJ et al. 2018.. MHCflurry: open-source class I MHC binding affinity prediction. Cell Syst; 7:129–32.
[2] Lundegaard et al. 2008. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res; W509-12.
[3] Jurtz et al. 2017. NetMHCpan 4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J Immuno; 199(9)3360-3368.
[4] Abella et al. 2019. APE-Gen: A Fast Method for Generating Ensembles of Bound Peptide-MHC Conformations. Molecules; 24(5):881
[5] Abella et al. 2020. Large-Scale Structure-Based Prediction of Stable Peptide Binding to Class I HLAs Using Random Forests. Front Immuno; 11:1583.