Deep Learning of the Immune Synapse

J.W. Sidhom[1,2,3], D.M. Pardoll[1,2], A.S. Baras[1,2,4]

[1]Bloomberg~Kimmel Institute for Cancer Immunotherapy

[2]The Sidney Kimmel Comprehensive Cancer Center

[3]Department of Biomedical Engineering

[4]Department of Pathology

The Johns Hopkins University School of Medicine

Abstract

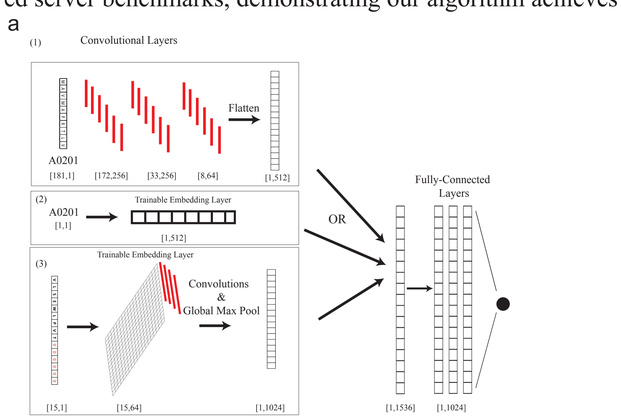
Artificial intelligence is poised to revolutionize every aspect of human life, finding applications in everything from self-driving cars to diagnosing cancer. In fact, almost any task that involves pattern recognition can be formulated in a way that modern AI algorithms can be used to achieve super-human performance. The immune synapse is a highly complex interaction between several proteins and peptides that allows for a constant surveillance of foreign invaders. However, modeling these interactions is extremely difficult as the combinations of interactions is simply intractable. In immune-oncology, the study of this interaction is crucial as anti-tumor responses rely on sensitive and specific recognition of tumor-specific antigens. Implications of accurately predicting and modeling these interactions in immune-oncology range from improved and potent vaccine design to biomarkers for predicting response to immunotherapy to furthering our understanding of immune recognition.

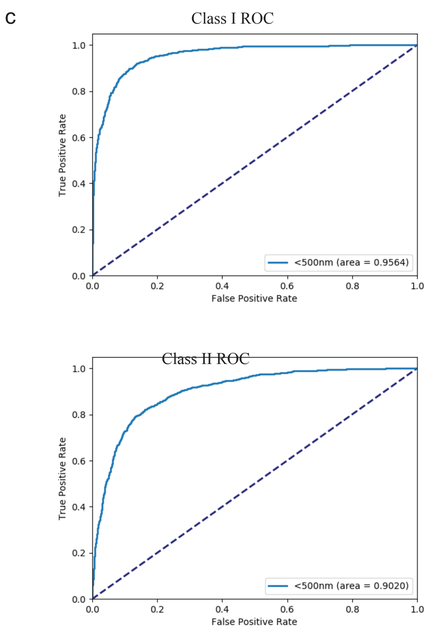
Our group has developed a variety of deep learning models to model the signal transmission within the immune synapse. We first present AI-MHC, an applied deep convolutional neural network for class-specific MHC binding algorithm that achieves state-of-the-art performance in both Class and Class II predictions. By incorporating 'meaning' of the allele within the network, we are able to model the interaction of allele and peptide within the context of a neural network. We take these concepts further in the development of DeepMANA, a deep learning framework which combines sequence-specific information about an allele/peptide pairing to not only predict binding affinity for any allele with a known protein sequence but also provide an antigen 'quality' score. We observe that in three immunotherapy clinical trials, these quality neoantigens are enriched in long-term survivors/responders. Finally, we present DeepTCR, a package of unsupervised and supervised deep learning algorithms to reveal structure in T-cell receptor sequencing that is predictive of various pathologies and therapies.

AI-MHC

The immune system has potential to present a wide variety of peptides to itself as a means of surveillance for pathogenic invaders. This means of surveillances allows the immune system to detect peptides derives from bacterial, viral, and even oncologic sources. However, given the breadth of the epitope repertoire, in order to study immune responses to these epitopes, investigators have relied on in-silico prediction algorithms to help narrow down the list of candidate epitopes, and current methods still have much in the way of improvement.

We present Allele-Integrated MHC (AI-MHC), a deep learning architecture with improved performance over the current state-of-the-art algorithms in human Class I and Class II MHC binding prediction. Our architecture utilizes a convolutional neural network that improves prediction accuracy by 1) allowing one neural network to be trained on all peptides for all alleles of a given class of MHC molecules by making the allele an input to the net and 2) introducing a global max pooling operation with an optimized kernel size that allows the architecture to achieve translational invariance in MHC-peptide binding analysis, making it suitable for sequence analytics where a frame of interest needs to be learned in a longer, variable length sequence. We assess AI-MHC against internal independent test sets and compare against all algorithms in the IEDB automated server benchmarks, demonstrating our algorithm achieves state-of-the-art for both Class I and Class II prediction.[1]





(a) AI-MHC is designed to take a peptide/allele pair which are transformed with either (1) convolutional layers or (2) trainable embedding layers learning vector representations of alleles and (3) amino acids. 1024 10-mer convolutions with global max pooling are applied to the sequence resulting in a [1,1024] feature map for each sequence. The sequence feature map is then concatenated to the [1,512] allele feature map. This long-form vector [1,1536] is then followed by 3 fully-connected layers with 50% dropout and utilizing leaky relu activations functions with a final output node with sigmoid activation.(b) Trained embedding layers were extracted from the network graph for amino acid and Class I embeddings and are visualized with clustermaps (c) ROC for Class I and Class II models.

DeepTCR

Deep learning algorithms have been utilized to achieve excellent performance in pattern-recognition tasks, such as in image and vocal recognition[3,4]. The ability to learn complex patterns in data has tremendous implications in the genomics world, where sequence motifs become learned 'features' that can be used to predict functionality, guiding our understanding of disease and basic biology[1,5–7]. T-cell receptor (TCR) sequencing assesses the diversity of the adaptive immune system, and while prior conventional biological sequence analysis tools have been insightful, they can miss signals in the data due to their rigidity[8-10]. We present DeepTCR, a broad collection of unsupervised and supervised deep learning methods able to uncover structure in highly complex and large TCR sequencing data. We demonstrate its utility across multiple basic science and clinical examples, including learning antigen-specific motifs, understanding immunotherapy-related shaping of repertoire, and learning a predictive signature in the peripheral blood of individuals with multiple sclerosis (MS). We further extract meaningful motifs from the trained network as a means of explaining the sequence concepts that have been learned to accomplish a given task. Our results show the flexibility and capacity for deep neural networks to handle the complexity of high-dimensional genomics data for both descriptive and predictive purposes.

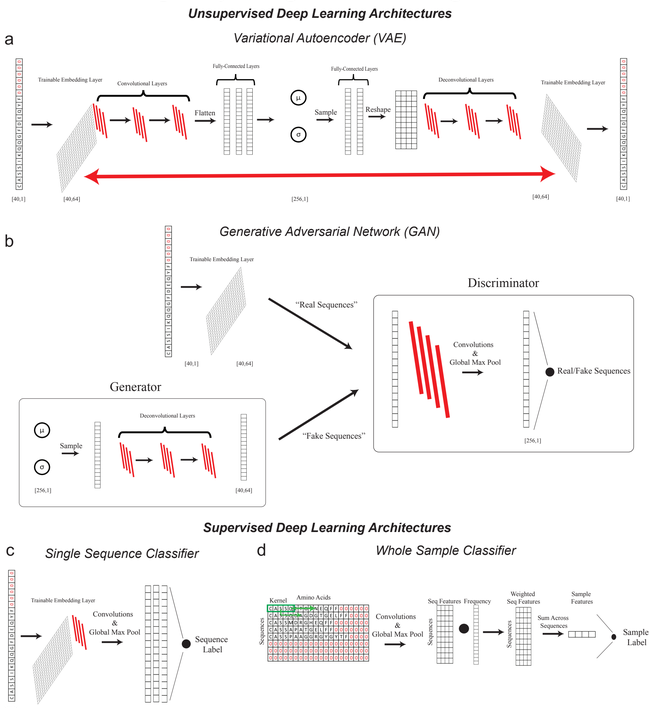


Figure 1. Deep Learning Architectures. (a) The variational autoencoder (VAE) is designed to take as a discrete input the amino acid sequence of the TCR sequence with a right zero-padding scheme. A trainable embedding layer is used to transform the sequence from discrete to continuous numerical domain. Convolutional and fully connected layers transform the sequence into a latent representation that is parametrized by a multi-dimensional unit gaussian. Reconstruction of the sequence occurs via fully connected and deconvolutional layers followed by the transposition of the same trainable embedding layer used at the beginning of the network. (b) The generative adversarial network (GAN) consists of the generator and discriminator, separate networks trained with separate objective functions. The generator samples from a multi-dimensional unit gaussian to create a 'fake' TCR sequence. The discriminator learns to distinguish 'real' from 'fake' sequences through one layer of convolutions with a global max pooling operation to provide translational invariance to the network. Of note, the generator's output is the continuous and not discrete representation of the TCR sequence. The latent space used for downstream analysis is the penultimate layer of the discriminator, here described as having dimensions of [256,1]. (c) The single sequence classifier follows a conventional convolutional neural network architecture consisting of one convolutional layer with global max pooling and three fully connected layers to a final classification layer. (d) The whole sample classifier utilizes a kernel that scans in a horizontal fashion across all sequences in the file resulting in a sequences-by-features tensor. This is then multiplied by the frequency vector for each sequence to derive weighted sequence features. These are then summed across the sequence space to compute sample level features that are fed into a classification layer.

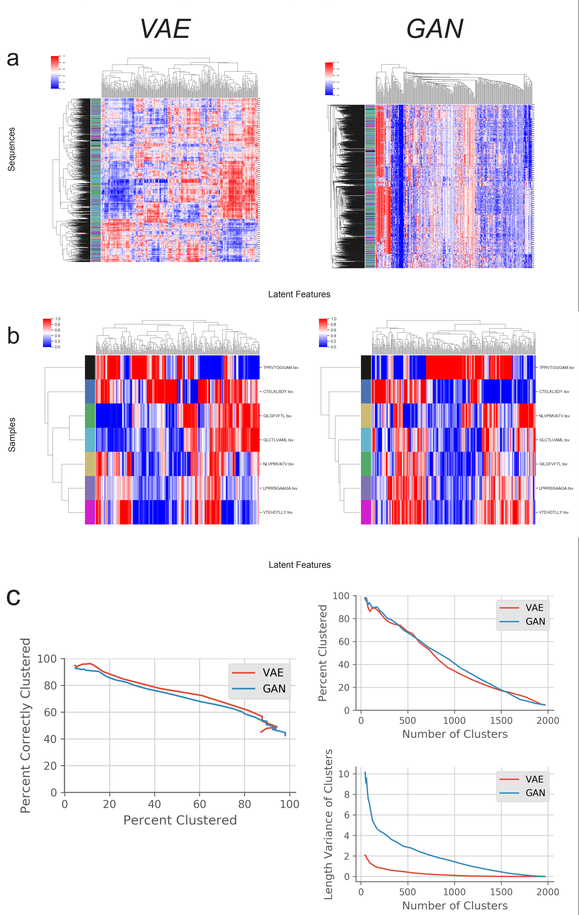
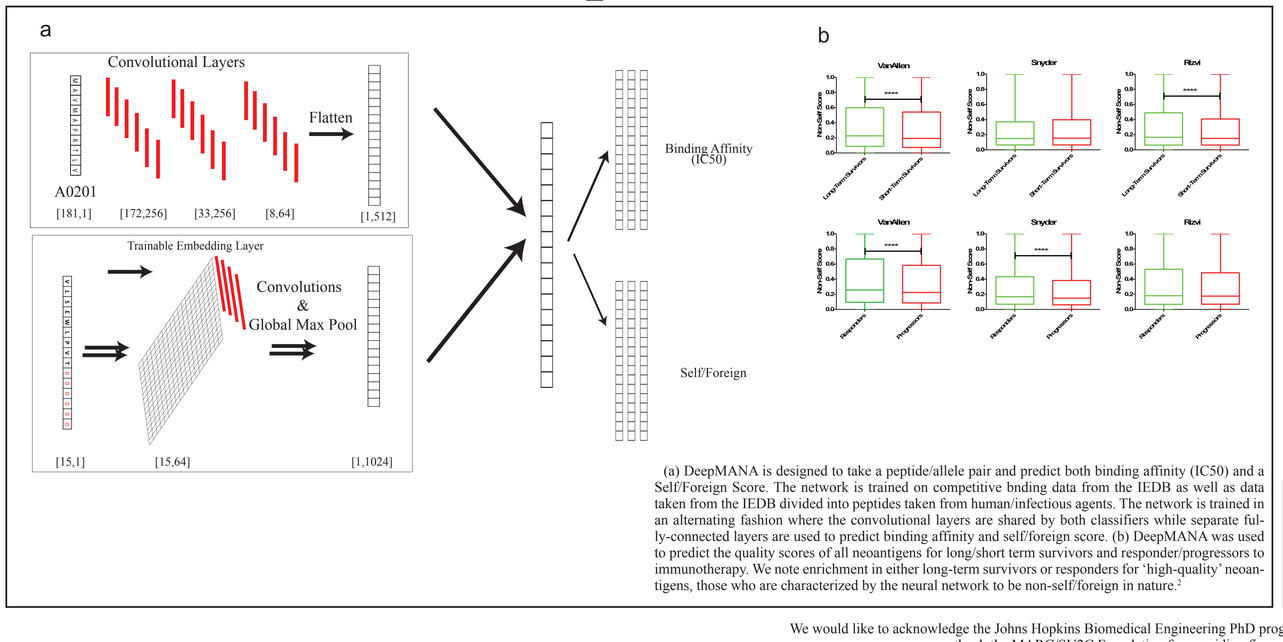
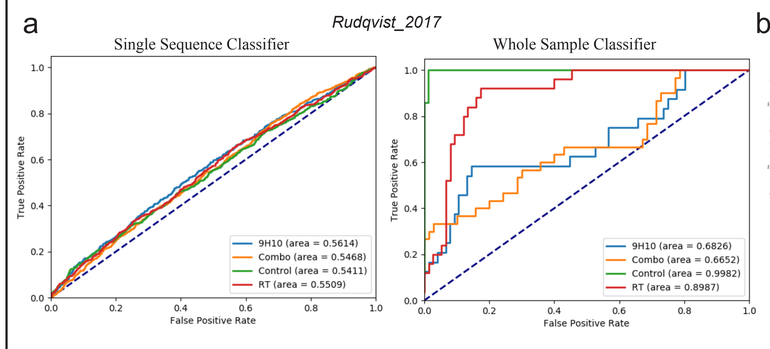


Figure 2. Unsupervised Learning Examples. (a) Heatmaps of sequences-by-features for anti-gen-specific sequences tetramer-sorted cells for 7 viral-specific antigens. (b) Heatmaps of samples-by-features for antigen-specific samples.(c) Clustering specificity by VAE & GAN (d) Characteristics of clustering solutions as applied to VAE & GAN

DeepMANA



(a) DeepMANA is designed to take a peptide/allele pair and predict both binding affinity (IC50) and a Self/Foreign Score. The network is trained on competitive bnding data from the IEDB as well as data taken from the IEDB divided into peptides taken from human/infectious agents. The network is trained in an alternating fashion where the convolutional layers are shared by both classifiers while separate fully-connected layers are used to predict binding affinity and self/foreign score. (b) DeepMANA was used to predict the quality scores of all neoantigens for long/short term survivors and responder/progressors to immunotherapy. We note enrichment in either long-term survivors or responders for 'high-quality' neoantigens, those who are characterized by the neural network to be non-self/foreign in nature.[2]



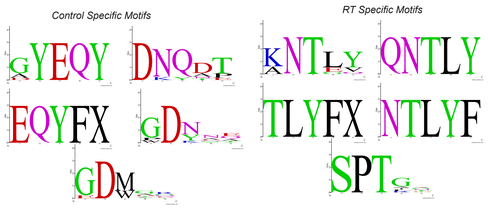


Figure 3. Supervised Learning Examples. (a) Receiver Operating Characteristic (ROC) curves for single sequence and whole sample classifier for murine TIL cohorts treated by various immunotherapies.[11] (b) Representative motifs learned by whole sample classifier for cohorts that had highly predictive structural signatures.

References

[1]Sidhom, J. W., Pardoll, D., & Baras, A. (2018). AI-MHC: an allele-integrated deep learning framework for improving Class I & Class II HLA-binding predictions. bioRxiv, 318881.

[2]Luksza, M., Riaz, N., Makarov, V., Balachandran, V. P., Hellmann, M. D., Solovyov, A., ... & Wolchok, J. D. (2017). A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature, 551(7681), 517.

[3]LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436 (2015).

[4]Krizhevsky, A., Sutskever, I. & in neural, H. G. Imagenet classification with deep convolutional neural networks. (2012).

[5]Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology 33, 831–838

[6]Zeng, H., Edwards, Liu, G. & Bioinformatics, G. D. Convolutional neural network architectures for predicting DNA–protein binding. (2016).

[7]Han, Y. & bioinformatics, K. D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. (2017).

[8]Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. Nature 547, 94–98 (2017).

[9]Sidhom, J.-W. et al. ImmunoMap: A Bioinformatics Tool for T-Cell Repertoire Analysis. Cancer Immunology Research 6, canimm.0114.2017 (2017).

[10]Dash, P. et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 547, 89–93 (2017).

[11]Rudqvist, N. P., Pilones, K. A., Lhuillier, C., Wennerberg, E., Sidhom, J. W., Emerson, R. O., ... & Demaria, S. (2018). Radiotherapy and CTLA-4 blockade shape the TCR repertoire of tumor-infiltrating T cells. Cancer Immunol Res, 6, 139-50.

We would like to acknowledge the Johns Hopkins Biomedical Engineering PhD program and the Johns Hopkins Medical Scientist Training Program for supporting this work. In addition to training support, we also thank the MARC/SU2C Foundation for providing financial support for developing algorithmic pipelines to understand responses to cancer immunotherapy.