Immunogenic peptide discovery for T cell immunity with deep learning

Jesse Elliott

CS 886: deep learning for biotechnology

David R. Cheriton School of Computer Science
Institute for Quantum Computing
University of Waterloo

Immunogenic peptide discovery for T cell immunity

Which peptides will elicit a T cell immune response?

Applications

- Vaccines for T cell immunity (for viruses and cancer treatments)

Advantages of T cell immunity over antibodies

1. Harder for a virus to mutate and evade T cell immunity
2. T cell immunity lasts longer
    - Antibodies last for months whereas T cell immunity lasts for years

Preliminaries

The immune system

1. The innate immune system
2. The adaptive immune system

The immune system

Caveats

- The immune system is complicated
- What follows is a high level overview, leaving out many details

The immune system (main functions)

What does the immune system do?

1. Prevents pathogens from entering the body
2. Kills pathogens that have entered the body

Pathogens

- Proteins, viruses, bacteria, parasites, fungi, etc.

The innate immune system (overview)

All pathogens are treated equal

What does the immune system do?

1. Prevents pathogens from entering the body
2. Kills pathogens that have entered the body

(1) first line of defense

- Skin
- Mucous membranes
- Stomach acid

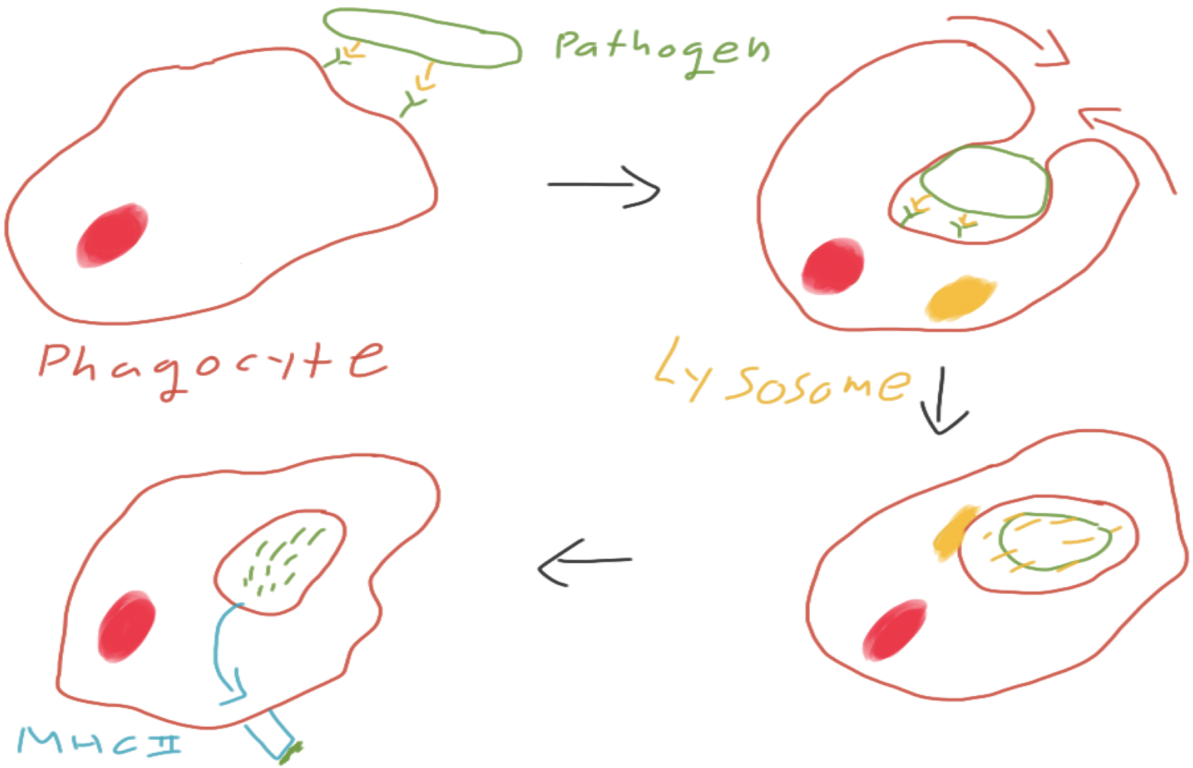
(2) second line of defense

- Phagocytes
  - Latin, essentially meaning eating cells
  - White blood cells that non-specifically eat pathogens

Phagocytosis

Phagocyte

- Pathogen eating white blood cell (non-specific)



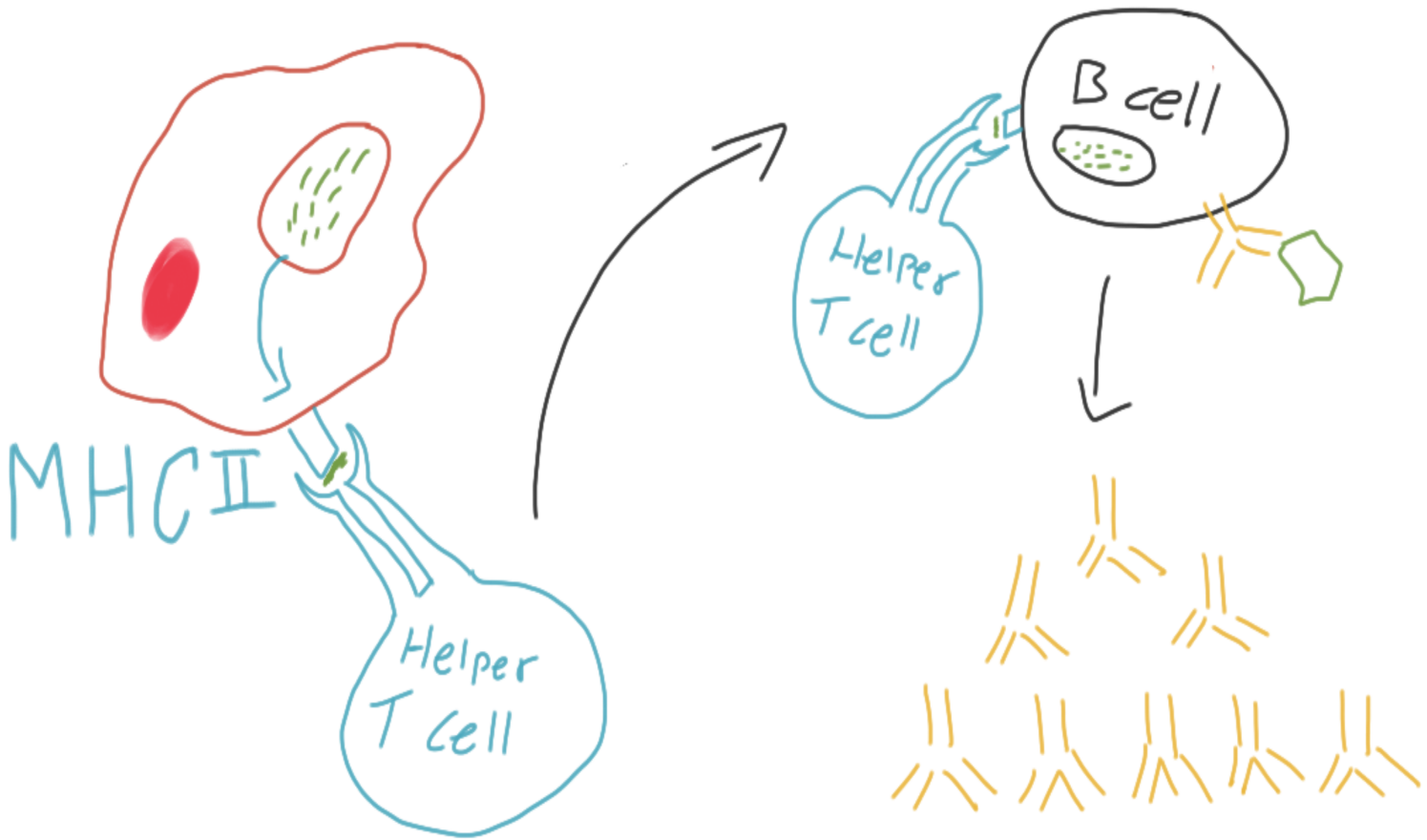
Antigen presenting cells (APCs)

- APCs intermediate the innate and adaptive immune systems

Major histocompatibility complex (MHC)

- Histo means tissues; refers to organ and tissue transplantation
- Genes coding cell surface proteins activating adaptive immunity

MHCII



The adaptive immune system

Types of adaptive immunity

1. B cells
2. T cells

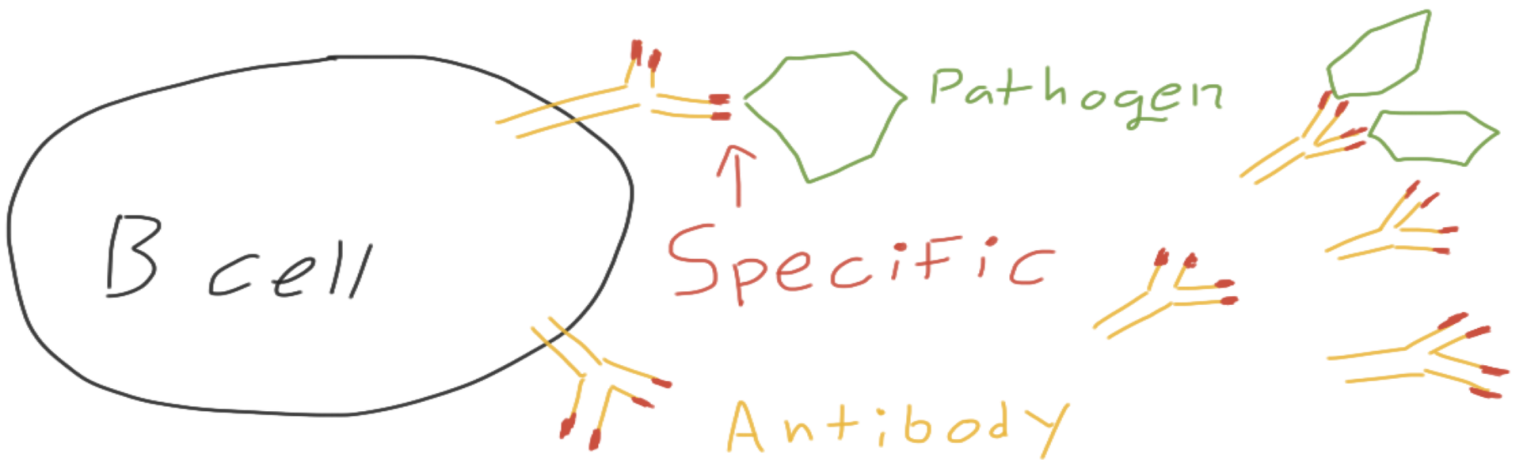
B cells

B cell (B lymphocytes)

- Produce antibodies
  - Y shaped proteins that neutralize specific pathogens
  - Also called immunoglobulins

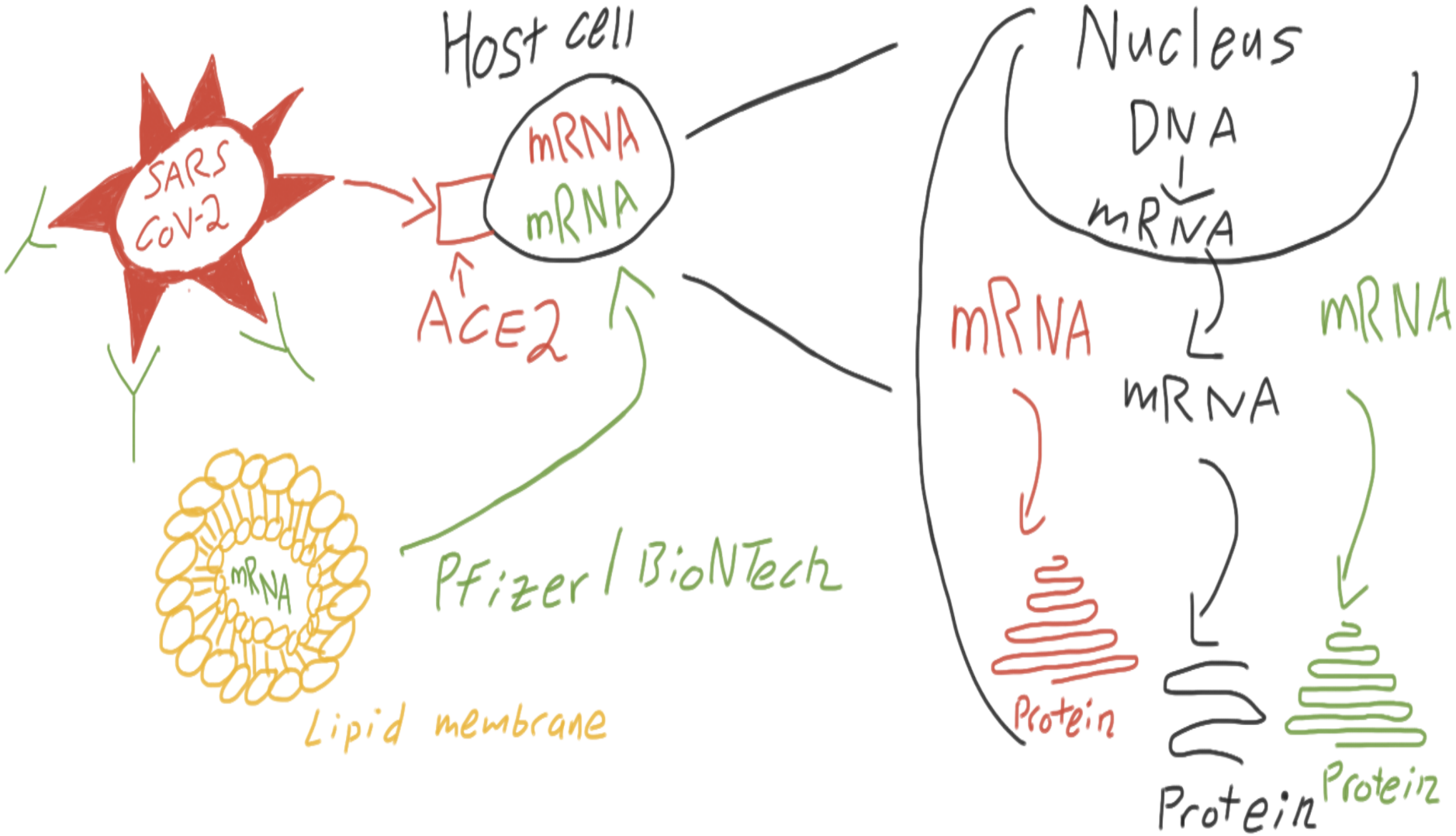
Humeral immunity

- In the blood



B is for bursa of Fabricius

Pfizer / BioNTech



T cells

T cell (T lymphocytes)

1. T helper cell
   - Help / activate B cells
2. T suppressor cell
   - Suppress / deactivate B cells
3. T cytotoxic/killer cell
   - Destroy abnormal cells

Cell mediated immunity

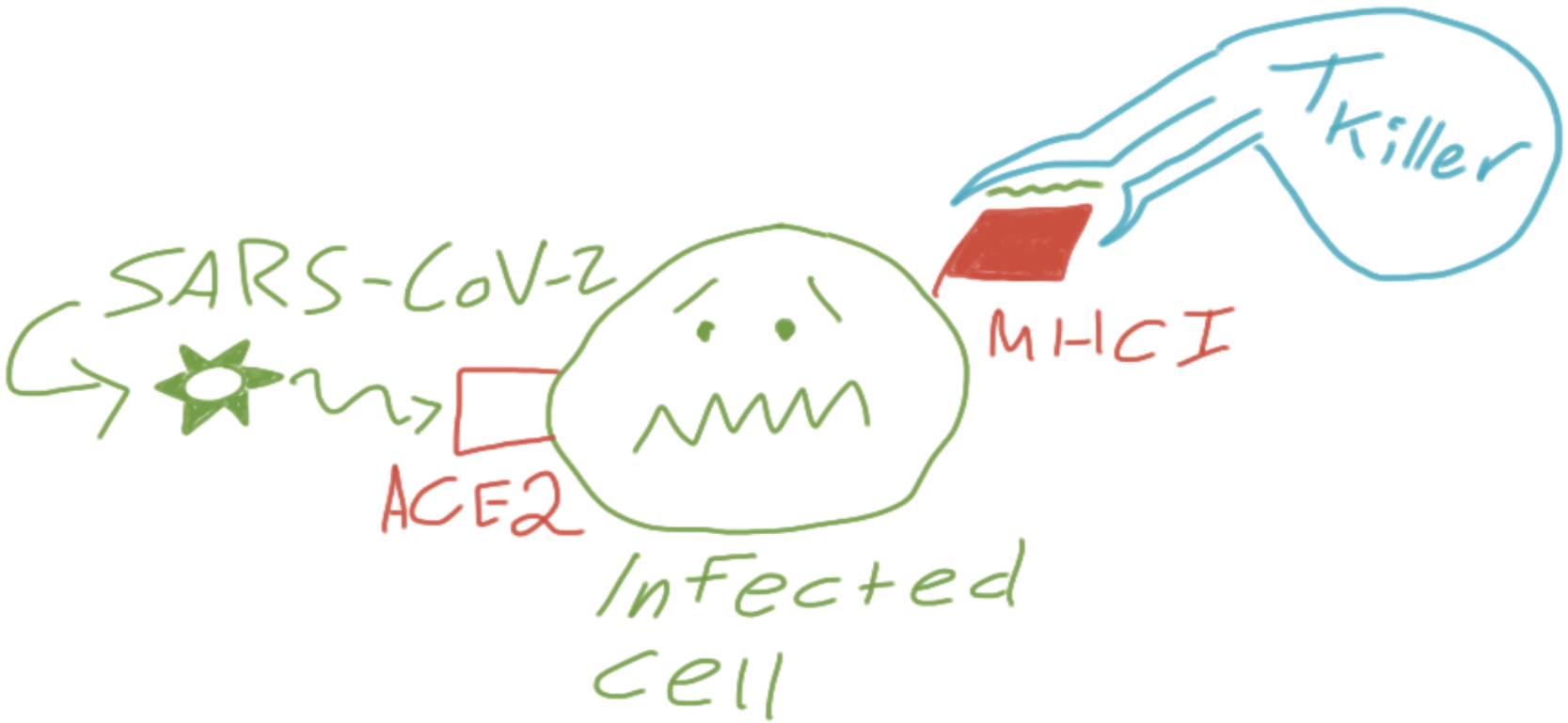
- Kill infected cells

T is for thymus because T cells mature in the thymus

Cytotoxic / killer T cells

MHC-I molecule

- Killer T cells communicate with MHC-I molecules
- MHC-I molecules present viral peptides on the cell surface



Genetic variability of T cells

MHC molecules are highly genetically variable; thousands of variants exist

Human Leukocyte Antigen (HLA)

- Human version of MHC
- Genes located on chromosome 6
- Among the most genetically variable regions on the human genome
- HLA-alleles give different MHC which bind with different peptides

Deep learning for immunogenic peptide discovery

Want peptides that do 2 things

1. Bind to MHC-I (binding affinity)
2. MHC-I peptide complex is recognized by TCR (immunogenicity)

Why deep learning?

1. Predicting binding affinity
2. Predicting immunogenicity

I mainly focus on work that uses deep learning to predict immunogenicity

Previous work - support vector machine classification

$$SVM: \ (\text{peptide}) \rightarrow \begin{cases} 1, & \text{immunogenic} \\ 0, & \text{not immunogenic} \end{cases}$$

POPI: prediction of peptide immunogenicity (2007)

Chun-Wei Tung, Shinn-Ying Ho

POPISK: T-cell reactivity prediction using support vector machines and string kernels (2011)

Chun-Wei Tung, Matthias Ziehm, Andreas Kämper, Oliver Kohlbacher, Shinn-Ying Ho

Issues

- Small training datasets
- Limited consideration of HLA-alleles

Previous work - recurrent neural network

DeepHLApan: a deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity (2019)

Jingcheng Wu,Wenzhe Wang,Jiucheng Zhang,Binbin Zhou,Wenyi Zhao,Zhixi Su,Xun Gu,Jian Wu,Zhan Zhou, Shuqing Chen

- Algorithms for binding affinity and immunogenicity
- 3 layer bidirectional Gated Recurrent Unit with an attention layer

Note

- Encoding of the amino acid sequence does not incorporate physicochemical properties

Main paper for discussion

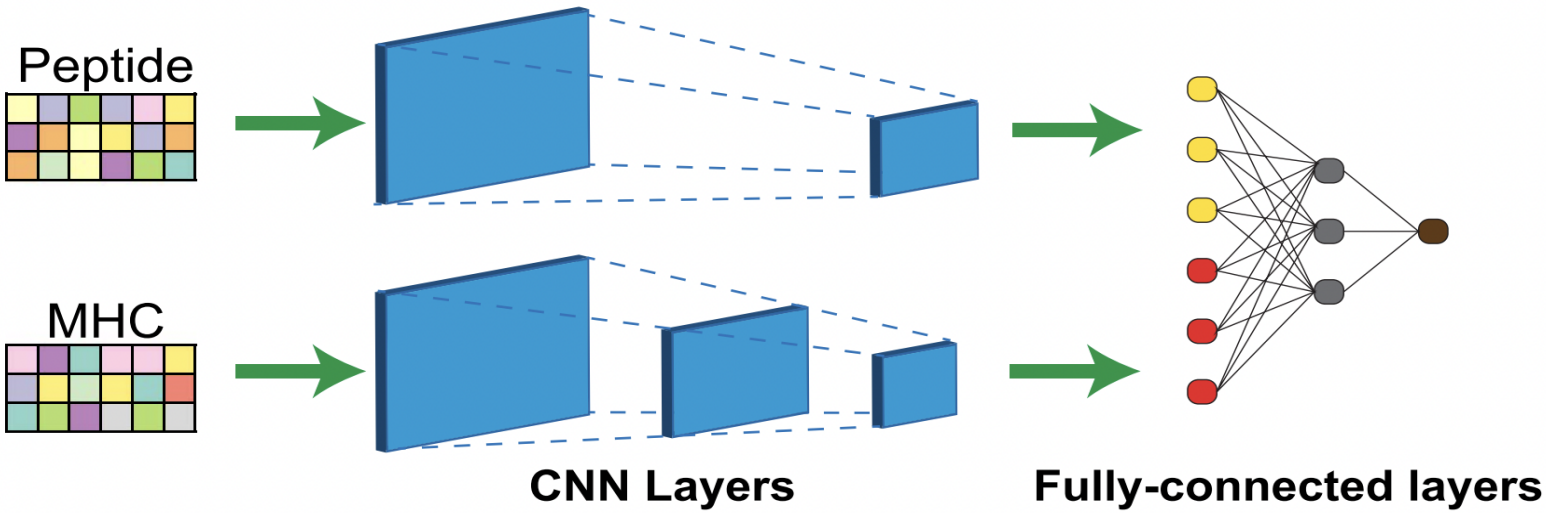
DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity (2021)

Guangyuan Li, Balaji Iyer, V B Surya Prasath, Yizhao Ni, Nathan Salomonis

Main contributions

1. DeepImmuno-CNN: predicts immunogenicity of peptide-HLA pairs
2. DeepImmuno-GAN: simulates synthetic immunogenic peptides
   - Large synthetic training datasets
   - Help us learn rules governing which peptides are immunogenic and why

DeepImmuno-CNN



- Two consecutive convolutional layers
- Two fully connected dense layers
- Input: peptide-HLA pair
- Output: continuous predictive value in [0,1]

DeepImmuno-CNN

Training data: labeled by sampling from a beta distribution

$$(peptide, HLA) \rightarrow d \sim \mathrm{Beta}(\alpha + S, \beta + F)$$

Beta distribution: Beta(α, β)

- Continuous distribution defined on [0, 1] with parameters α (# successes) and β (# fails)

Example: bin of red and blue balls

- 10 balls chosen with 5 red; the proportion of red $\sim$ Beta(5,5)

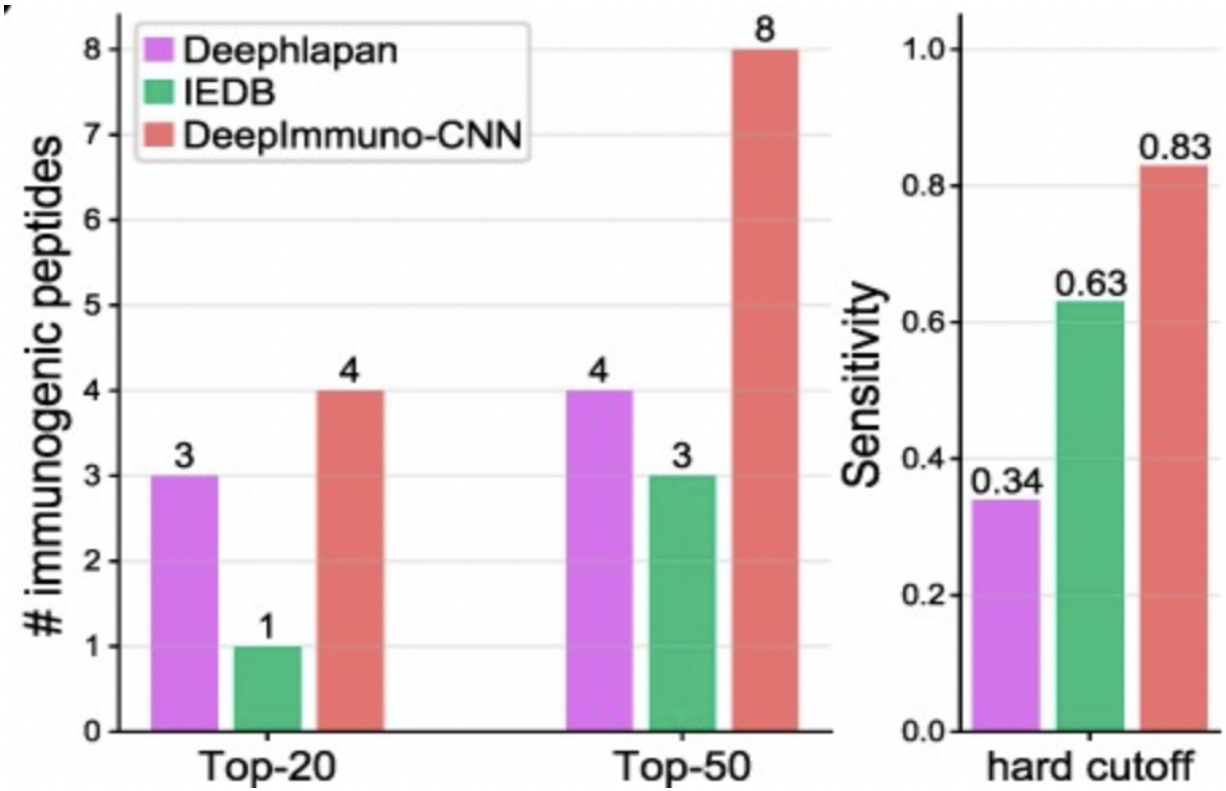
The immune epitope database (IEDB) and analysis resource

DeepImmuno-CNN training dataset: $>$ 9000 assays obtained from IEDB

The immune epitope database (IEDB)

- Systematically characterizes the biochemical properties of over 30 000 MHC-I-bound immunogenic peptides (with associated HLA alleles)
- Suit of algorithms for predicting binding affinity and immunogenicity
- Maintained by the *National Institute for Allergy and Infection*
- Created to assist biomedical researchers in the development of new vaccines, diagnostics, and therapeutics
- Established in 2003

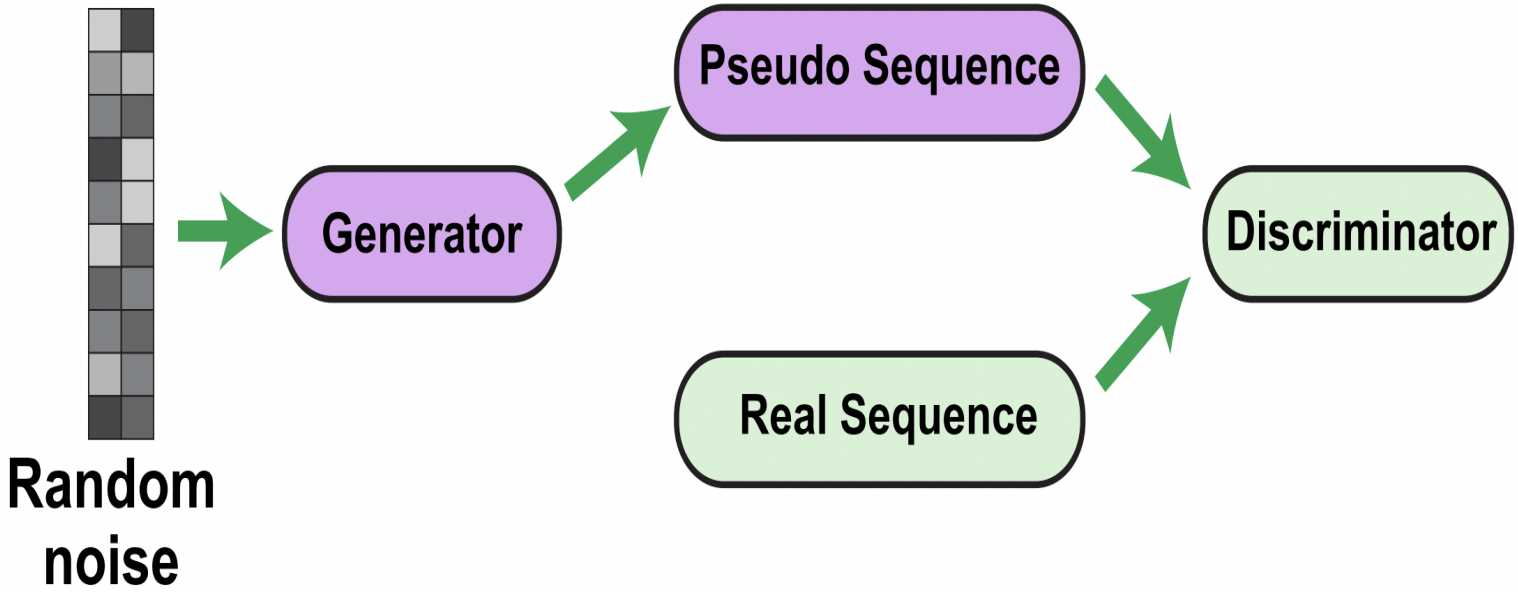
Tumor neoantigen dataset



Number of true-positives with top 20 or top 50 predictions (Left)

Sensitivity using a 0.5 scoring threshold (Right)

DeepImmuno-GAN (synthetic immunogenic peptides)



- Generates pseudo-sequences to convince the discriminator
- Discriminator uses real peptide sequences to distinguish the difference

Applications

- Large synthetic training datasets
- Help us learn rules governing which peptides are immunogenic and why

DeepImmuno-GAN (synthetic immunogenic peptides)

Pseudo-sequence	Real sequences	Similarity
MLIGLIPVLV	MLIGIPVYV	0.84
KILLGFQINV	KLLGQINLV	0.84
FLLKLTLGLV	FLLKLTPLL	0.84
ALLAVKLVRV	ALSLAAVLVV	0.80
QLRSYGLRAV	QLRSLGLNAV	0.80
ALLAVLVSFV	ALSLAAVLVV	0.80

Generative pseudo-sequences and their most similar counterparts in HLA-A$*$0201 immunogenic peptides

End of presentation

Thank you, please ask questions or give comments if you have any

Comparison of models

ML models

- ElasticNet, KNN, SVM, Random Forest, AdaBoost

DL models

- CNN, ResNet, GNN

3 testing datasets

- Dengue virus, tumor neoantigens, and SARS-CoV-2

- Random Forest-based regressor had slightly better RMSE
- AdaBoost regression performed the best in the dengue virus dataset (average accuracy = 0.91)
- CNN model achieved superior performance in the neoantigen dataset
- All models achieved similar results on the SARS-CoV-2 dataset
- ResNet model overfit

Comparison of models

Supplemental Table 2

Model	Encoding	Validation	Dengue	Neoantigen			COVID-19 (convalesent)		COVID-19 (unexposed)	
		RMSE	Accuracy	top20	top50	Recall	Recall	Precision	Recall	Precision
ElasticNet	aaindex+paratope	0.29±0.006	0.85±0.004	1.2±0.4	5.9±0.7	0.83±0.015	0.73±0.016	0.27±0.007	0.78±0.050	0.09±0.006
	onehot+paratope	0.30±0.004	0.92±0.022	3.8±0.6	9.3±1.001	0.83±0.009	0.716±0.020	0.24±0.001	0.75	0.08±0.002
	aaindex+psedo34	0.29±0.006	0.86±0.003	1.4±0.490	5.7±0.640	0.83±0.011	0.728±0.016	0.27±0.006	0.775±0.050	0.09±0.026
SVM	aaindex+paratope	0.35±0.015	0.82±0.001	1.4±1.020	4.9±1.513	0.8	0.68±1.110	0.25	0.75	0.09
	onehot+paratope	0.33±0.029	0.82	3.4±0.66	0.3±0.781	0.8	0.68±1.11	0.25	0.75	0.8
	aaindex+psedo34	0.34±0.02	0.83±0.001	1.6±1.02	4.4±1.36	0.8	0.68	0.25±0.002	0.75	0.08±0.009
KNN	aaindex+paratope	0.28±0.006	0.82±0.005	2	4.0±1.0	0.79±0.032	0.72±0.016	0.27±0.005	0.86±0.038	0.10±0.004
	onehot+paratope	0.28±0.007	0.85±0.006	1.8±0.600	3.9±1.044	0.71±0.036	0.708±0.036	0.25±0.009	0.75	0.09±0.002
	aaindex+psedo34	0.29±0.005	0.81±0.005	2	4.9±0.094	0.80±0.017	0.71±0.005	0.27±0.005	0.85±0.050	0.10±0.006
Random Forest	aaindex+paratope	**0.26±0.006**	0.86±0.009	1.9±0.943	**6.0±1.0**	**0.84±0.042**	0.73±0.003	0.27±0.010	**0.9±0.050**	**0.11±0.006**
	onehot+paratope	0.26±0.007	0.83±0.012	2±0.89	4.4±0.8	0.79±0.029	0.724±0.022	0.26±0.008	0.875±0.056	0.10±0.006
	aaindex+psedo34	0.26±0.006	0.86±0.006	0.4±0.049	2.7±1.1	0.85±0.021	0.74±0.027	0.27±0.009	0.95±0.06	0.11±0.007
Adaboost	aaindex+paratope	0.30±0.005	**0.91±0.036**	1.5±0.806	4.3±1.418	0.82±0.014	0.74±0.027	0.25±0.007	0.75	0.08±0.003
	onehot+paratope	0.30±0.007	0.91±0.067	1.8±0.98	5.2±1.99	0.83±0.017	0.73±0.037	0.25±0.003	0.75	0.08±0.005
	aaindex+psedo34	0.30±0.005	0.86±0.050	0.7±0.64	2.6±1.02	0.81±0.014	0.7±0.03	0.25±0.005	0.75	0.08±0.005
CNN	aaindex+paratope	0.28±0.005	0.86±0.009	**2.9±1.044**	5.9±1.221	0.8±0.036	**0.74±0.051**	**0.27±0.013**	0.8±0.061	0.09±0.008
	onehot+paratope	0.28±0.005	0.85±0.021	2.7±0.781	6.4±1.02	0.78±0.061	0.78±0.08	0.28±0.015	0.83±0.08	0.09±0.007
	aaindex+psedo34	0.28±0.005	0.86±0.009	2.9±1.044	5.9±1.221	0.8±0.036	0.74±0.051	0.27±0.013	0.8±0.061	0.9±0.008
ResNet	aaindex+paratope	0.30±0.001	0.82±0.034	2.6±0.917	5.0±1.612	0.76±0.063	0.712±0.053	0.27±0.019	0.83±0.083	0.10±0.011
	onehot+paratope	0.31±0.005	0.81±0.030	1.9±1.04	5.3±1.269	0.71±0.088	0.69±0.069	0.26±0.025	0.79±0.098	0.09±0.011
	aaindex+psedo34	0.30±0.001	0.82±0.034	2.6±0.917	5.0±1.612	0.76±0.063	0.71±0.053	0.27±0.019	0.83±0.083	0.10±0.011

Considering its overall performance and the complexity of the model, the CNN was chosen for further analysis