bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

HLA class I binding prediction via convolutional neural networks

Yeeleng S. Vang and Xiaohui Xie
Department of Computer Science
University of California, Irvine, CA 92697, USA
{ysvang,xhx}@ics.uci.edu

Abstract

Many biological processes are governed by protein-ligand interactions. Of such is the recognition of self and nonself cells by the immune system. This immune response process is regulated by the major histocompatibility complex (MHC) protein which is encoded by the human leukocyte antigen (HLA) complex. Understanding the binding potential between MHC and peptides is crucial to our understanding of the functioning of the immune system, which in turns will broaden our understanding of autoimmune diseases and vaccine design.

We introduce a new distributed representation of amino acids, named HLA-Vec, that can be used for a variety of downstream proteomic machine learning tasks. We then propose a deep convolutional neural network architecture, named HLA-CNN, for the task of HLA class I-peptide binding prediction. Experimental results show combining the new distributed representation with our HLA-CNN architecture acheives state-of-the-art results in the vast majority of the latest two Immune Epitope Database (IEDB) weekly automated benchmark datasets. Codes are available at https://github.com/uci-cbcl/HLA-bind.

1 Introduction

The major histocompatibility complex (MHC) are cell surface proteins used to bind intracellular peptide fragments and display them on cell surface for recognition by T-cells [Janeway *et al.*, 2001]. In humans, the human leukocyte antigens (HLA) gene complex encodes these MHC proteins. HLAs displays a high degree of polymorphism, a variability maintained through the need to successfully process a wide range of foreign peptides [Jin *et al.*, 2003, Williams, 2001].

The HLA gene lies in chromosome 6p21 and is comprised of 7.6Mb [Simmonds *et al.*, 2007]. There are different classes of HLAs including class I, II, and III corresponding to their location in the encoding region. HLA class I is one of two, the other being class II, primary classes of HLA. Its function is to present peptides from inside cells to be recognized either as self or nonself as part of the immune system. Foreign antigens presented by class I HLAs attracts killer T-cells and provoke an immune response. Similarly, class II HLAs are only found on antigen-presenting cells, such as mononuclear phagocytes and B cells, and presents antigen from extracellular proteins [Ulvestad *et al.*, 1994]. Unlike class I and II, class III HLAs encode proteins important for inflammation.

The focus of this paper is on HLA class I proteins. As these molecules are highly specific, they are able to bind with only a tiny fraction of the peptides available through the antigen presenting pathway [Nielsen *et al.*, 2016, Yewdell, 1999]. This specificity makes binding to the HLA protein the most critical step in antigen presentation. Due to the importance of binding, accurate prediction models can shed understanding to adverse drug reactions and autoimmune diseases [Gebe *et al.*, 2002, Illing *et al.*, 2012], and lead to the design of more effective protein therapy and vaccines [Chirino *et al.*, 2004, van der Burg *et al.*, 2006].

Given the importance of MHC to the immune response, many algorithms have been developed

bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for the task of MHC-peptide binding prediction. The following list is by no means exhaustive but a small sample of previously proposed models. Wang et al. proposed using quantitative structure-activity relationship (QSAR) modelling from various amino acid descriptors with linear regression models [Wang *et al.*, 2015]. Kim et al. derived an amino acid similarity matrix [Kim *et al.*, 2009]. Luo et al. proposed both a colored and non-colored bipartite networks [Luo *et al.*, 2016]. Shallow and high-order artificial neural networks were proposed from various labs [Hoof *et al.*, 2009, Koch *et al.*, 2013, Kuksa, 2015, Nielsen *et al.*, 2003]. Of these approaches, NetMHC has been shown to achieve state-of-the-art for MHC-peptide binding prediction [Nielsen *et al.*, 2016].

In this article, we apply machine learning techniques from the natural language processing (NLP) domain to tackle the task of MHC-peptide binding prediction. Specifically, we introduce a new distributed representation of amino acids, named HLA-Vec, that maps amino acids to a 15-dimensional vector space. We combine this vector space representation with a deep convolutional neural network (CNN) architecture, named HLA-CNN, for the task of HLA class I-peptide binding prediction. Finally, we provide evidence that shows HLA-CNN achieves state-of-the-art results for the majority of different allele subtypes from the IEDB weekly automated benchmark datasets.

2 Methods

2.1 Dataset

To control for data pre-processing variabilities, we decided to use an existing post-processed training dataset so prediction algorithms could be more directly compared. The dataset used was filtered, processed, and prepared by Luo et al. [Luo *et al.*, 2016]. This dataset contained HLA class I binding data currated from four widely used, publicly available MHC datasets: IEDB [Vita *et al.*, 2015], AntiJen [Toseland *et al.*, 2005], MHCBN [Lata *et al.*, 2009], and SYFPEITHI [Rammensee *et al.*, 1999]. Target indicator indicating binding or non-binding was readily given as one of the column in the processed dataset. Peptides that contained unknown or indiscernible amino acids, denoted as "X" or "B", were removed from the dataset prior to training. Dataset was split into 70% training set and 30% validation set.

The test datasets were obtained from

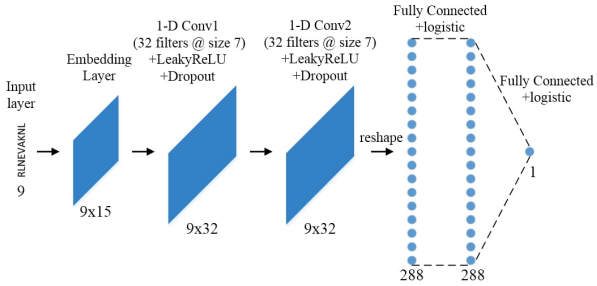

Figure 1: We illustrate our CNN architecture for MHC-peptide binding prediction of size 9-mers. The input is the peptide. The embedding layer substitues the individual amino acids with their 15-dimensional vector space representation. This is followed by two 1-dimensional convolutional layers preserving input length using 32 filters of size 7. The output of the 2nd convolutional layer is reshape into a 1-dimensional vector and is fully connected to the next layer of the same size. This fully connected layer is then fully connected to a logistic output unit. The architecture is generalizable to allele subtypes of any length.

IEDB automatic server benchmark page http://tools.iedb.org/auto_bench/mhci/weekly/). Allele subtypes with less than 500 training examples were excluded from testing. The lack of training data is a well-known weakness of deep neural networks as the model may not converge to a solution or worst yet, may overfit to the small training set. Indicators of binding were given has either binary values or ic50 (half maximal inhibitory concentration) measurements. Binary indicator was used directly. For test sets given in ic50 measurements, a standard threshold, ic50 ¡ 500 nM, is used to denote binding.

2.2 Distributed Representation

Distributed representation has been successfully used in NLP to train word embeddings, the mapping of words to real-value vector space representations. More generally, distributed representation is a means to represent an item by its relationship to other items. In word embeddings, this means semantically similar words are mapped near each other in the distributed representation vector space [Mikolov *et al.*, 2013]. The resulting distributed representation can then be used much like how BLOSUM is used for sequence alignment of proteins [Henikoff *et al.*, 1992] or peptide binding prediction by NetMHCpan [Andreatta

bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: HLA-Vec, an amino acids distributed representation.

Amino Acid	dim-1	dim-2	dim-3	dim-4	dim-5	dim-6	dim-7	dim-8	dim-9	dim-10	dim-11	dim-12	dim-13	dim-14	dim-15
A	-0.1428	0.1384	0.0579	-0.2463	0.3611	0.2930	-0.2692	-0.1532	0.0249	-0.0250	0.4739	-0.1261	-0.1988	0.0109	-0.0323
C	0.1210	-0.4359	-0.2869	-0.3863	0.2308	0.0282	0.0265	-0.0629	-0.0820	0.2711	0.2992	-0.3216	0.0350	-0.1972	0.0464
E	-0.0669	0.2425	0.0402	-0.2557	-0.0065	0.1357	0.0091	-0.2227	-0.2217	0.0975	0.4011	-0.0615	-0.0142	-0.2462	-0.1409
D	-0.0493	0.0821	-0.0815	-0.3480	0.0629	0.2017	0.2364	-0.1319	-0.0762	0.2678	0.1717	-0.1500	0.0241	-0.1535	-0.2351
G	-0.1611	0.0342	-0.1203	-0.1648	0.2604	0.0739	0.1568	-0.2627	0.0374	0.0746	0.2992	-0.3769	-0.1973	-0.0003	-0.1831
F	-0.1002	-0.0711	-0.2272	-0.1740	0.2519	0.1076	0.1501	-0.1994	-0.0486	0.0774	0.1696	-0.0822	0.2303	-0.1641	-0.2655
I	-0.0833	0.1719	-0.2545	-0.2451	0.1372	0.3516	0.0905	-0.0902	-0.1880	0.0864	0.0773	0.0309	0.1380	-0.2591	-0.1420
H	-0.1433	-0.0003	-0.0744	-0.1195	0.3056	0.1037	0.0642	-0.0514	-0.1960	0.2619	0.1837	-0.2322	0.1123	-0.0715	-0.2034
K	0.0276	0.1958	-0.2127	-0.1873	0.0951	0.0930	0.0423	-0.2705	-0.1871	0.2797	0.2675	0.0294	0.0429	-0.2417	-0.0357
M	-0.1812	0.1389	-0.1602	-0.1305	0.3983	0.2286	-0.2404	-0.1206	0.1616	0.3216	0.0079	-0.2107	-0.0639	-0.2552	-0.0892
L	-0.0360	0.0927	-0.1824	-0.1546	0.2046	-0.0066	-0.0146	-0.2175	-0.1835	0.2056	0.4694	0.0770	0.0804	-0.1424	0.0520
N	-0.1220	0.1053	-0.1168	-0.4092	-0.0007	0.2341	0.0629	-0.0610	-0.1055	0.4091	0.1039	0.0990	0.2457	-0.1396	-0.2956
Q	-0.5482	0.0352	0.1479	-0.0171	-0.1638	0.0976	-0.0539	-0.3098	-0.1891	0.0823	0.3988	-0.1479	0.0769	-0.2872	-0.1089
P	-0.3671	-0.1098	-0.0392	0.0031	0.2176	0.3222	0.1557	-0.1623	0.0569	0.1854	0.2744	-0.2758	0.2775	0.0526	0.0957
S	-0.0774	-0.0416	-0.2532	-0.1159	0.2320	0.0761	-0.0995	-0.2774	-0.0892	0.2454	0.1238	-0.1930	0.0999	-0.1710	-0.1671
R	0.3854	0.1272	-0.3518	-0.1442	0.2487	0.0564	0.1701	-0.1434	0.1015	-0.0507	0.2773	-0.0669	0.2507	-0.0338	-0.0685
T	-0.0935	0.0087	-0.1558	-0.1983	0.2365	0.2426	0.0244	-0.0749	-0.1608	0.0807	0.2357	-0.1303	0.1860	-0.1256	-0.0830
W	-0.4829	-0.0159	0.0106	0.0676	0.3279	-0.1073	-0.0050	-0.1282	-0.1045	-0.0425	0.1982	-0.2086	-0.0252	-0.4396	-0.3651
V	-0.11540	.0944	-0.1744	-0.0475	0.2863	0.3909	0.1128	-0.1018	-0.1815	0.0061	0.1972	-0.1604	0.0812	-0.2151	0.1363
Y	-0.1308	-0.0410	-0.1395	0.0534	0.3133	0.2197	0.1469	-0.1309	-0.3230	0.2696	0.0919	-0.0462	0.0193	-0.2942	-0.0820

This table lists the 15-dimensional vector space distributed representation of amino acid trained unsupervised on HLA class I peptides of all allele subtypes and lengths from the training dataset. The dimensions are arbitrary and have no physicochemical interpretation.

et al., 2015]. That is, we encode amino acids with their vector space distributed representation to be useable by downstream machine learning algorithms.

Recently, distributed representation had been explored for bioinformatics applications. Specifically, 3-gram protein distributed representation of size 100-dimensions was used to encode proteins for protein family classification and identifying disordered sequences, resulting in state-of-the-arts performance [Asgari *et al.*, 2015]. The distributed representation was further showed to grouped 3-gram proteins with similar physicochemical property closer to each other by mapping the 100-dimensional space to 2-D.

The two main neural probabilistic language models commonly used to train a distributed representation are the continuous skip-gram model and continuous bag-of-words (CBOW) model [Mikolov *et al.*, 2013]. Both models are similar and is often thought of as inverse of one another. In the skip-gram model, the adjacent context-words are predicted based on the center (target) word. Conversely in the CBOW model, the center word is predicted based on adjacent context-words. In this paper, the skip-gram model is used. The interested reader is encourage to consult the relevant references for further details of the CBOW model.

A short overview of the skip-gram model is given here for completeness. As originally formulated by Mikolov et al. [Mikolov *et al.*, 2013], in the skip-gram model, given a sequence of words $w_1, w_2, ..., w_T$, the objective is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_j) \qquad (1)$$

where c is the sliding window size and $p(w_{t+j}|w_t)$ is defined as:

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\mathsf{T}} v_{w_I}\right)}{\sum_{w=1}^{W}\exp({v'_w}^{\mathsf{T}} v_{w_I})} \qquad (2)$$

Here, v_w and v'_w are two vector space representations of the word w. The subscripts O and I correspond to the output (context-words) word and input (target) word respectively. W is the total number of unique words in the vocabulary. In typical NLP text corpus with large vocabulary, calculating the gradient of the log probability becomes impractical. An approximation to the log probability is obtained by replacing every $\log p(w_O|w_I)$ with

$$\log\sigma({v'_{w_O}}^{\mathsf{T}} v_{w_I}) + \sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}[\log\sigma({v'_{w_i}}^{\mathsf{T}} v_{w_I})] \qquad (3)$$

where $\sigma(x) = 1/(1+\exp(-x))$ and k are negative samples. This was motivated by the idea that a good model should be able to differentiate real data from false (negative) ones.

bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

By formulating protein data as standard sequence data like sentences in a text corpus, standard NLP algorithms can be readily applied. More concretely, individual peptides are treated as individual sentences and amino acids are treated as words. In this paper, the skip-gram model is used with 5 negative samples and the projected vector space of 15-dimensions is chosen. The 15-dimensional vector space distributed representation, HLA-Vec, is summarized in Table 1.

2.3 Convolutional neural network

Convolutional neural networks (CNN) have been studied since the late 1980s and have made a comeback in recent years along with the renewed interested in artificial neural networks, and in particular of the deep architecture varieties. Much of the recent fervor is spurned in part by both accessibility to large training datasets consisting of over millions of training examples and advances in cheap computing power needed to train these deep network architectures in a reasonable amount of time. Although originally proposed for the task of image classification [LeCun *et al.*, 1989, Krizhevsky, 2012, Simonyan *et al.*, 2014], CNN have been found to work well for general sequence data such as sentences [Kalchbrenner *et al.*, 2014, Kim, 2014]. It is with this insight that we propose a convolutional neural network for the task of MHC-peptide binding prediction.

The CNN architecture we propose in this paper consists of both convolutional and fully connected (dense) layers. Convolutional layers preserve local spatial information [Taylor *et al.*, 2010] and thus is well suited for studying peptides where spatial locations of the amino acids are critical for bonding.

Our CNN model, dubbed HLA-CNN, can be seen in Fig. 1. The input into HLA-CNN network is the character string of the peptide, a 9-mer peptide in this example. The input feeds into the embedding layer that substitutes each amino acid with their 15-dimensional vector space representation. The output encoding is a 2-dimensional matrix of size 9x15. The vector space matrix is then 1-dimensionally convolved with 32 filters of size 7 and returning the same output length as input, resulting in a matrix of size 9x32. The activation unit use is the leaky rectified linear units (LeakyReLU) with default learning rate of 0.3. LeakyReLU is similar to rectified linear units except there is no zero region which results in non-zero gradient over the entire domain [Maas *et al.*, 2013]. Dropout is used after each of the convolutional layers. Dropout acts as regularization to prevent overfitting by randomly dropping a percentage of the units from the CNN during training [Srivastava *et al.*, 2014]. This has the effect of preventing co-adaptation between neurons, the state where two or more neurons detect the same feature. In our architecture, the dropout percentage is set to 25%. The output then feeds into a second convolutional layer with the same filter size, activation unit, and dropout as the first convolutional layer. The 9x32 matrix outputted by the second convolutional layer is reshaped into a single 1-D vector of size 288 which is fully connected to another layer of the same size with sigmoid activation units. This dense layer is then fully connected to a logistic regression output unit to make a prediction.

The loss function used is the binary cross entropy function and the optimizer used was the Adam optimizer with learning rate 0.004. We used a variable batch size instead of a fixed one, choosing instead to force all allele subtypes to have 100 batches no matter the size of the training set. The convolutional layers's filters are initialized by scaling a random Gaussian distribution by the sum of edges coming in and going out of those layers [Glorot *et al.*, 2010]. Finally, the embedding layer of HLA-CNN is initialized to the previously learned HLA-Vec distributed representation with the caveat that the embedding layer is allowed to be updated during the supervised binding prediction training for each allele subtypes. This allows for the distributed representation to be fined-tuned for each allele subtypes uniquely and for the task of peptide binding specifically. The number of epoch was less important as we arbitrarily set max epoch to 100 but enforce early stoppage if the loss function stops improving for 2 epochs. Solutions were found to have converged under 40 epochs for all test sets.

The dataset was most abundant in 9-mer HLA-A*02:01 allele (10547 samples) therefore this specific 9-mer subtype is used for network architectural design and hyperparameter tuning. Dataset split of 70% training and 30% validation was used to determine the optimal architecture and hyperparamters. While the network architecture was designed using a single allele subtype of length

bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2: Performance comparison of NetMHCpan, sNebula, and HLA-CNN on IEDB datasets.

Dataset	IEDB	HLA	Peptide length	Peptide count	Measurement	NetMHCpan		sNebula		HLA-CNN	
						SRCC	AUC	SRCC	AUC	SRCC	AUC
2016-05-03/2016-02-19	1029824	A*02:01	9	77	binary	0.071	0.546	0.060	0.539	**0.178**	**0.560**
2015-08-07	1029125	B*27:05	9	21	binary	**0.751**	**0.959**	**0.752**	**0.959**	0.571	0.898
2015-08-07	1029061	B*57:01	9	26	ic50	**0.612**	**0.943**	0.169	0.575	0.395	0.841
2015-08-07	1028928	A*02:01	9	13	binary	**0.570**	**0.955**	0.539	0.909	**0.778**	**0.955**
2015-08-07	1028928	B*07:02	9	12	binary	**0.648**	**1.000**	0.522	0.900	**1.000**	**1.000**
2015-08-07	315174	B*27:03	9	11	binary	0.657	0.893	0.179	0.607	**1.000**	**1.000**
2015-08-07	1028790	A*02:01	9	55	ic50	0.615	0.574	**0.505**	**0.778**	0.228	0.590
2015-08-07	1028790	A*02:01	10	35	ic50	0.407	0.677	**0.432**	**0.704**	0.269	0.653
2015-08-07	1028790	B*02:02	9	55	ic50	0.582	0.713	0.372	0.680	**0.527**	**0.796**
2015-08-07	1028790	A*02:03	9	55	ic50	0.539	0.696	0.477	0.629	**0.550**	**0.777**
2015-08-07	1028790	A*02:03	10	35	ic50	0.208	0.750	0.419	0.697	**0.632**	**0.888**
2015-08-07	1028790	A*02:06	9	55	ic50	0.630	0.770	**0.510**	**0.848**	0.716	0.836
2015-08-07	1028790	A*02:06	10	35	ic50	0.572	0.768	0.525	0.680	**0.742**	**0.866**
2015-08-07	1028790	A*68:02	9	55	ic50	0.534	0.806	0.482	0.713	**0.824**	**0.909**
2015-08-07	1028790	A*68:02	10	35	ic50	0.272	0.620	0.591	0.813	**0.749**	**0.954**

These benchmark datasets came from IEDB and encompasses the two most recent datsets. Allele subtypes with fewer than 500 training examples were excluded from these test datasets. SRCC stands for Spearman's rank correlation coefficient and AUC stands for area under the receiver operating characteristic curve.

9, HLA-CNN framework is robust enough to accept and make prediction for allele subtypes of any length.

Each test datasets of different allele subtypes and peptide lengths are treated as completely separate tests. For a specific test dataset, the training dataset is filtered on the allele subtype and peptide length. The resulting smaller training subset is then used to train the HLA-CNN model. Due to the random nature of initialization in the deep learning software framework used, five prediction scores are made for each test sets. The final prediction used for evaluation purposes is taken as the average predicted score of the five predictions. Two commonly used evaluation metric for peptide binding prediction task are the Spearman's rank correlation coefficient (SRCC) and area under the receiver operating characteristic curve (AUC). The state-of-the-art NetMHCpan [Andreatta *et al.*, 2015], a shallow feed forward neural network, and a more recently developed bipartite network-based algorithm, sNebula [Luo *et al.*, 2016], will be used to compared the performance of our proposed HLA-CNN prediction model.

3 Results

We have introduced the HLA class I dataset. We formulated the HLA class I peptide data as an equivalence of text data used in NLP machine learning tasks. We have proposed a model to learn a vector space distributed representation of amino acids from this HLA class I dataset. We have also describe our deep learning method and how it takes advantage of this new distributed representation of amino acids to solve the problem of HLA class I-peptide binding prediction. Next, we show the result of the learned distributed representation followed by the performance of our model against the state-of-the-art prediction model and another recently developed model.

3.1 Distributed Representation

The 15-dimensional distributed representation of amino acids is shown in Table 1. Each of the 15 dimensions have no corresponding physicochemical equivalence or interpretation. They are simply the result of the algorithm and our choice of size for the representation. Various other dimensional size were explored, however, 15-dimensions gave the best results on 10-fold cross-validation of HLA-A*02:01 subtype.

To understand the learned, 15-dimensional HLA-Vec distributed representation of the twenty amino acids, we visualize this vector space in 2-D using a dimension reduction techniqued called t-distributed stochastic neighboring embedding (t-SNE) [Maaten *et al.*, 2008]. t-SNE is capable of preserving local structure of the data, e.g. points closer to each other in the high-dimensional space are grouped closer together in the low 2-D space.

In Fig 2, we see the 2-D mapping of HLA-Vec colored by various physicochemical properties, including hydrophobicity, normalized van der waals volume, polarity, mass, volume, and net charge [Asgari *et al.*, 2015] from the Amino acid Physicochemical properties Database (APDbase)

bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

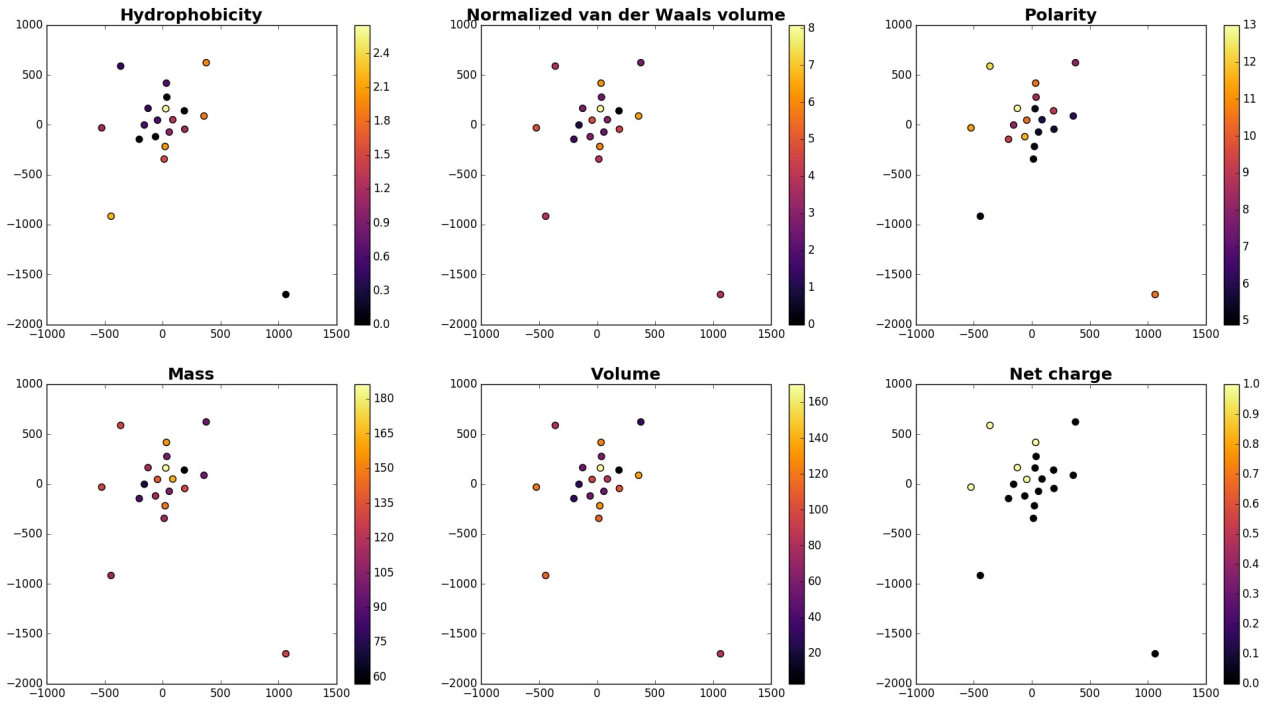

Figure 2: In these plots, each point represents the 2-D mapping of an amino acid from the 15-dimensional distributed representation using t-SNE. The color indicates the scale of each physicochemical property.

[Mathura *et al.*, 2005]. As can be seen, hydrophobicity, polarity, and net charge, factors important for covalent chemical bonding, can be visually distinguished into groups. This gives validation to distributed representation as an effective method to encode amino acids that also preserves important physicochemical properties.

3.2 HLA-peptide binding prediction

The results of our HLA-CNN prediction model against NetMHCpan and sNebula on the two latest IEDB benchmarks are shown in Table 2. As AUC is a better measure of the goodness of binary predictors compared to SRCC, for evaluation purposes between models, we say one algorithm is superior to another if it scores higher on the AUC metric.

On these latest IEDB benchmark datasets, our algorithm achieved state-of-the-art results in 10 of 15 (66.7%) test datasets. This is in contrast to NetMHCpan, which acheived state-of-the-art results in only 4 out of 15 (26.7%) and sNebula in 4 out of 15 (26.7%). In many of the allele subtypes, our algorithm achieved significant AUC improvements (greater than 10%) over the two existing models. In fact, for B*07:02 and B*27:03 subtypes, our model achieved perfect SRCC and AUC scores. In the 10 test sets where our model achieved state-of-the-art results, our model averaged a 7.8% improvement over the previous state-of-the-art. SRCC performance also increased on the vast majority of test sets as well.

In Fig 3, the ROC curves are shown for all five predictions of the HLA-A*68:02 9-mer subtype as an example of the improvement our model gives over the previous state-of-the-part. As can be seen, all five curves are outperforming NetMHCpan's curve at almost all thresholds.

The results suggests that HLA-CNN can accurately predict HLA class I-peptide binding and outperforms the current state-of-the-art algorithms. The results also confirmed that the hyperparamters of HLA-CNN learned on the HLA-A*02:01 9-mer subtype generalizes well to cover a variety of other allele subtypes and peptide lengths. This validates the robustness of our algorithm as different networks did not have to be specifically design for each allele subtypes.

bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

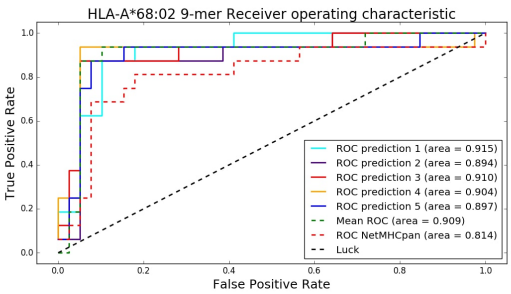

Figure 3: The ROC curves for HLA-A*68:02 9-mer test set is shown for all five predictions and their mean compared against NetMHCpan. Our model shows improvement across all five predictions and the average prediction.

4 Conclusion

In this work, we have described how machine learning techniques from the NLP domain could be applied to HLA class I-peptide binding prediction. We presented a method to extract a vector space distributed representation of amino acids from HLA peptide data that preserved property critical for covalent bonding. Using this vector space representation, we proposed a deep CNN architecture for the purpose of HLA class I-peptide binding prediction. This framework is capable of making prediction f

While our network established new state-of-the-art results for many of these test sets, by no means does it mean this is the best our model could ever do on these datasets. As with any artificial neural network-based machine learning algorithms, our model will improve as more training data becomes available. Therefore, as genomic researchers publish more HLA class I peptide data, our algorithm can be re-trained at a future date and improved results can be expected.

On future work, allele-specific affinity thresholds instead of a general binding affinity ic50 threshold of 500 nM can be used to identify peptide binders in different subtypes. This approach had shown superior predictive efficacy in previous work [Paul *et al.*, 2013]. From an architecture design standpoint, one possibility to extend the network is to replace the dense layer with a convolutional layer, thereby creating a fully convolutional network (FCN). The motivation being since convolutional layers preserve spatial information in the peptide, perhaps a FCN could improve performance over the existing network if all layers in the network had this capability. Another option is to tackle this peptide binding problem using more advanced NLP artificial neural networks such as Long Short Term Memory (LSTM) recurrent neural networks models which has the ability to remember values for long sequences. This would allow us to train a single model to classify binding prediction for peptides of any subtype and length.

References

Andreatta, M. and Nielsen, M. (2015) Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, p.btv639.

Asgari, E. and Mofrad, M. R. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PloS one*, *10*(11), e0141287.

Chirino, A.J., Ary, M.L. and Marshall, S.A. (2004) Minimizing the immunogenicity of protein therapeutics. *Drug discovery today, 9*(2), pp.82-90.

Gebe, J.A., Swanson, E., and Kwok, W.W. (2002) HLA Class II peptidebinding and autoimmunity. *Tissue antigens, 59*(2), pp.78-87.

Glorot, X. and Bengio, Y. (2010 May) Understanding the difficulty of training deep feedforward neural networks. In *Aistats* (Vol. 9, pp. 249-256).

Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), pp.10915-10919.

Hoof, I., Peters, B., Sidney, J., Pedersen, L. E., Sette, A., Lund, O., Buus, S., and Nielsen, M. (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, **61**(1), pp.1-13.

Illing, P. T., Vivian, J. P., Dudek, N. L., Kostenko, L., Chen, Z., Bharadwaj, M., Miles, J. J., Kjer-Nielsen, L., Gras, S., Williamson, N.A ., and Burrows, S. R., 2012. Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature, 486*(7404), pp.554-558.

Janeway, C. A., Jr, Travers,P., Walport,M., et al. (2001) Antigen Presentation to T Lymphocytes. *Immunobiology: The Immune System in Health and Disease*. 5th edn. Garland Science, New York.

Jin, P. and Wang, E. (2003) Polymorphism in clinical immunology-from HLA typing to immunogenetic profiling, *Journal of translational medicine*, **1:8**, doi:10.1186/1479-5876-1-8.

bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Kim, Y., Sidney, J., Pinilla, C., Sette, A. and Peters, B. (2009) Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC bioinformatics*, *10*(1), p.1.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kim, Y., Sidney, J., Pinilla, C., Sette, A. and Peters, B. (2009) Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC bioinformatics, 10*(1), p.1.

Koch, C. P., Perna, A. M., Pillong, M., Todoroff, N. K., Wrede, P., Folkers, G., Hiss, J. A., and Schneider, G. (2013) Scrutinizing MHC-I binding peptides and their limits of variation. *PLoS Comput Biol, 9*(6), p.e1003088.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems.* (pp. 1097-1105).

Kuksa, P. P., Min, M. R., Dugar, R. and Gerstein, M. (2015) High-order neural networks and kernel methods for peptide-MHC binding prediction. *Bioinformatics*, p.btv371.

Lata, S., Bhasin, M., and Raghava, G. P. (2009) MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res.*, *Notes* **2**, 61, doi:10.1186/1756-0500-2-61

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), pp.541-551.

Lundegaard, C., Lund, O., and Nielsen, M. (2008) Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*, *24*(11), pp.1397-1398.

Luo, H., Ye, H., Ng, H. W., Shi, L., Tong, W., Mattes, W., Mendrick, D., and Hong, H. (2015). Understanding and predicting binding between human leukocyte antigens (HLAs) and peptides by network analysis. *BMC bioinformatics 16*, *Suppl 13*, **S9**

Luo, H., Ye, H., Ng, H. W., Sakkiah, S., Mendrick, D. L., and Hong, H. (2016) sNebula, a network-based algorithm to predict binding between human leukocyte antigens and peptides. *Scientific Reports*, **6**

Maas, A.L., Hannun, A.Y., and Ng, A.Y. (2013 June). Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML* (Vol. 30, No. 1).

Maaten, L.V.D. and Hinton, G. (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(Nov), pp.2579-2605.

Mathura, V.S. and Kolippakkam, D. (2005) Apdbase: Amino acid physico-chemical properties database. *Bioinformation, 1*(1), pp.2-4.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013) Distributed representations of words and phrases and their compositionality. *Advances in neural informational processing systems.*, 3111-3119

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013) Efficient estimation of word representations in vector space. *ICLR Workshop*.

Nielsen, M., Lundegaard, C., Worning, P., Lauemller, S. L., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2003) Reliable prediction of Tcell epitopes using neural networks with novel sequence representations. *Protein Science*, *12*(5), pp.1007-1017.

Nielsen, M. and Andreatta, M. (2016) NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome medicine, 8*(1), p.1.

Paul, S., Weiskopf, D., Angelo, M.A., Sidney, J., Peters, B., and Sette, A. (2013) HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *The Journal of Immunology, 191*(12), pp.5831-5839.

Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., and Stevanoic, S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50**, 213-219

Simmonds, M.J. and Gough, S.C.L. (2007) The HLA region and autoimmune disease: associations and mechanisms of action. *Current genomics, 8*(7), pp.453-465.

Simonyan, K. and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), pp.1929-1958.

Taylor, G.W., Fergus, R., LeCun, Y., and Bregler, C. (2010) Convolutional learning of spatio-temporal features. In *European conference on computer vision* (pp. 140-153). Springer Berlin Heidelberg.

Toseland, C. P., Clayton, D. J., McSparron, H., Hemsley, S. L., Blythe, M. J., Paine, K., Doytchinova, I. A., Guan, P., Hattotuwagama, C. K. and Flower, D. R. (2005) AntiJen: a quantitative immunology

bioRxiv preprint doi: https://doi.org/10.1101/099358; this version posted January 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res.*, **1**,4, doi:10.1186/1745-1-4

Ulvestad, E., Williams, K., B, L., Trapp, B., Antel, J., and Mrk, S. (1994) HLA class II molecules (HLA-DR,-DP,-DQ) on cells in the human CNS studied in situ and in vitro. *Immunology, 82*(4), p.535.

van der Burg, S. H., Bijker, M. S., Welters, M. J., Offringa, R. and Melief, C. J. (2006) Improved peptide vaccine strategies, creating synthetic artificial infections to maximize immune efficacy. *Advanced drug delivery reviews, 58*(8), pp.916-930.

Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, Jtacl. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A. and Peters, B. (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, **43**, D405-D412.

Wang, Y., Zhou, P., Lin, Y., Shu, M., Hu, Y., Xia, Q., and Lin, Z. (2015) Quantitative prediction of class I MHC/epitope binding affinity using QSAR modeling derived from amino acid structural information. *Combinatorial chemistry & high throughput screening, 18*(1), pp.75-82.

Williams, T. M. (2001) Human leukocyte antigen gene polymorphism and the histocompatibility laboratory, *The Journal of Molecular Diagnostics*, **3.3**, 98-104.

Yewdell, J.W. and Bennink, J.R. (1999) Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses 1. *Annual review of immunology, 17*(1), pp.51-88.