Deep learning of binding interactions for peptide-MHC class I complex revealed specific amino acid properties decisive in peptide binding

JaeUng Hyun[1], Kwoneel Kim[1,2]
[1]Department of Biomedical and Pharmaceutical Sciences, Kyung Hee University, Seoul, Republic of Korea
[2]Department of Biology, Kyung Hee University, Seoul, Republic of Korea

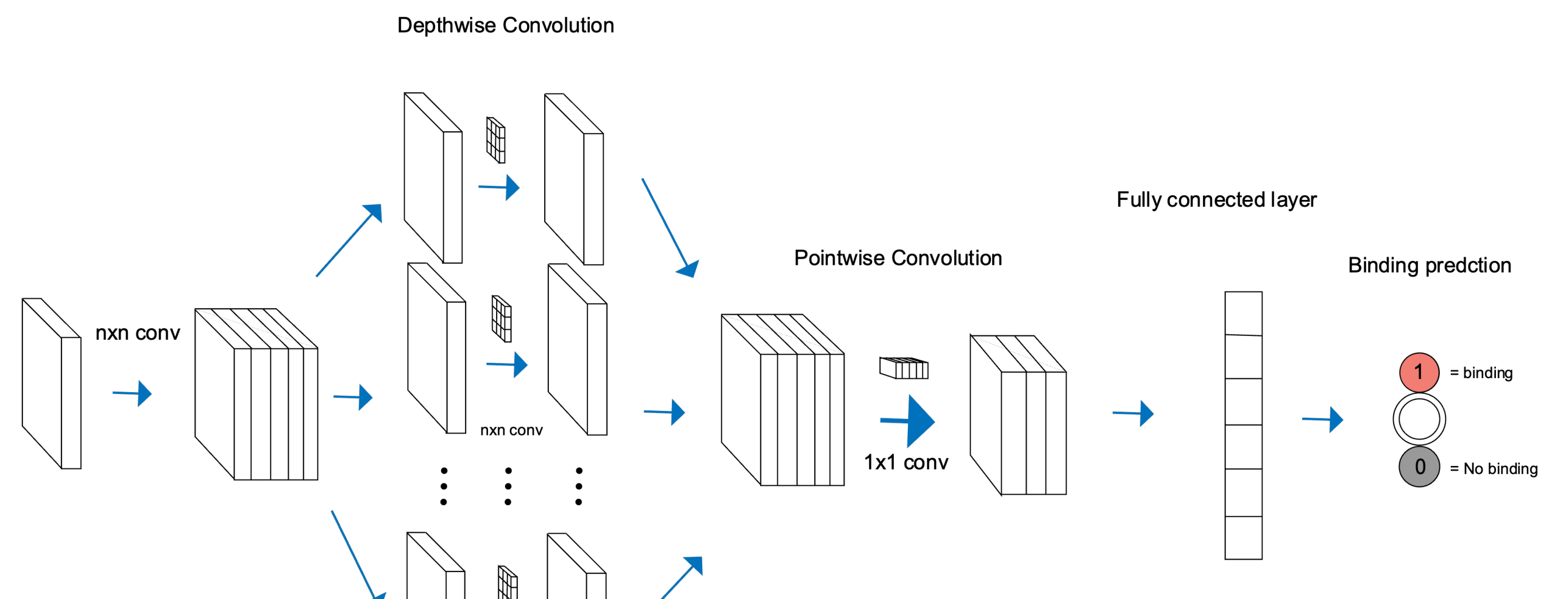
Introduction

The Major Histocompatibility Complex (MHC) plays crucial role in immune system of vertebrates. Major function of MHC is to recognize antigens derived from self-protein or pathogens by binding and presenting on cell surface for recognition by T-cells. The MHC genes are known that they are the most polymorphic over 19,500 distinct class I alleles and over 7300 distinct class II alleles as of April of 2020 [1]. Therefore, it is very difficult to predict the binding pattern of peptide-MHC molecule accurately. Recently, studies to predict peptide-MHC binding through neural network have been actively conducted [2-5, 10,11]. Most models trained based on peptide sequences have limitations in accurately predicting binding. This is because peptide sequence has only sequence information and no directive information such as interaction energy and physiochemical features [6]. In this study, we developed the model which is called DeepNeo-MHC. DeepNeo-MHC is trained by the amino acid interaction map consisted of physiochemical features among amino acids. Amino acids interaction map, in addition, is improved that was built in our previous work and changed CNN model to modern architecture called as EfficientNet structure [7, 8].

Method

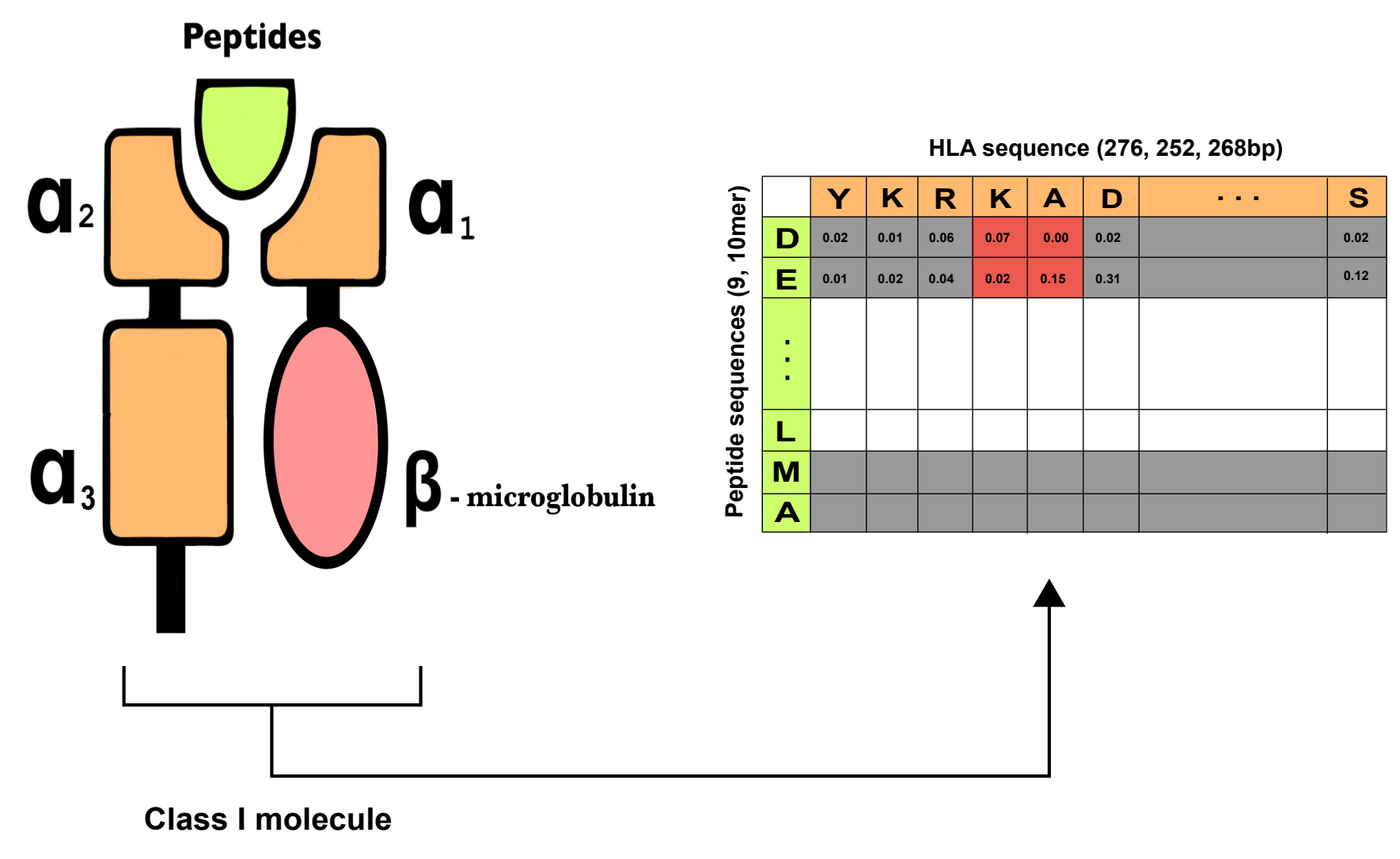
Improve convolutional neural network architecture

In last decade, performance of deep learning has improved significantly in computer vision. Especially, CNN (convolutional neural network) outperformed cognitive ability of human in ImageNet competition. LeNet we used in previous work was one of the earliest convolutional neural network. LeNet has limitation to extract features from interaction map because it has only 5 layers. Recent architectures normally have more dozens of layers and improved architecture, which is sufficient to extract features. Therefore, we chose EfficientNet structure which is introduced by Mingxing Tan et al. in 2019 [8]. For our data shape, input dimension was modified from 3 channels to 1 channel for class I data (See below figure), and 2 channels for class II data. Kernel size of each layer was also changed.



Modify amino acid interaction map

Interaction energy represent as a smaller value as the force of attraction and as a larger value as the force of the repulsive force acts [6]. We made interaction map with interaction energy between HLA sequence and peptide sequence to capture interactions (see below figure). The interaction energy value was converted to a positive value for smooth model training.



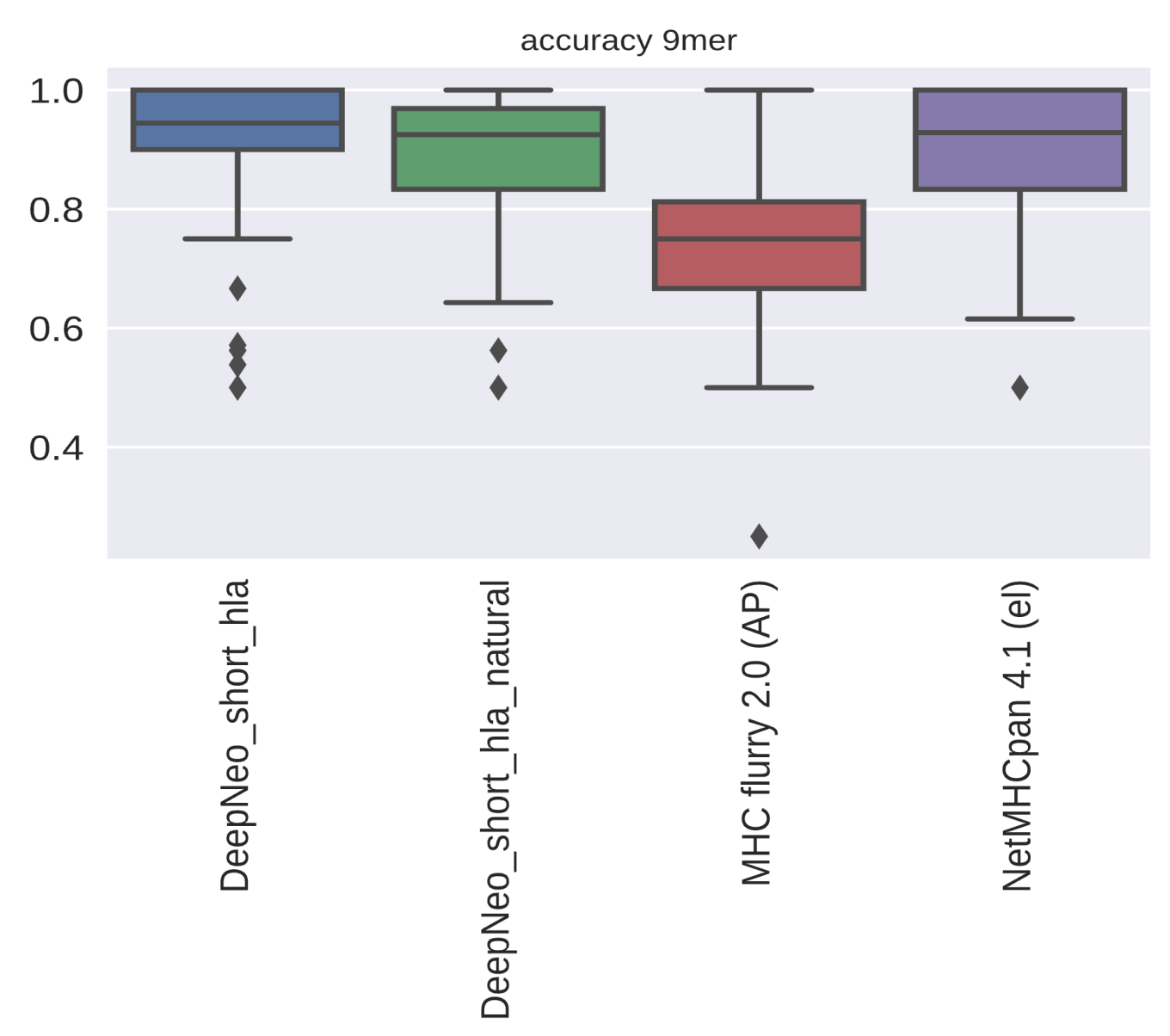
	Y	K	R	K	A	D	· · ·	S
D	0.02	0.01	0.06	0.07	0.00	0.02		0.02
E	0.01	0.02	0.04	0.02	0.15	0.31		0.12
⋮								
L								
M								
A								

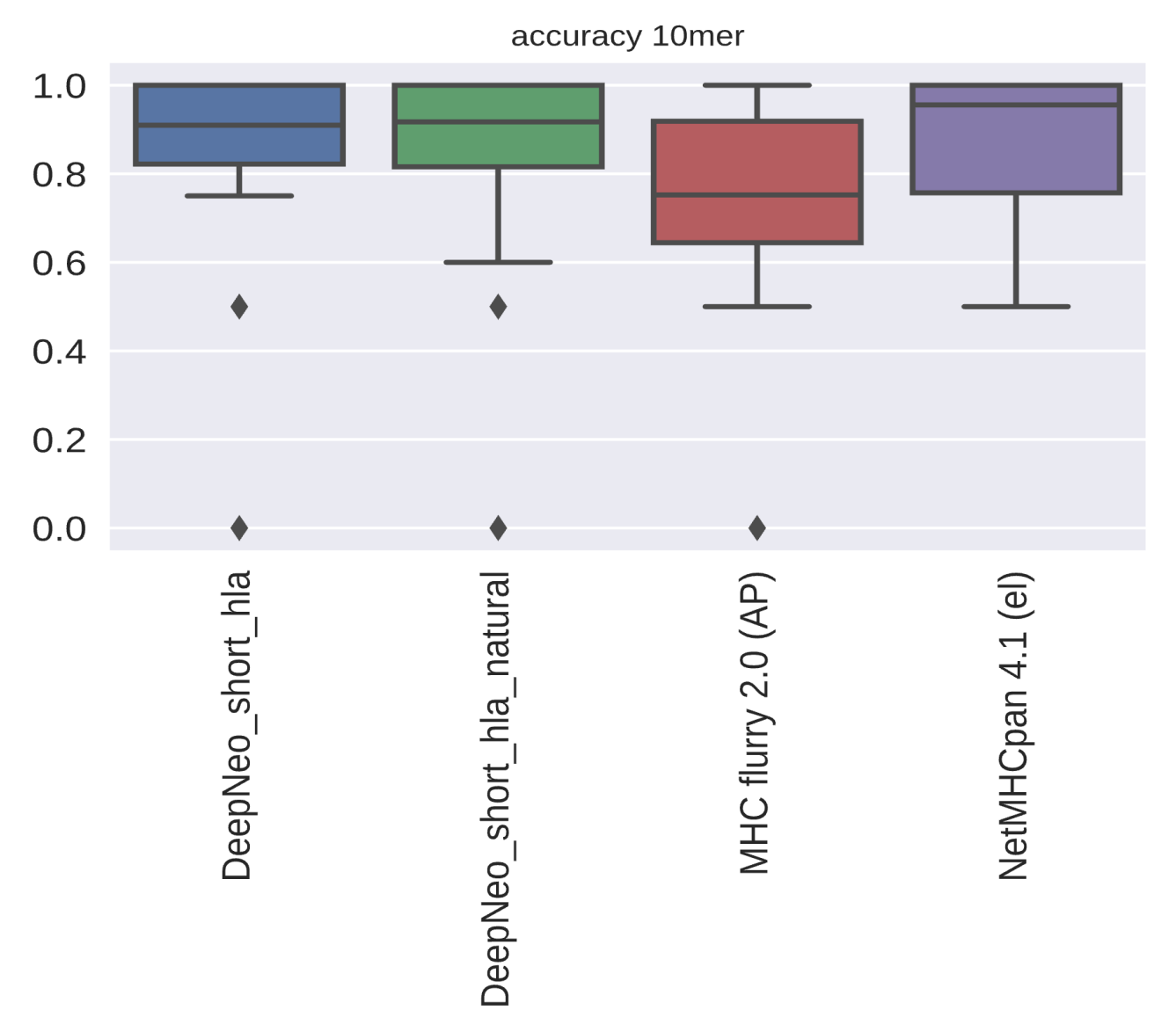
IEDB peptide-MHC binding datasets

All data was taken from the IEDB database (https://www.iedb.org, mhc_ligand_full(single_file.zip))[9]. This data were curated to MHC allele class = I, Assay Group = ('ligand presentation', 'half maximal inhibitory concentration (IC50)', 'qualitative binding', '3D structure', '50% dissociation temperature', 'half life'), Method/Technique = ('cellular MHC/mass spectrometry', 'purified MHC/competitive/radioactivity', 'binding assay', 'cellular MHC/direct/fluorescence', 'x-ray crystallography', 'purified MHC/direct/fluorescence', 'purified MHC/competitive/fluorescence') and length of Description = 9, 10. For nonbinding peptide data, we generated random peptides as many as binding peptides. For comparison, we also generate false set from natrual protein sequence. The number of curated dataset is 803844.

Benchmark on huge external datasets

For performacne checking, IEDB weekly benchmark dataset and 15% of curated dataset was used. This external dataset is consisted of 125088 peptides across 94 HLA alleles. For comparison, NetMHCpan 4.1 (EL) and MHCFlurry 2.0 (AP) were used together for performance measurement [10-11]. DeepNeo shows better performance than NetMHCpan 4.1 and MHCFlurry 2.0 (AP).

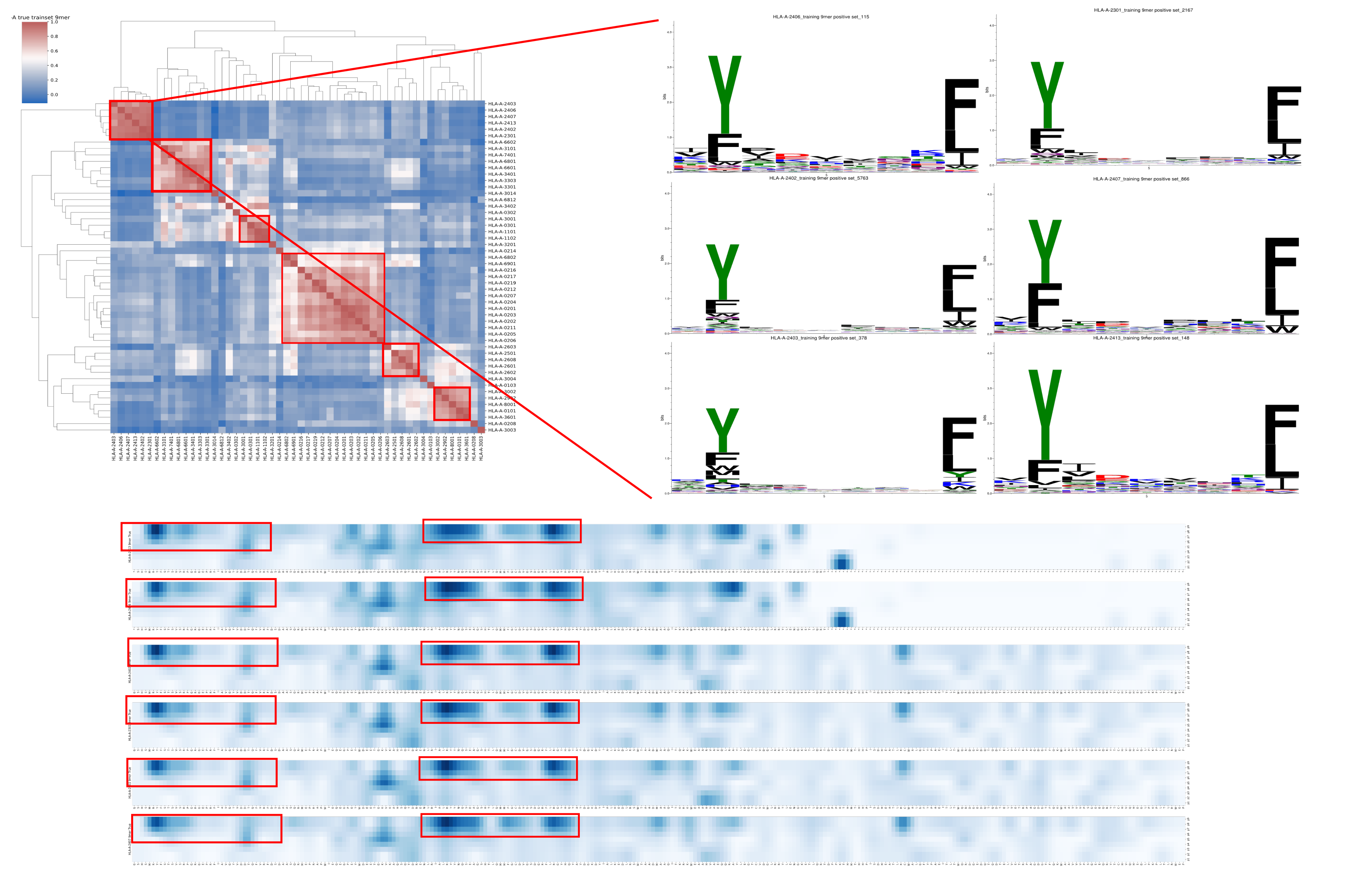




Result

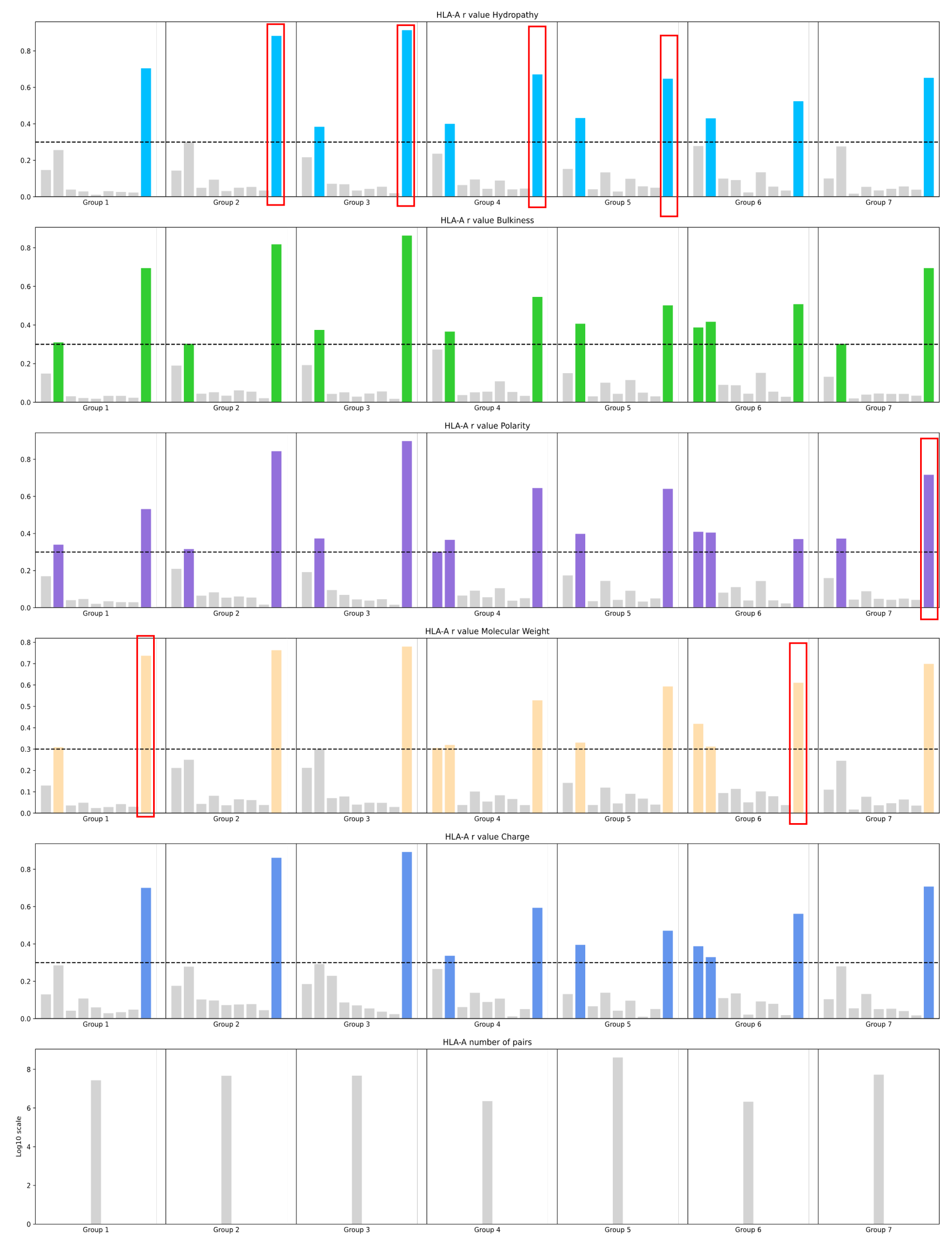
Same HLA cluster shows similar GradCAM results

GradCAM is one of explainable AI[11]. GradCAM shows weighted signals in amino acid interaction map. We think if sequence logos of each allele are similar, GradCAM results will be similar. Thus, we conducted clustering based on position frequency matrix of binding peptides. We analyzed a cluster group. GradCam results within one group were found to be similar.



Correlation analysis between GradCAM result and physiochemical properties of amino acids by position

It has been studied the secondary anchor and last anchor position were related in stability of binding between peptide and MHC class I molecules. Correlation analysis was conducted about physiochemical properties of amino acids and GradCAM results by position. In this study, we used hydropathy, bulkiness, polarity, molecular weight and charge value as physiochemical properties [12-13] . We hypothesized that certain amino acids properties contributes to bind MHC class I with in cluster (see above figure). We found different correlation patterns between groups and positions. Although the correlation pattern for each group is different, it is generally strong in p2,9. This result supports the previously reported studies.



Reference

[1] Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE, IPD-IMGT/HLA Database, Nucleic Acids Research (2020) **48**:D948-55
[2] Jurtz,V., Paul,S., Andreatta,M., Marcatili,P., Peters,B. and Nielsen,M. (2017) NetMHCpan-4.0: improved peptide-MHC class i interaction predictions integrating eluted ligand and peptide binding affinity data. J. Immunol., 199, 3360–3368.
[3] Sarkizova, S., Klaeger, S., Le, P.M. et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat Biotechnol **38,** 199–209 (2020).
[4] Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. (2017) Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. PLoS Comput Biol 13(8): e1005725.
[5] Racle, J., Michaux, J., Rockinger, G.A. et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat Biotechnol **37,** 1283–1286 (2019).
[6] Jha AN, Vishveshwara S, Banavar JR. Amino acid interaction preferences in proteins. Protein Sci. 2010;19(3):603-616.
[7] Kim, K., Kim, H.S., Kim, J.Y. et al. Predicting clinical benefit of immunotherapy by antigenic or functional mutations affecting tumour immunogenicity. Nat Commun **11,** 951 (2020).
[8] Mingxing Tan, Quoc V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks , arXiv:1905.11946 (2019).
[9] Vita R, Mahajan S, Overton JA, Dhanda SK et al. The Immune Epitope Database (IEDB): 2018 update, Nucleic Acids Res. 2018 Oct 24. doi: 10.1093/nar/gky1006.
[10] Timothy J. O'Donnell, Alex Rubinsteyn, Uri Laserson, MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing, Cell Systems, Volume 11, Issue 1, 2020,, (2015)
[11] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, Morten Nielsen, 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 Research, Volume 48, Issue W1, 02 July 2020, Pages W449–W454
[11] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, arXiv:1610.02391v4 (2016).
[12] Biro, J. Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theor Biol Med Model* 3, 15 (2006).
[13] Diego Chowell, Sri Krishna et al, TCR contact residue hydrophobicity is a hallmark of immunogenic CD8$^+$ T cell epitopes, PNAS April 7, 2015 112 (14) E1754-E1762;