Cancer Neoantigen Prediction using Deep Learning Approach

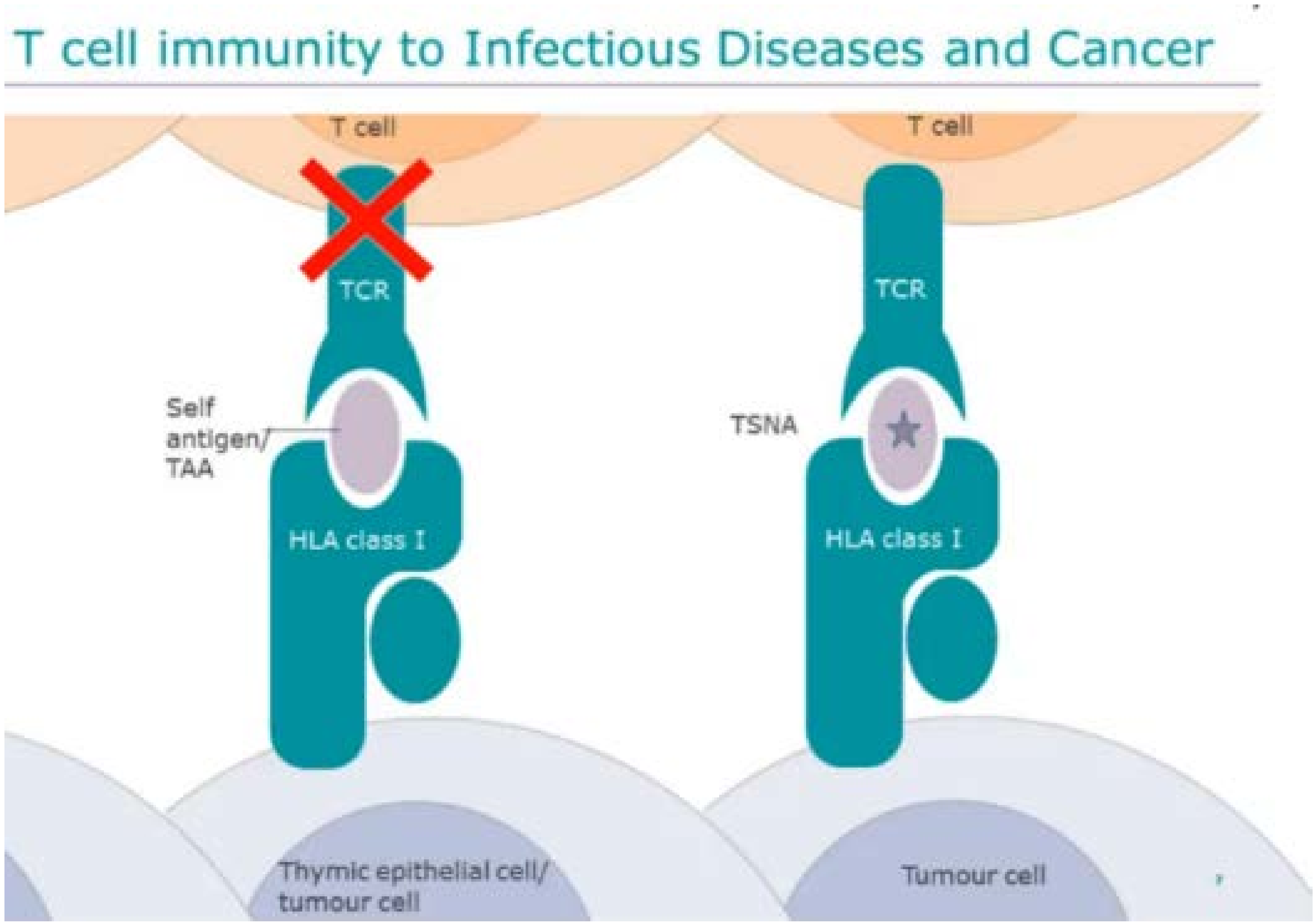
GNBF6010 - Research Project, CUHK

Supervised by Professor Sun Hao

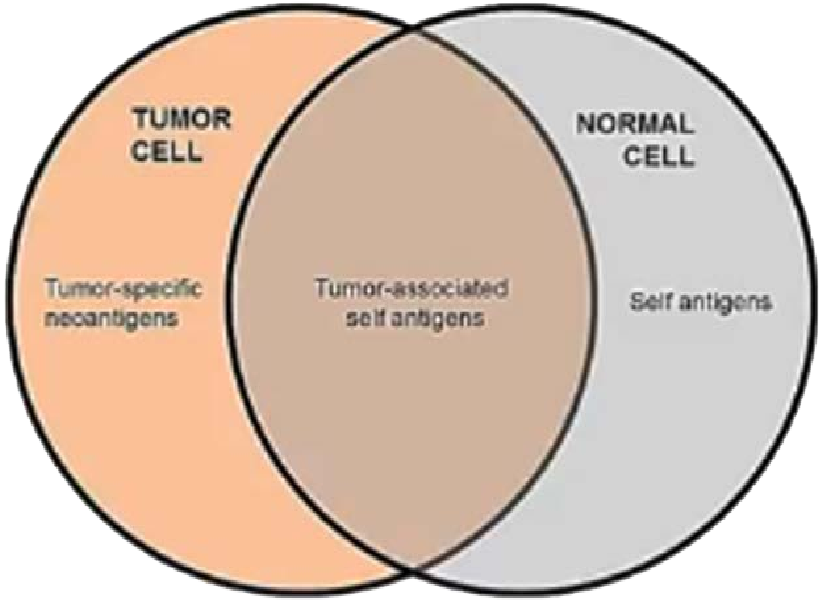
Presented by HO Wan Ping, Brian

15th May 2021

About Hunting Neoantigen

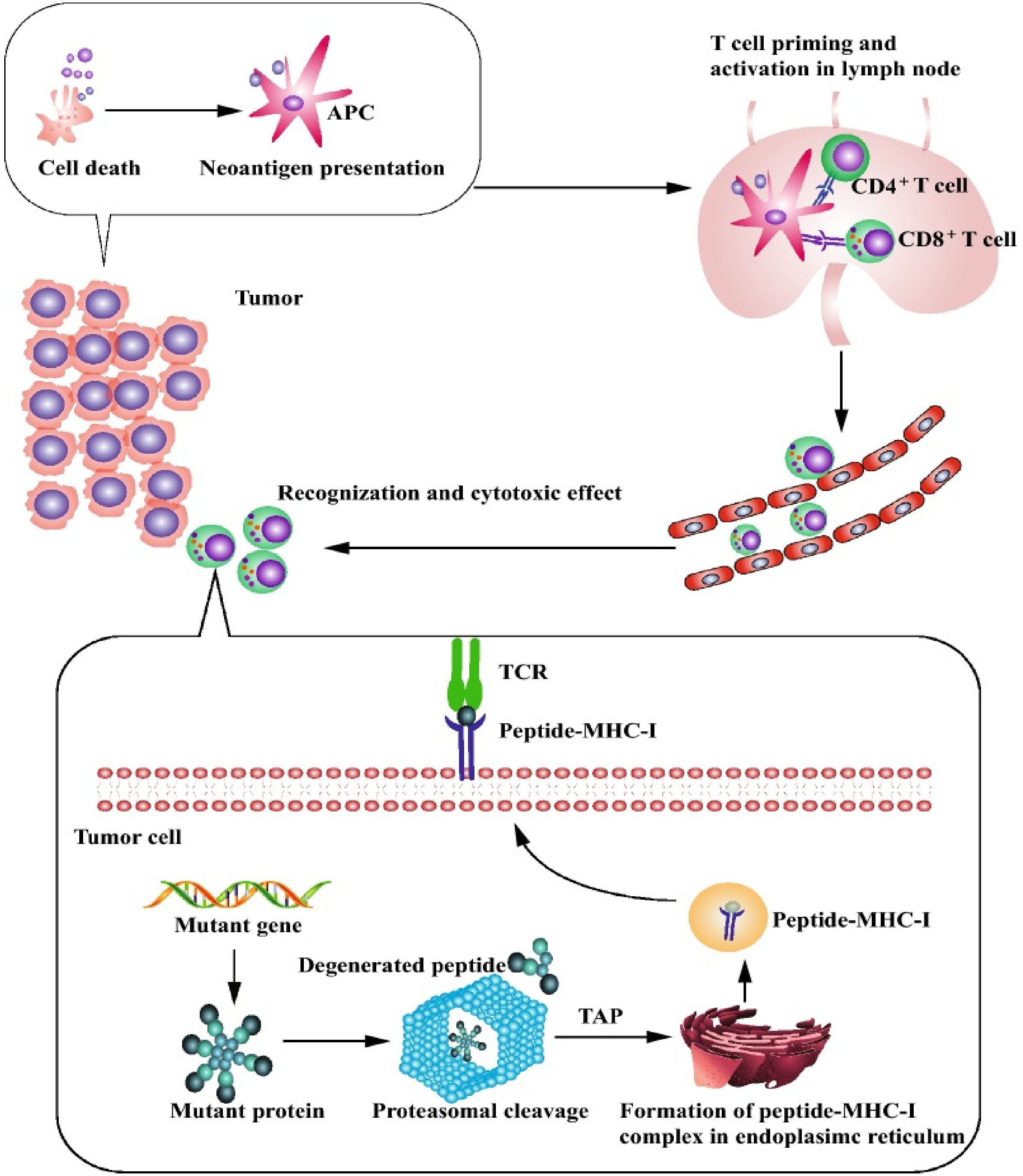


T cell immunity to Infectious Diseases and Cancer

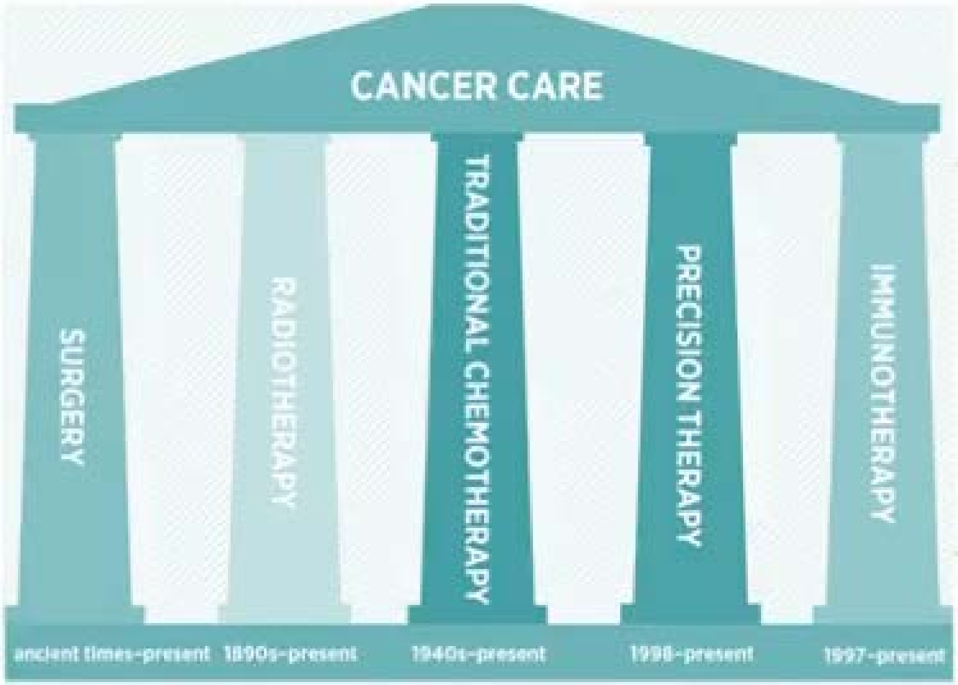


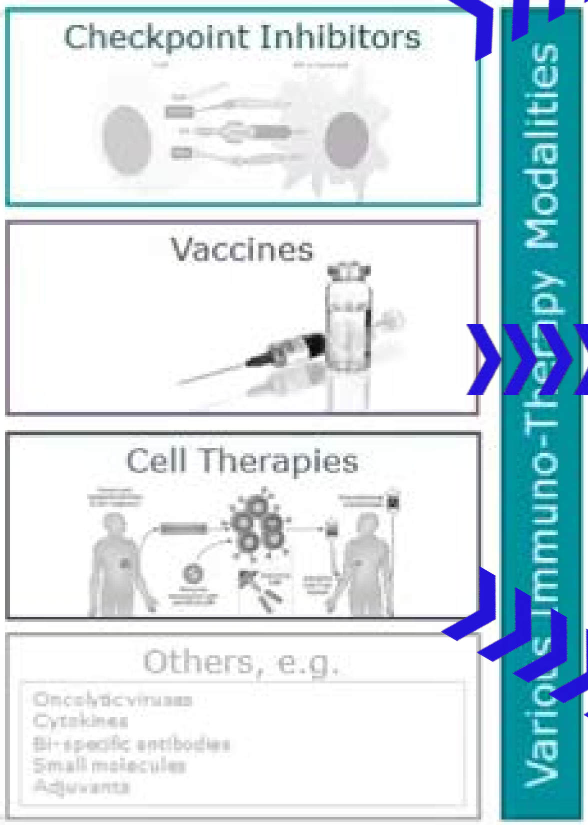
Background

- Tumor ***neoantigens*** generated by somatic mutations producing **novel peptides** that are bound & presented by HLA on cancer cell surface, to be ***recognized as foreign by T-cell leading to immune response***
- **All HLAs Classes:** ***Class I & II significant***
- ***HLA class I: the most selective*** requirement for a peptide to be presented
- HLA genes: ***Highly polymorphic***
  - i.e. > 17,000 HLA alleles reported in IMGT/HLA DB (Mar 2018)



Significance to Cancer Care




Various Immuno-Therapy Modalities

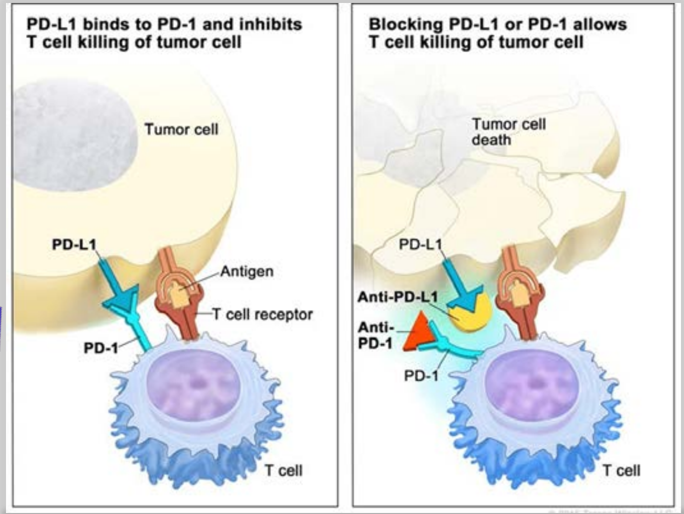
Checkpoint Inhibitors

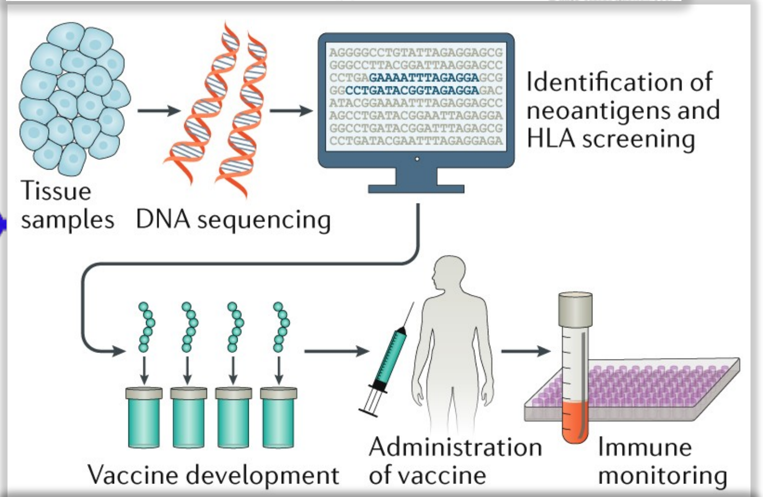
Vaccines

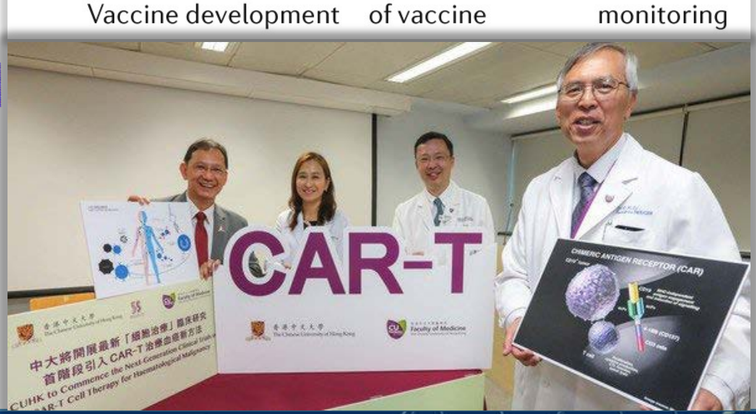
Cell Therapies

Others, e.g.

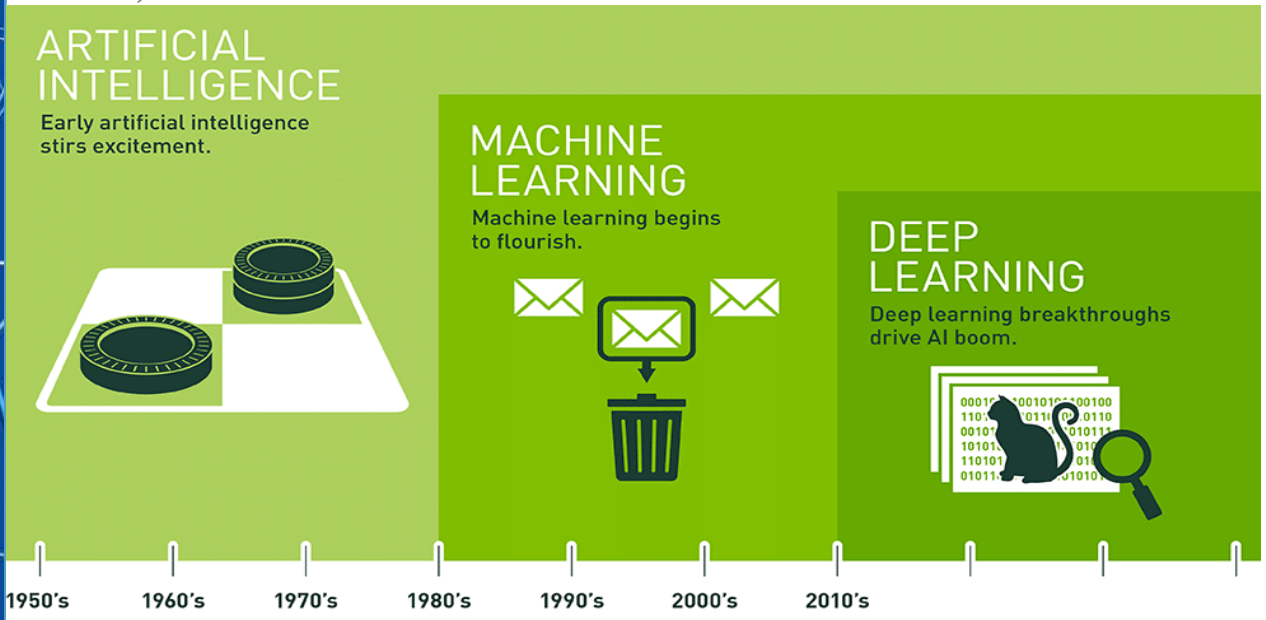
Oncolytic viruses
Cytokines
Bi-specific antibodies
Small molecules
Adjuvants

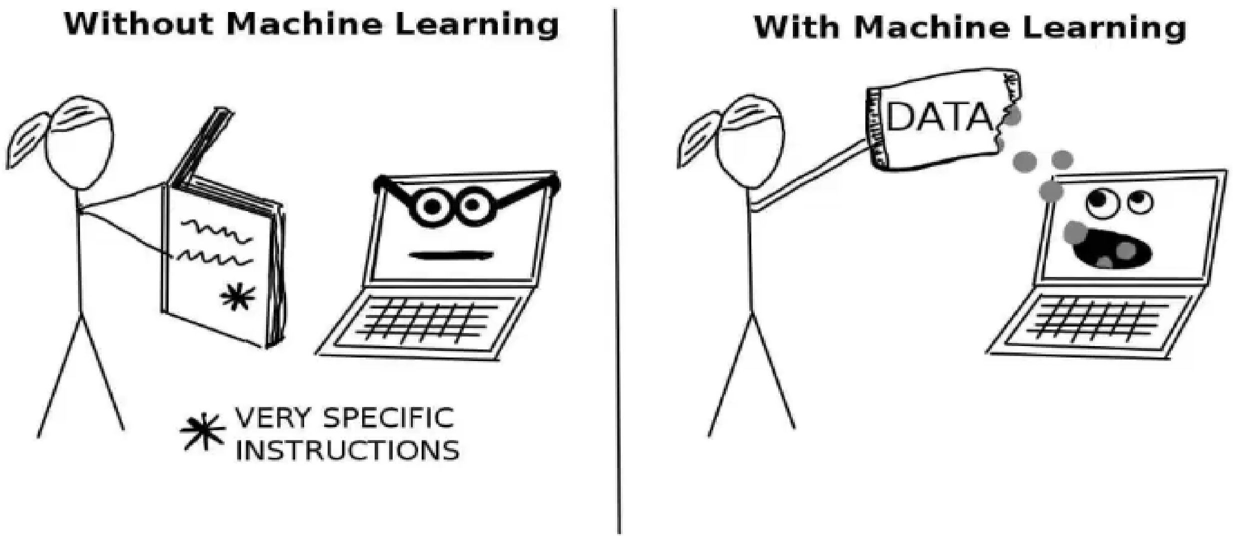

PD-L1 binds to PD-1 and inhibits T cell killing of tumor cell
Tumor cell
PD-L1
Antigen
T cell receptor
PD-1
T cell
Blocking PD-L1 or PD-1 allows T cell killing of tumor cell
Tumor cell death
PD-L1
Anti-PD-L1
Anti-PD-1
PD-1
T cell

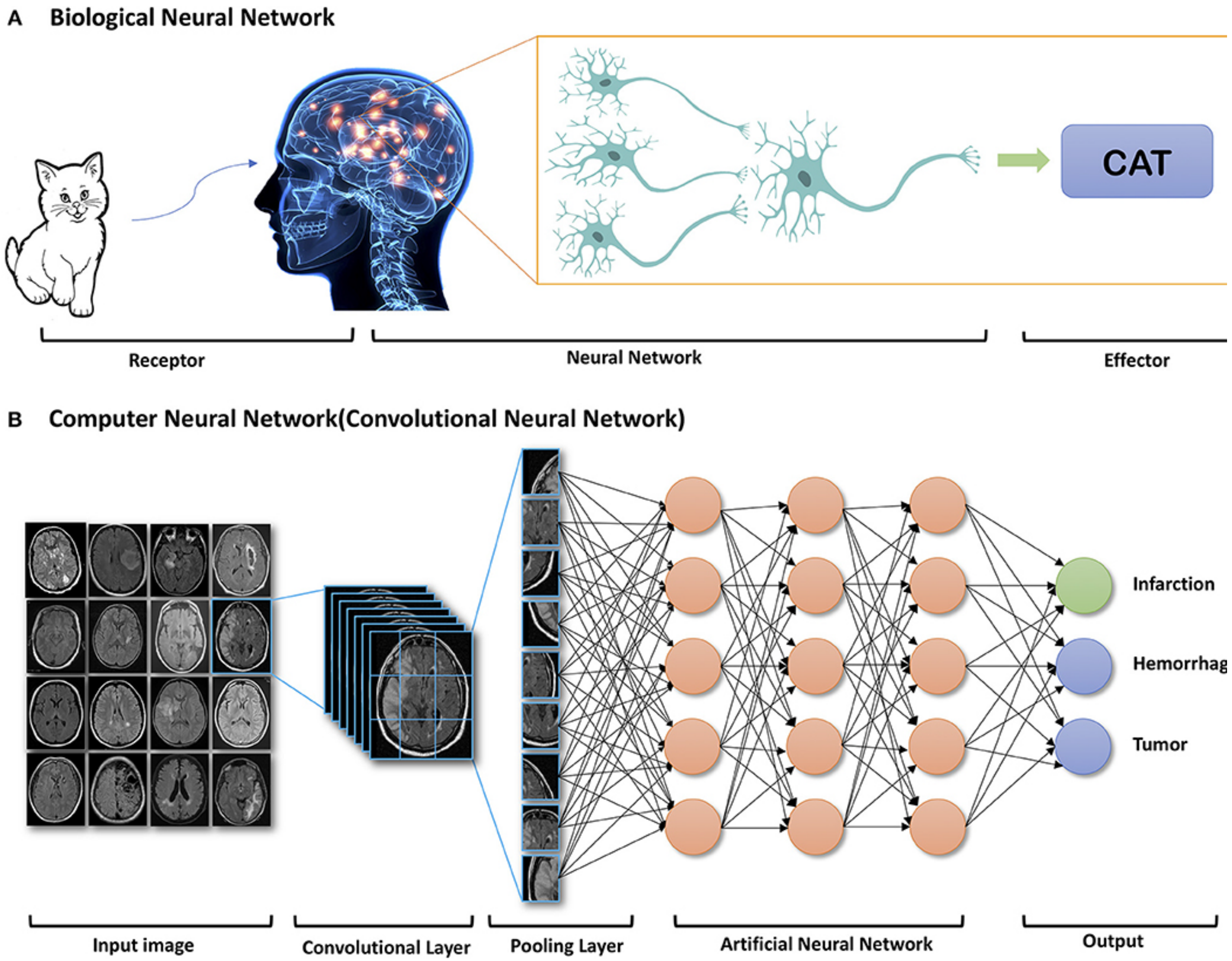

Tissue samples
DNA sequencing
Identification of neoantigens and HLA screening
Vaccine development
Administration of vaccine
Immune monitoring



What is Deep Learning?


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's



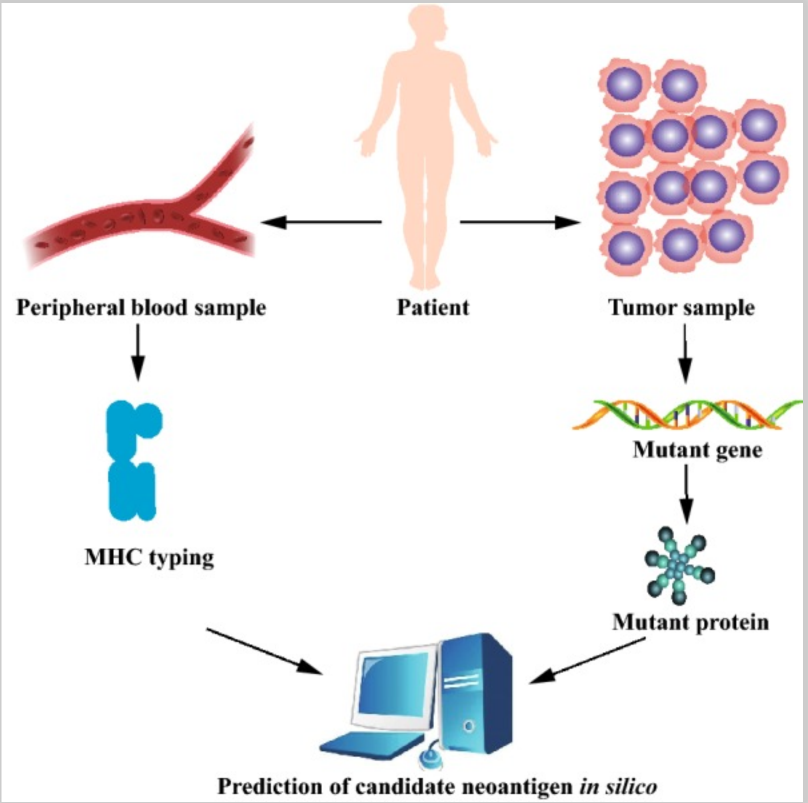

A Biological Neural Network
CAT
Receptor
Neural Network
Effector
B Computer Neural Network(Convolutional Neural Network)
Infarction
Hemorrhag
Tumor
Input image
Convolutional Layer
Pooling Layer
Artificial Neural Network
Output

Why Deep Learning?

- ***Experimental*** identification of peptide-MHC or neoantigens:
  - ***Costly*** invasive difficult-to-obtain clinical specimens
  - ***Time-consuming*** screening of hundreds to thousands of synthetic peptides or tandem minigenes, which may be only relevant to specific HLA alleles
  - *Clinically **unfeasible***
- **Computer-assisted** binding predictions
  - ***Cost-effective*** & ***faster*** alternative
  - *Accurate:* best prediction achieved by neural network-based pan-specific models
- **Abundance** of binding affinities **data** in databases *e.g. IEDB, SYEPEITHI & MHCBN*

☒

☑



What is happening? Other Predictors

- Categories
  1. ***Allele-specific*** *e.g. NetMHC & SMM* vs.

     Pan-specific *e.g. MHCFlurry*
  2. ***HLA class I*** and or ***class II*** *e.g. NetMHCPan, PickPocket*
- Other Predictors' Features
  - *Architectural Variety* in Deep Learning

    e.g. FFNN, RNN, CNN, Autoencoder, LTSM etc.
  - ***Other processing pathway factors*** in HLA class I

    e.g. Proteosomal cleavage & transporter-associated antigen processing (TAP)-mediated peptide transport
  - Inclusion of ***other type of data*** for training

    e.g. mass spectrometry (MS)-based HLA peptidome data, transcriptomic data (RNA-Seq), structural data (Hi-C data) (DeepAntigen)
- General ***Good Binding Affinity Predictions***
- ***Low*** Prediction in final *Immunogenicity e.g.* ***< 5%*** *(Bulik-Sullivan et al.)*

NeonMHC2
MHCherryPan
DeepLigand
AI-MHC
ConvMHC
MARIA
DeepSeqPan
ACME
MHCSeqNet
DeepHLApan
MHCflurry
PUFFIN
DeepMHC
HLA-CNN
DeepNeo
DeepSeqPanII
DeepAttentionPan
MHCnuggets
USMPep

Deep Learning Model in progress

1 Allele specific model by CUHK PhD Student

2 Optimizer: RMSprop algorithm

3 Probabilistic losses: cross-entropy loss

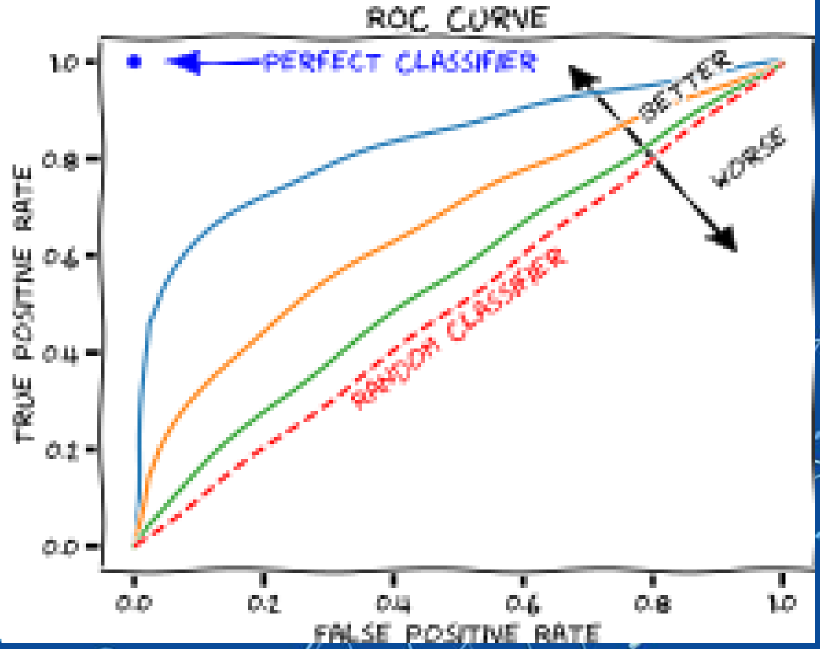
4 Metrics= 'accuracy‘

5 url:https://github.com/wenwenwendy/MiS/tree/main/predict_function

Method

- *Black Box Testing* - Collected data from IEDB & other predictors' papers
- Focus on prediction of HLA-peptide pair *binding affinity ONLY* by below predictors
  1. *DeepSeqpan* (default trained CNN model)
  2. *DeepHLA* (default trained RNN model)
  3. *MHCFlurry* (default trained FFNN model)
  4. *MiS* (ANN in progress by CUHK PhD student)
- Predictor's ***ROC*** (Receiver Operator Characteristic) graphs & ***AUC*** (*area under the curve*) *comparisons*

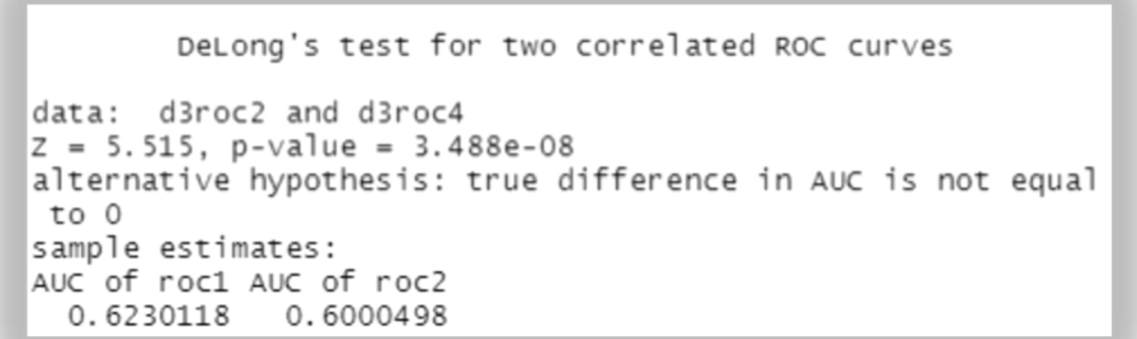
Data Collection



Data Source & Evaluation Metrics

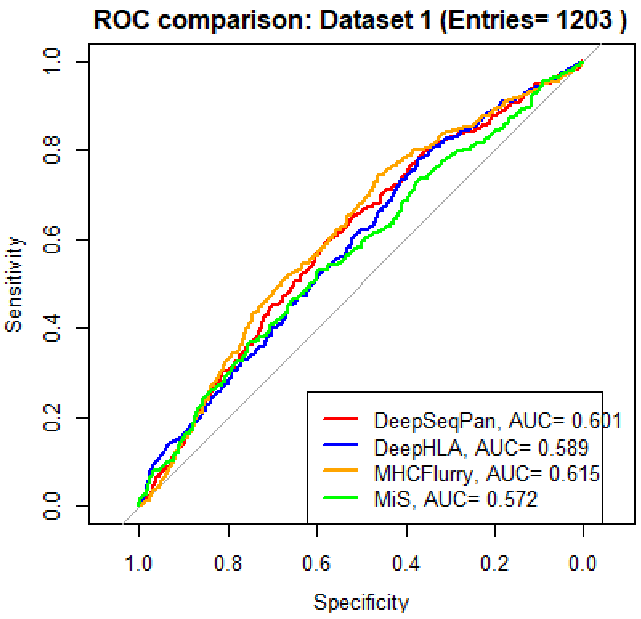
	Source	Past Usage	Features	Sub-total	Filters	Total (For Evaluation)
Dataset 1	IEDB	Training Data (DeepAntigen) Evaluation Data (DeepAntigen)	1) Homo Sapian Only, mapped to hg19 2) T-Cell Assay Immunogenicity (May 2018) 3) Peptide Length = 9 4) MHC-1 Subtype Only	4,339	Unsupported HLA Types	1,203
Dataset 2	IEDB (May 2018)	Training Data (DeepHLA)	1) Collected 280,525 binding data pairs 2) HLA-A, B C Subtypes = 81 HLA Alleles 3) Peptide Length = 8-15 4) Balancing allele proportion by creating 156,552 pseudo-HLA-peptide pairs from Ensembl database (38) with binding data predicted by DeepHLA Basic Model	437,077	Unsupported HLA Types Peptide length =**9**	303,616
Dataset 3	IEDB (May 2018)	Training Data (DeepHLA)	1) HLA-Pepitde pairs with Immunogenicity data 2) 7212 pairs immunogenic, of which 3013 related to HL-A02:01 3) Peptide Length = 8-15	32,785	Unsupported HLA Types Peptide length =9	17,195
Dataset 5	IEDB Weekly Benchmark Data (April 2021 – Mar 2014) URL: tools.iedb.org/auto_bench/mhci	Evaluation Data For Many Predictors (NetMHCons, IEDB Consensus, DeepSeqPan. SMM. NetMHCpan. MHCFlurry, Pickpocket, ANN3.4. SMMPMBEC. ANN3.4)	1) SLA molecules from Sus scrofa excluded 2) Pairs with binary binding data only 3) Peptide Length = 9-11	19,177	Unsupported HLA Type Peptide length =9	15,079

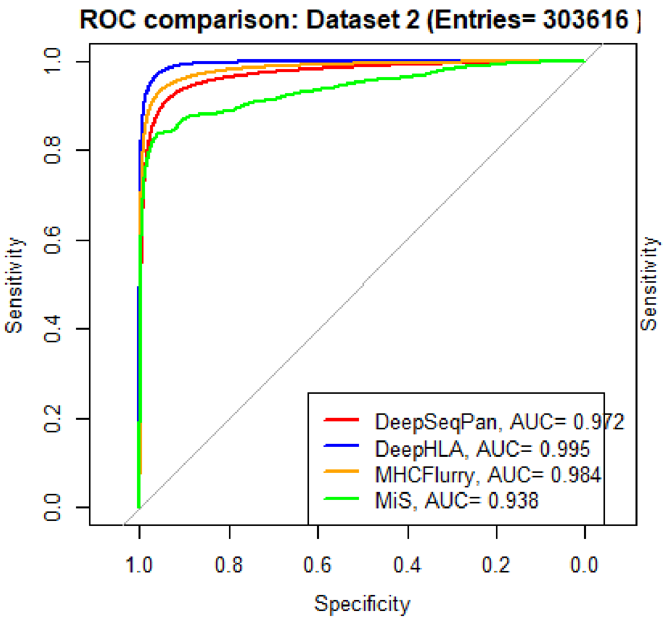
- **DeLong Test**: A nonparametric test comparing >= 2 AUC correlated ROC curves. (*DeLong et al.)*

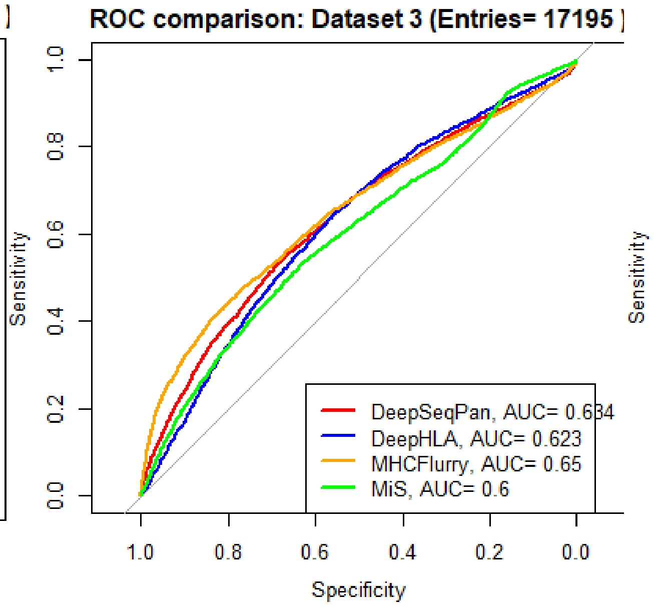


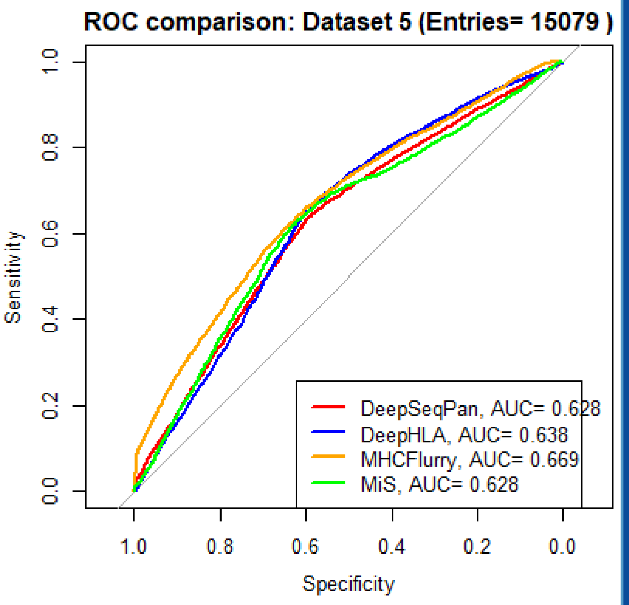
```
        DeLong's test for two correlated ROC curves

data:  d3roc2 and d3roc4
Z = 5.515, p-value = 3.488e-08
alternative hypothesis: true difference in AUC is not equal
 to 0
sample estimates:
AUC of roc1 AUC of roc2
  0.6230118   0.6000498
```


ROC comparison: Dataset 1 (Entries= 1203)

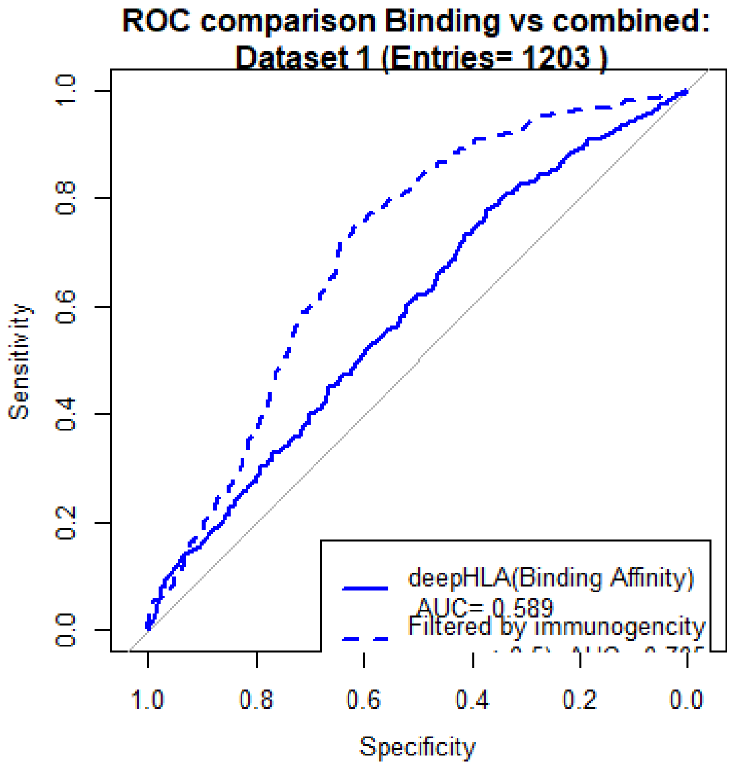

ROC comparison: Dataset 2 (Entries= 303616)

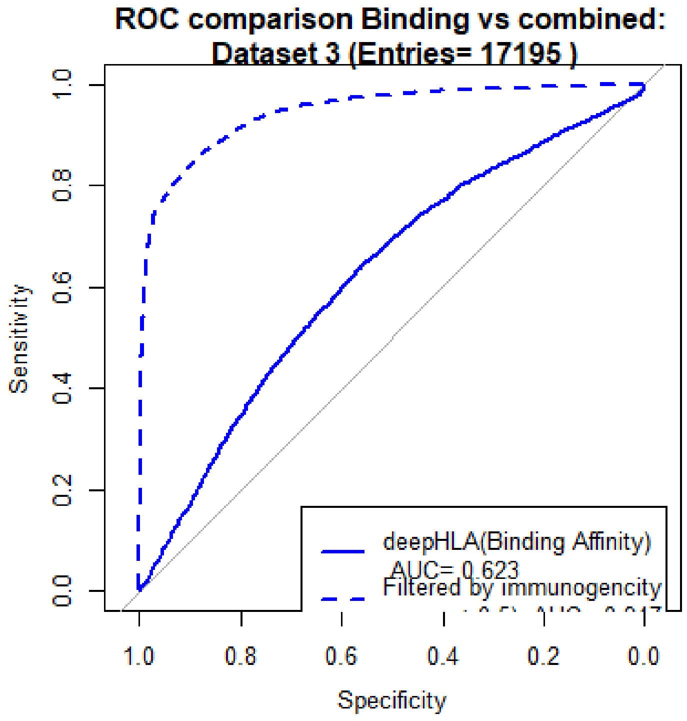

ROC comparison: Dataset 3 (Entries= 17195)

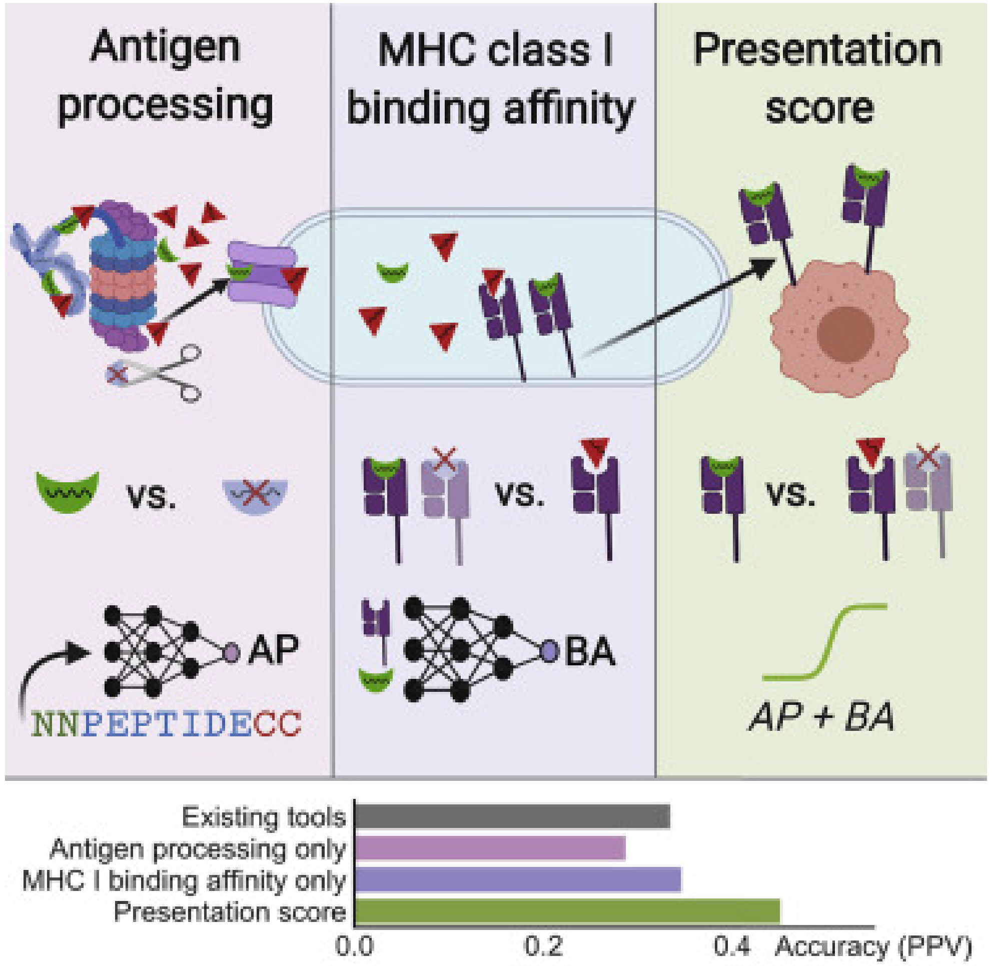

ROC comparison: Dataset 5 (Entries= 15079)

	DATASET 1		DATASET 2		DATASET 3		DATASET 5	
	AUC	vs. MiS's AUC H_A P-Value (DeLong's Test)	AUC	vs. MiS's AUC H_A P-Value (DeLong's Test)	AUC	vs. MiS's AUC H_A P-Value (DeLong's Test)	AUC	vs. MiS's AUC H_A P-Value (DeLong's Test)
DeepSeqPan	0.601	0.09515	0.972	2.20E-16	0.634	6.75E-12	0.628	0.8074
DeepHLA	0.589	0.2225	0.995	2.20E-16	0.623	3.49E-08	0.638	8.20E-05
MHCFlurry	0.615	0.005728	0.984	2.20E-16	0.65	2.20E-16	0.669	2.20E-16
MiS	0.572	N/A	0.938	N/A	0.6	N/A	0.628	N/A

DeepHLA Better Performance with extra Immunogenicity Prediction Model







Discussion

- Comparison of ROCs shows **MiS ‘s fair or marginally trailing binding prediction** performance relative to other predictors
- Potential Prediction Enhancement with other factors in the pathway
  - e.g. abundance of proteins, antigen processing & proteosomal cleavage
- To also predict **other HLA Class 1 Alleles** & **immunogenicity** i.e. Cell Surface presentation

Thank You & Q&A