OnionMHC: A Deep Learning Model for Peptide - HLA-A*02:01 Binding Predictions using both Structure and Sequence Feature Sets

Shikhar Saxena
Nanyang Technological University
Sambhavi Animesh
National University of Singapore
Melissa Fullwood
Nanyang Technological University
Yuguang Mu (ygmu@ntu.edu.sg)
Nanyang Technological University

Research Article

Keywords: Peptide-MHC binding, Structure, Sequence, Binding Affinity, Deep Learning

Posted Date: December 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-124695/v1

OnionMHC: a deep learning model for peptide - HLA-A*02:01 binding predictions using both structure and sequence feature sets

Shikhar Saxena[1], Sambhavi Animesh[2], Melissa J. Fullwood[1,2], and Yuguang Mu*[1]

[1] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551

[2] Cancer Science Institute, National University of Singapore, 14 Medical Drive, Singapore, 117599.

*Corresponding author: ygmu@ntu.edu.sg

Abstract:

Background:

The peptide binding to Major Histocompatibility Complex (MHC) proteins is an important step in the antigen-presentation pathway. Thus, predicting the binding potential of peptides with MHC is essential for the design of peptide-based therapeutics. Most of the available machine learning-based models predict the peptide-MHC binding based on the sequence of amino acids alone. Given the importance of structural information in determining the stability of the complex, here we have utilized both the complex structure and the peptide sequence features to predict the binding affinity of peptides to human receptor HLA-A*02:01. To our knowledge, no such model has been developed for the human HLA receptor before that incorporates both structure and sequence-based features.

Results:

We have applied machine learning techniques through the natural language processing (NLP) and convolutional neural network to design a model that performs comparably with the existing state-of-the-art models. Our model shows that the information from both sequence and structure domains results in enhanced performance in the binding prediction compared to the information from one domain alone. The testing results in 18 weekly benchmark datasets provided by the Immune Epitope Database (IEDB) as

well as experimentally validated peptides from the whole-exome sequencing analysis of the breast cancer patients indicate that our model has achieved state-of-the-art performance.

Conclusion:

We have developed a deep-learning model (OnionMHC) that incorporates both structure as well as sequence-based features to predict the binding affinity of peptides with human receptor HLA-A*02:01. The model demonstrates state-of-the-art performance on the IEDB benchmark dataset as well as the experimentally validated peptides. The model can be used in the screening of potential neo-epitopes for the development of cancer vaccines or designing peptides for peptide-based therapeutics. OnionMHC is freely available at https://github.com/shikhar249/OnionMHC

Keywords: Peptide-MHC binding; Structure; Sequence; Binding Affinity; Deep Learning

Background:

Neo-epitopes are the unique peptide signatures produced by cancer cells, some of which are presented on their outer surface and recognized by T cells [1, 2]. For the successful presentation of these antigenic peptides, the peptide should be able to bind to the Major Histocompatibility Complex (MHC) receptor with enough strength (IC_{50} < 500nM) in most cases. This makes peptide binding to MHC the most crucial and selective step in antigen presentation[3]. Thus, there is a need to develop computational algorithms that can accurately predict the binding affinity of peptides to MHC.

There have been several approaches made in this regard, mainly in terms of allele-specific or pan-specific models. In allele-specific models, the model was trained on a specific allele[4] whereas in a pan-specific version the model was trained on multiple alleles data [5]. These models were mainly based on either the sequence-based features[5, 6] or the structure-based features[7, 8]. Generally, most of the previous

approaches employed the sequence-based features for predicting the binding affinity. Different neural network models like convolutional neural network models [6, 9, 10], attention-based neural network models [11] have been employed with different ways of sequence information encoding. Other studies have used the structure-based features only for predicting the binding affinity [7, 12]. The information provided by the peptide-MHC 3D structures can potentially bring to light yet unknown drivers of T-cell activation thus, facilitating more accurate predictions. However, this direction is progressing at a very slow pace, primarily because 1) a very limited number of experimentally determined peptide-MHC complex structures are available; and 2) the structural modeling is computationally expensive, rendering it feasible only on a small scale. To our knowledge, there is no such model that combines both structure and sequence-based features to predict the binding affinity of peptides with human MHC receptors. Recently a study demonstrated the potential of combined structure and sequence feature sets in predicting the binding affinity of peptides to mouse allele H-2^D [13]. However, this study used a very simple machine learning algorithm of the support vector machines which is not sophisticated enough to capture all the intrinsically complex patterns. Furthermore, the authors have not made the model publicly available, so it is difficult to compare with their model.

In this study, we have used both complex structure and sequence feature sets to predict the binding affinity of the peptides to widely studied human HLA-A*02:01 alleles. The structural features we used are the contact pairs between the atoms in the peptide and the MHC receptor. The contact pairs based structural features are known to perform well for the small molecule and protein complex in the PDBbind dataset[14]. Here, the similar contact-pairs based structural features for the peptide-MHC complex were combined with sequence-based features. Different encodings of raw peptide sequences and different architectures, such as Long Short-Term Memory (LSTM) based neural network models, were tested. Our model's performance is comparable and even slightly better than the current state-of-the-art model NetMHCpan4.0[5], HLAthena [15], and the convolutional neural network model ConvMHC [9] on the

independent IEDB benchmark dataset. All these models are the sequence-based models that take solely the sequence of peptides to make the binding affinity predictions. We have also performed molecular dynamics simulations to demonstrate the stability of the predicted models. Moreover, our model has been used to predict the binding affinity of the peptides derived from the neoantigen analysis of the exome sequencing samples from the breast cancer patients which were further validated experimentally. The exome sequencing and neoantigen analysis will be reported in an upcoming manuscript.

OnionMHC is publicly available and can be accessed at https://github.com/shikhar249/OnionMHC

Methods:

Dataset collection

The peptide sequences and their binding affinities to HLA-A*02:01 were downloaded from IEDB [16](http://tools.iedb.org/main/datasets/). This dataset was compiled by combining BD2009, BD2013, and the blind data [17] and it also included additional data that was not publicly available at that time. The detailed dataset used for the training is provided in Table S3, Additional File 4. Although the HLA-A*02:01 receptors can bind to peptides of varying lengths, we only considered 9 mers for this study as 9 mers are the most preferred binders to HLA-A*02:01. The dataset contains 9051 peptides of length 9 residues with their corresponding IC_{50} values with HLA-A*02:01 receptors. The peptides with IC_{50} less than 500 nM were considered as the positive binders rendering 3273 peptides as positive and 5778 peptides as negative towards binding to HLA-A*02:01. The same dataset from IEDB was utilized to train several other peptide-MHC binding affinity predicting models[5, 9-11, 18, 19] , thus making comparison straightforward. The comparison of these binding affinity predicting models was done on the weekly benchmark data provided by IEDB (http://tools.iedb.org/auto_bench/mhci/weekly/) which is different from the dataset used for training.

Structure Modelling

The structure of HLA-A*02:01 with the template peptide backbone of 9 residues was taken from PDB[20](PDB ID: 1AKJ). The nonameric peptides from PDB structures show high conservation in their backbones[7]. The side chains of different peptides were modeled on the same template structure using SCWRL4 software[21]. It has been shown that SCWRL4 can reproduce experimentally determined X-ray structures of peptide – HLA complexes with the lowest root mean square deviation[22]. The structures obtained from SCWRL4 were further energy minimized with the default settings in GROMACS[23] to take care of any bad contacts thus, generating atomic-level structural models for different peptides with HLA-A*02:01. This approach of structure generation from peptide sequences is similar to the one used in a previously published study [7].

Structure feature generation: contact pair counts

In this study, the structure features are characterized by the contact pair counts between the peptide (ligand) and the HLA receptor (protein). The structure-based features are adapted from a recent study characterizing the binding of small molecule ligands with protein receptors [14]. In this case, each protein atom or ligand atom can be categorized into one of the 8 element types:

$$\boldsymbol{E_L} = [\boldsymbol{C_{aliphatic}}, \boldsymbol{C_{aromatic}}, \boldsymbol{O_{carbonyl}}, \boldsymbol{O_{hydroxyl}}, \boldsymbol{N}, \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{Du}] \quad \mathbf{1}$$

which would generate a total of 64 (8_{ligand} X $8_{protein}$) element-pair combinations between the ligand and the protein (Du represents all other remaining elements). Here, we defined 60 shells of width ~0.5 Å around each ligand atom and counted the number of different protein atoms within each of these 60 shells. The center of the first shell was set at the distance of 1 Å from the ligand atom and the center of the last shell has a distance of 30.5 Å to the ligand atom. Such a broad distance range covers both local and non-local interaction information between the ligand and the protein.

The atom-based contact number information between the ligand and the protein was then converted into element-pair specific and distance shell specific contact numbers. This was done by summing up all

the contact numbers between the ligand atoms and the protein atoms for a given elemental pair in a given shell resulting in 64 element-pair specific contact numbers for each of the 60 shells.

Thus, one peptide-HLA complex can be represented by an image matrix of the size (60 X 64) where 60 is the number of shells and 64 is the number of element-pair combinations (8 X 8) in each shell. The more detailed information on the structure-based features used here can be found in our previous study [14].

Sequence feature generation: peptide encoding:

Before inputting to the neural network, the raw peptide sequences were encoded into the numerical format. Although, there are several ways in which a peptide sequence can be represented, here we tested the three most popular methods of representing peptide sequence in a deep learning model. The first one is a one-hot encoding, in which each peptide residue is represented as a one-hot encoded vector of dimension 20 (standard amino acids) with the value corresponding to the residue was set to 1 and the remaining 19 values set to 0 (Figure 1a). Secondly, in the BLOSUM62 encoding, each residue in a peptide is encoded as a 20-dimensional vector from the BLOSUM62 matrix[24] thus, generating a 9 X 20 matrix representation for the peptide (Figure 1b). For embedding, the amino acid sequence is converted into an embedding vector by assigning discrete numerical codes to the residues in the peptide (Figure 1c). Each of these discrete numerical variables in the embedding vector is then mapped to a vector of continuous numbers (of dimension 20) thus, resulting in the embedding matrix. Unlike one-hot encoding or BLOSUM62 encoding, the mapping in the embedding matrix can be improved by learning using a neural network on a supervised task. For CNN based neural networks, all these encodings were mapped as 20-channel 1-row 2D tensors similar to DeepMHC[10] whereas, for LSTM based neural networks, the sequence information was supplied as 1-channel 2D tensor with 9 rows and 20 columns.

a)

		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
R	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
F	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
T	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
D	7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

b)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
F	1	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
V	2	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
R	3	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
F	4	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
S	5	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	6	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
D	7	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
K	8	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2

c)

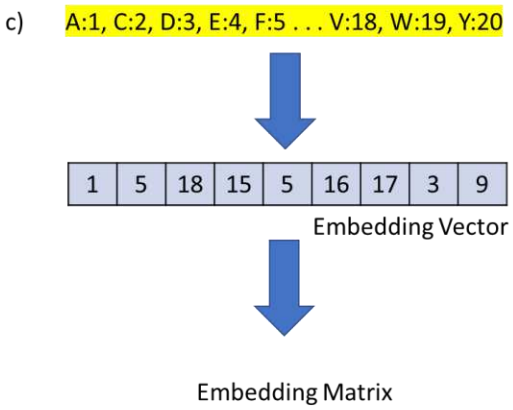


Figure 1: Illustration of encoding of peptide AFVRFSTDK with three different encoding techniques used in this study a) In one-hot encoding, each position is represented as a 20-dimensional vector with the value of corresponding residue kept as 1 and remaining 19 residues as 0 b) In BLOSUM62 encoding, each position is represented by the vector corresponding to the residue from the BLOSUM62 matrix c) In embedding vector, each residue is assigned an integer value and the peptide is represented as embedding vector formed out of those integer values.

Model Architecture:

The models used in this study consist of two modules: the structure module and the sequence module. The structure module consists of three convolution layers with each convolution layer is followed by a max-pooling layer. Batch-normalization and dropout layers are added after each convolution layer with "relu" as the activation function. The organization of the structural module was mainly inspired by our previous work, OnionNet[14]. For the sequence module, we tested both the convolution-based neural network and the LSTM-based neural network. The LSTM-based neural network in sequence module comprises of two LSTM-layers with each LSTM layer is followed by batch normalization and dropout layer (Figure 2a). The activation function used with LSTM layers is "tanh". The convolution-based neural network in sequence module comprises of three convolution layers with each convolution layer is

followed by a max-pooling layer (Figure 2b). The outputs of the structure and sequence blocks are received by the individual fully connected (FC) layer of 1024 neurons. The outputs are combined and fed into another FC layer of 512 neurons to predict the binding affinity. Deep learning models generally have millions of trainable parameters which is often larger than the number of training samples. This is because the input features often follow a very complicated relationship with target values and a network with few trainable parameters will not be able to represent the desired mapping. Having many parameters would make the network more flexible so that the features can be mapped properly to the target values. To reduce generalization error and prevent overfitting, we not only used implicit regularizations (e.g. dropout) but also employ explicit regularization L2 in our model with λ=0.001 [25-27].

We tested the three different peptide sequence encoding methods along with these two model architectures in the sequence module. The three sequence encoding methods have also been used to predict the binding affinity of peptides with MHC: one-hot encoding[10], BLOSUM62 encoding[28], and embedding vector representation embedding[6, 29]. Thus, we tested six models that differ in terms of peptide sequence encoding and the architecture of the sequence module. Since, the input features and the architecture of the structure module are the same in all the models, therefore, the tested models are named based on their peptide sequence encoding and network architecture used in the sequence module (Table 1).

S/No.	Model Name	Peptide encoding	Sequence Module Type
1	bls_cnn	BLOSUM62 encoding	CNN based
2	onehot_cnn	One-hot encoding	CNN based
3	emb_cnn	Embedding vector encoding	CNN based
4	bls_ lstm (OnionMHC)	BLOSUM62 encoding	LSTM based
5	onehot_lstm	One-hot encoding	LSTM based
6	emb_lstm	Embedding vector encoding	LSTM based

Table 1: Models with different peptide sequence encoding and the neural network architecture used in the sequence module. Three ways of peptide sequence encoding (one-hot, blosum62, embedding vector) combined with convolutional neural network

(CNN) or Long Short-Term Memory architectures in the sequence module resulted in a total of 6 different models that were tested in this study.

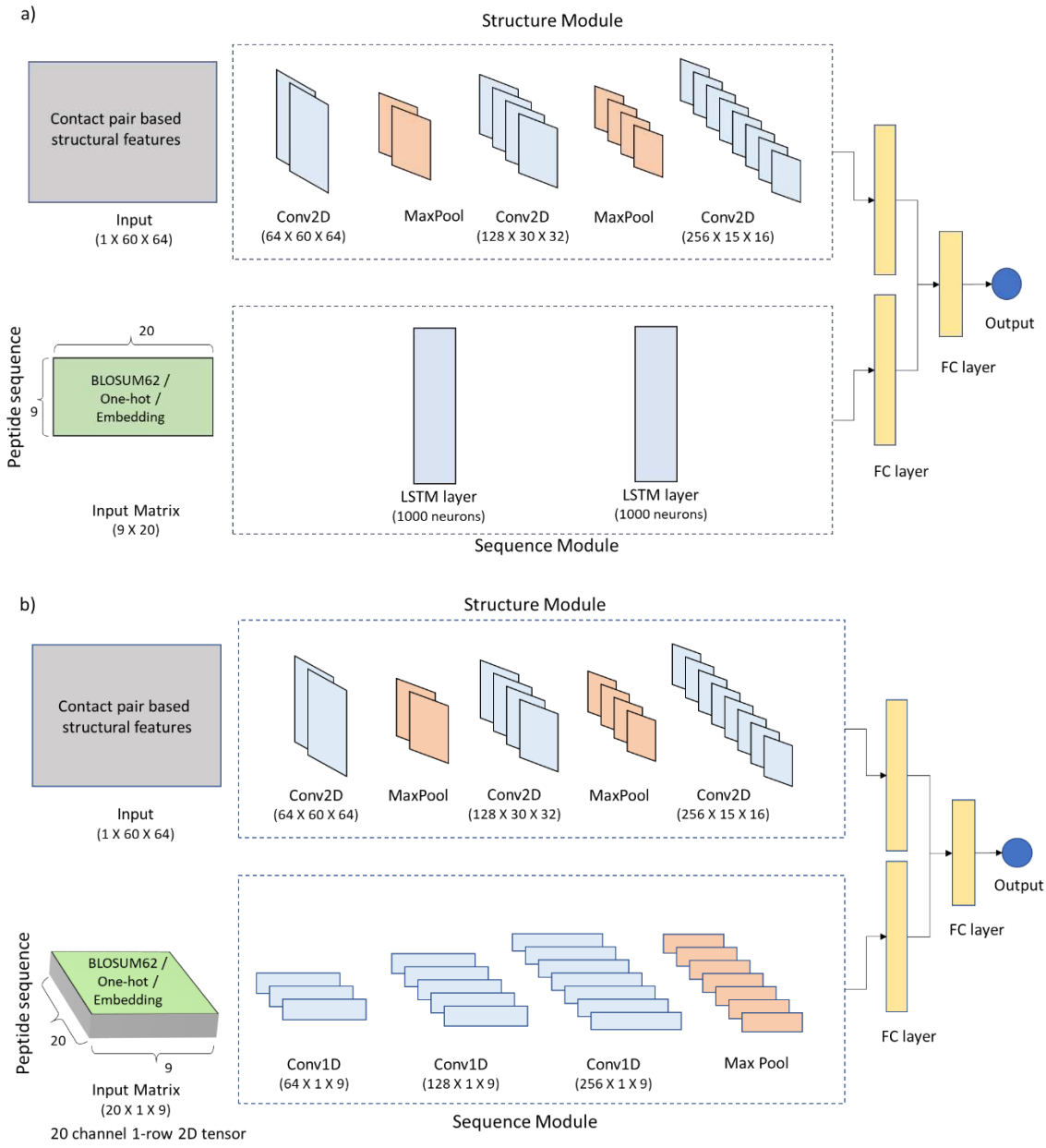


Figure 2: Architectures tested in this study with structure module based on Convolutional Neural Network: a) Long Short-Term Memory based neural network was used in sequence module with two lstm layers of 1000 neurons each and its output is combined with that of structure module using Fully Connected (FC) layer b) Convolutional Neural Network-based neural network was used in sequence module with three 1-D convolutional layers (of filter size 64, 128, 256), followed by maximum pooling and its output is also combined with that of structure module using Fully Connected (FC) layer

Model Training:

The binding affinities (in nM) in the collected dataset were transformed to numbers within 0 to 1 with a cap of 50,000 nM:

$$affintiy_{transformed} = min\left(1 - \frac{\log(affinity)}{\log(50000)}, 1\right) \quad (2)$$

as previously described in [5, 11, 28].

The hyperparameters involving the number of neurons, number of layers, and batch size were optimized using the five-fold cross-validation (Figure S1, Additional File 1). The model with optimized hyperparameters was trained with mean-squared-error used as the loss function.

$$Loss = \frac{1}{n}\sum_{i}^{n}(\tilde{y}_i - y_i)^2 \quad (3)$$

The dataset was randomly divided into 5-folds and the model was trained for 600 epochs with earlystopper of 30 epochs if the validation loss did not decrease more than 0.0001. The framework used for the model implementation was keras-tensorflow2.0 on NVIDIA-GeForce GTX 1080 Ti.

For the cross-validation, an ensemble of 5 deep neural networks was generated for each model by training the network on 4 folds and testing it on the remaining fold. For the final prediction on the benchmark dataset, two more trials of each of the 5 cross-validated neural networks were performed thus, resulting in 15 deep neural networks for each model. The average prediction score of all these 15 networks was taken as the final prediction score for a peptide. This methodology was inspired by the recently published model ACME[11].

Performance Metric:

The performance of the model was independently evaluated on the benchmark data obtained from IEDB and compared with the existing models such as NetMHCpan4.0, convMHC, HLAthena. The performance metrics used for comparison is the Area Under Receiver Operating Curve (AUC) and the Spearman Rank Correlation Coefficient (SRCC). This SRCC metric is the same as adopted by the weekly benchmark dataset provided by IEDB (http://tools.iedb.org/auto_bench/mhci/help). In the weekly benchmark dataset, the performance was reported in terms of SRCC even for binary data also. In that case, the SRCC calculation was done by keeping the target values as either 0 (non-binder) or 1 (binder) which is different from ic50 or half-life measurements.

Molecular Dynamics Simulation

The stability of the modeled complex structures was explored by performing molecular dynamics (MD) simulations using GROMACS[23]. The bonding and non-bonding parameters were derived from AMBER99SB-ILDN force field[30]. The simulation was performed in a cubic box with water molecules and ions (Na^+ and Cl^-) concentration of 0.15M at 300K. The system was equilibrated before carrying out the production simulation. Under NVT conditions, the equilibration was performed for 100 ps, then under NPT conditions for 100 ps using 2 fs time step, Berendsen thermostat (τ = 0.1 ps), and barostat (τ = 2 ps). The bonds and angles were constrained by the LINCS algorithm[31]. The production simulation was carried out under NPT conditions for 50 ns with repeats. A leap-frog algorithm[32] was used to integrate Newton's equation of motion. The cut-offs for both van der Waals (VDW) interaction and short-range electrostatic interaction were set to 1.0 nm, and the particle mesh Ewald (PME) method[33] was employed to deal with the long-range electrostatic interactions.

Cell culture

T2 cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 media (Hyclone) supplemented with 20% heat-inactivated Fetal Bovine Serum (FBS; Hyclone) and 1% penicillin/streptomycin (Hyclone) at 37°C in a CO_2 incubator.

Peptide binding assay

Peptides, as shown in Table 5 were synthesized from Peptide 2.0, USA, in a lyophilized form with >95% purity. The peptides were dissolved in either sterile water or 10% DMSO depending on the sequences. The binding of the candidate peptides to the HLA-A*02:01 molecules was assessed by measuring the induction of surface expression of HLA-A*02:01 molecules on peptide-pulsed T2 cells following an established protocol[34] with a few modifications. The T2 cells were seeded in a 24 well plate and then treated with the peptides at 100 µM concentration. The peptide-pulsed T2 cells were then incubated at 37°C in a CO_2 incubator for 12 to 16h. The cells were then stained with HLA-A02:01 specific mAb conjugated to FITC for 30 min at room temperature and then analyzed by flow cytometer. The fluorescence index (FI) was calculated using the following formula:

$$\mathrm{FI} = \frac{\left[\text{mean fluorescence intensity } (\mathrm{MFI})_{\text{sample}} - \mathrm{MFI}_{\text{background}}\right]}{\mathrm{MFI}_{\text{background}}}$$

where $MFI_{background}$ represents the value without peptide. FI > 1.5 indicated that the peptide had a high affinity for HLA-A*02:01 molecules, 1.0 < FI < 1.5 indicated that the peptide had a moderate affinity for the HLA-A*02:01 molecules, and 0.5 < FI < 1.0 indicated that the peptide had low affinity for the HLA-A*02:01 molecules.

Results and Discussions

Five-fold Cross-Validation:

We evaluated the performance of all the six models using the five-fold cross-validation. The entire data was divided into five-folds and 5 different models were generated with each model tested on one of the 5 folds and trained on the remaining 4 folds. The performance evaluation of the above-mentioned models (Table 1) was done in terms of both AUC and SRCC. Overall, the LSTM-based neural network in the sequence module achieved better performance with our structure module compared to the CNN-based neural network in the sequence module as shown in Table 2 (Table S1, Additional File 2). The peptide encoding does not have much difference in the model performance with the LSTM-based neural network in the sequence module. However, for the sequence module with a CNN-based neural network, the one-hot encoding performed a bit worse than the blosum62 encoding or the embedding vector representation.

	emb_lstm	onehot_lstm	bls_lstm	emb_cnn	onehot_cnn	bls_cnn
AUC	0.955	0.953	0.954	0.943	0.878	0.941
SRCC	0.816	0.816	0.821	0.798	0.677	0.799

Table 2: Summary of five-fold cross-validation measured in terms of area under the receiver operating curve (AUC) and Spearman Rank Correlation Coefficient (SRCC)

Independent Evaluation:

For independent evaluation, we downloaded the weekly benchmark dataset (since 21/03/2014) from IEDB (http://tools.iedb.org/auto_bench/mhci/weekly/) for all the 9-mer peptides belonging to HLA-A*02:01 receptors. The performance of all models was evaluated on the weekly benchmark dataset in terms of AUC and SRCC (Figure 3, 4). Among all the models generated, bls_lstm (OnionMHC) achieved the best performance both in terms of AUC and SRCC on the IEDB benchmark dataset (Figure 3). The

performance of OnionMHC was compared with other different algorithms (Figure 4) which were already evaluated in previous work [35]. We also compared the performance of our model with the current state-of-the-art baseline model NetMHCpan4.0 (http://www.cbs.dtu.dk/services/NetMHCpan/) and the convolution-based neural network model convMHC**.** The standalone version of NetMHCpan4.0 was downloaded from http://www.cbs.dtu.dk/services/NetMHCpan/ and installed as instructed. The predictions of the extracted benchmark dataset were made and AUC and SRCC were calculated. NetMHCpan4 gives two types of outputs which are the binding probability of peptide with HLA receptor (NetMHCpan4-L) and the binding affinity of the peptide with the receptor (NetMHCpan4-B). For convMHC, the predictions of the extracted benchmark dataset were made from the webserver (http://jumong.kaist.ac.kr:8080/convmhc). We also compared our model with the recently published HLAthena[15] which can be accessed from the webserver (http://hlathena.tools/). On average, our model achieved a better performance both in terms of AUC and SRCC compared to NetMHCpan4-L, convMHC, and HLAthena and the performance was comparable to NetMHCpan4-B (Table 3).

The IEDB benchmark dataset consists of data of three measurement types: binary, ic50, and t1/2. In one of the references, we found some discrepancies with the peptides identified as positives. The reference id 1026840 showed the same peptides with two different measurements: ic50 and t1/2. With the conventional cut-offs of ic50 (< 500 nM) and t1/2 (> 120 min) for a peptide to be counted as positive, the numbers of binders were different (positives) in both measurements. There were 14 positives out of 24 peptides according to ic50 cut-off and 7 positives out of 24 according to t1/2 measurement. Because of this discrepancy, we did not consider this reference id for evaluating our model.

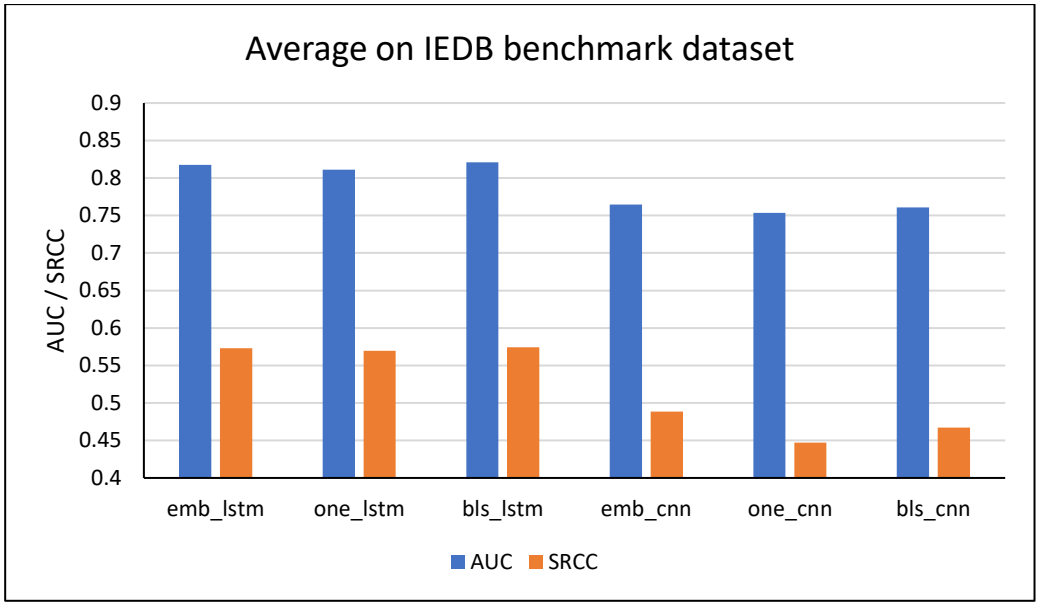


Figure 3: Performance evaluation of the six generated models on the IEDB benchmark dataset

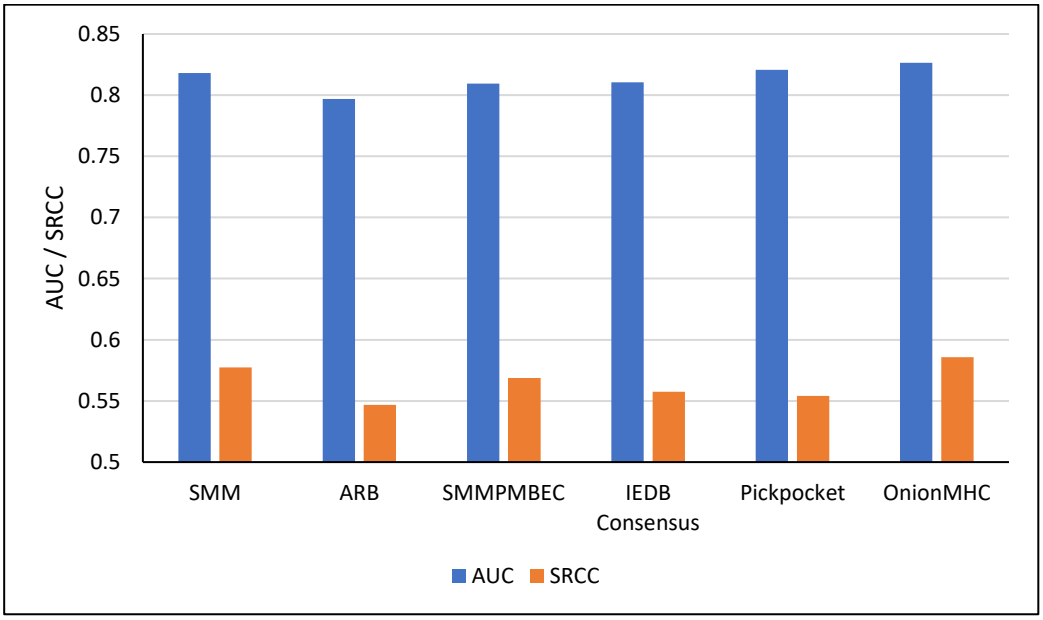


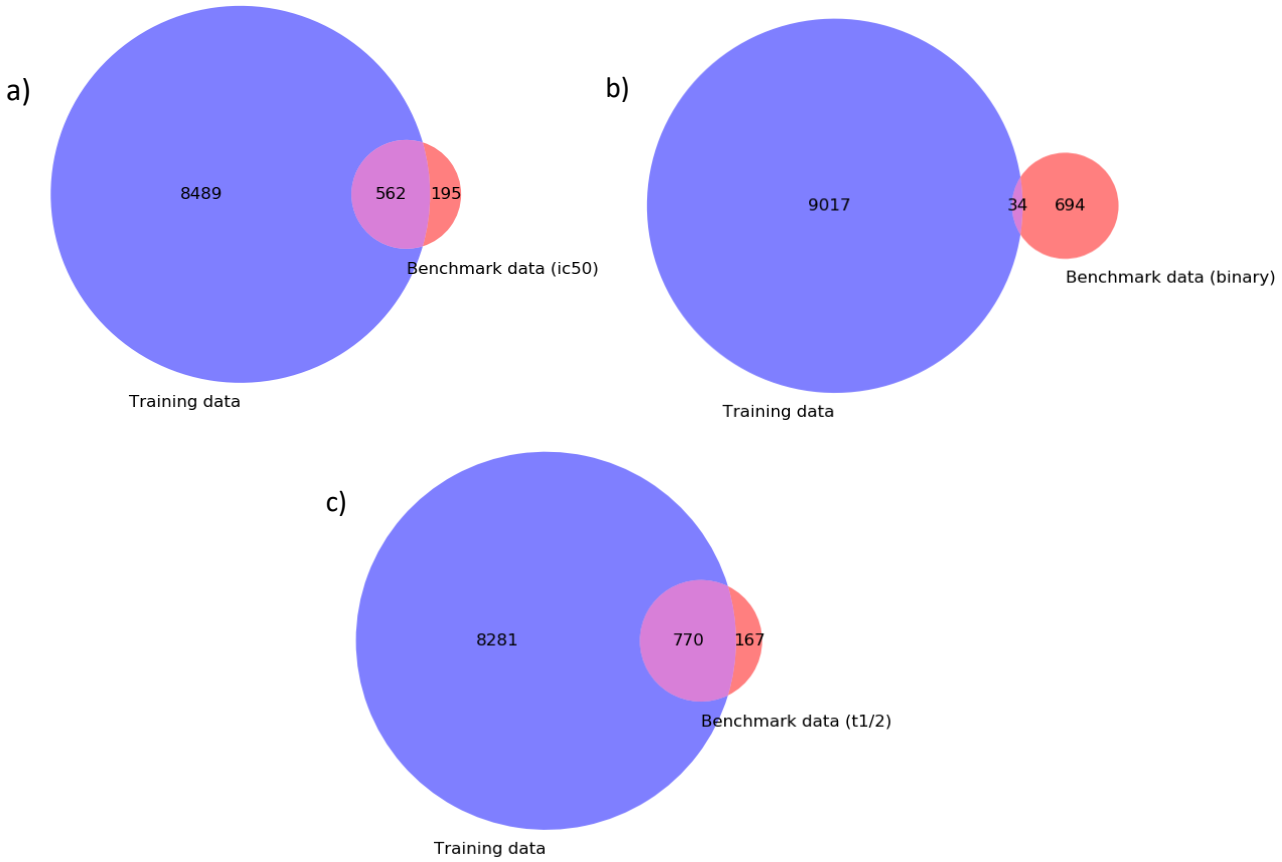
Figure 4: Performance comparison of OnionMHC with different prediction algorithms (Trolle et. al 2015) on the most recent IEDB benchmark dataset, measured in terms of Spearman Rank Correlation Coefficient (SRCC) and area under receiver operating curve (AUC)

IEDB Ref.	Peptide Count	Positive Count	Measurement type	NetMHCpan4 - L		NetMHCpan4 - B		convMHC		HLAthena		OnionMHC	
				AUC	SRCC	AUC	SRCC	AUC	SRCC	AUC	SRCC	AUC	SRCC
1033576	191	13	binary	0.619	0.104	0.63	0.113	0.63	0.09	0.596	0.084	0.666	0.145
1029824	77	56	binary	0.479	-0.033	0.556	0.087	0.588	0.14	0.526	0.039	0.565	0.101
1028928	13	11	binary	0.955	0.57	0.955	0.57	0.932	0.542	0.955	0.57	0.955	0.57
1027588	19	8	binary	0.829	0.564	0.841	0.584	0.705	0.35	0.875	0.642	0.773	0.467
1027471	45	6	binary	0.859	0.423	0.838	0.398	0.823	0.381	0.889	0.458	0.774	0.322
1027079	18	13	binary	0.846	0.538	0.846	0.538	0.815	0.49	0.815	0.49	0.769	0.418
1026840	357	76	binary	0.877	0.534	0.904	0.573	0.872	0.529	0.819	0.452	0.886	0.547
1026941	10	6	binary	1	0.853	0.917	0.711	1	0.853	0.875	0.64	1	0.853
1031894	434	368	ic50	0.776	0.531	0.901	0.743	0.82	0.603	0.743	0.479	0.932	0.8
1031072	28	23	ic50	0.974	0.739	1	0.875	0.957	0.772	0.913	0.727	0.974	0.856
1033071	113	41	ic50	0.866	0.77	0.876	0.812	0.851	0.759	0.871	0.787	0.872	0.819
1028790	55	47	ic50	0.598	0.628	0.582	0.626	0.612	0.607	0.516	0.508	0.588	0.643
1028553	22	2	ic50	0.7	0.432	0.85	0.574	0.85	0.603	0.575	0.229	0.925	0.568
1028554	44	7	ic50	0.838	0.539	0.876	0.721	0.753	0.417	0.9	0.641	0.884	0.679
1024516	51	45	ic50	0.663	0.685	0.63	0.667	0.693	0.658	0.556	0.558	0.678	0.728
1026941	10	6	ic50	1	0.864	0.917	0.717	1	0.422	0.875	0.724	1	0.824
1026371	85	49	t12	0.781	0.564	0.811	0.566	0.784	0.517	0.793	0.55	0.803	0.543
1028285	881	549	t12	0.822	0.656	0.865	0.721	0.829	0.658	0.799	0.626	0.831	0.66
Average				**0.805**	**0.553**	**0.822**	**0.588**	**0.806**	**0.521**	**0.772**	**0.511**	**0.826**	**0.586**

*Table 3: AUC Performance evaluation of OnionMHC with NetMHCpan and ConvMHC on IEDB benchmark dataset for HLA-A*02:01 with peptide length 9*

All these models (OnionMHC, NetMHCpan4, HLAthena, and convMHC) utilized the data from IEDB (http://tools.iedb.org/main/datasets/) in their training set and their evaluation was done on the benchmark data also provided by IEDB under three measurement values (binary, ic50, and t1/2). However, there is an overlap between these two datasets, so we removed the overlapping peptides for

the three measurement values (Figure 5) and re-calculated the spearman correlation (SRCC) and area under the receiver operating curve (AUC) on the benchmark peptides. These benchmark peptides have not been seen by OnionMHC during the training or the validation. Among the datasets with three measurement values, OnionMHC performed notably better than convMHC, NetMHCpan4-L, and HLAthena on ic50 measurement values, and equally well with NetMHCpan4-B (Table 4, Figure 6). For the data with the binary form, OnionMHC still performed better than HLAthena and similarly well with convMHC, NetMHCpan4-L, and NetMHCpan4-B. In the half-time measurement dataset, we found that the peptide TVYDINNEV had two different measurement values of 120 min and 170.4 min. With the conventional cut-off of >120 min for classifying a peptide as a binder, this peptide would have created a discrepancy, so we removed it from the analysis. For the half-time measurement dataset, OnionMHC did not perform as well as other models however, the performance was still comparable.



315

316 *Figure 5: Overlapping peptides between the training data and benchmark data with measurement value: a) binary; b) ic50; c)*
317 *half-life (t1/2)*

	OnionMHC	**NetMHCpan4.0-L**	**NetMHCpan4.0-B**	**convMHC**	**Hlathena predictor**
For ic50 only (195 peptides)					
SRCC	0.790	0.716	0.787	0.697	0.741
AUC	0.885	0.857	0.882	0.834	0.869
For binary only (694 peptides)					
SRCC	0.529	0.524	0.539	0.519	0.477
AUC	0.860	0.856	0.866	0.852	0.824

For half-life only (167 peptides)					
SRCC	0.465	0.618	0.603	0.532	0.611
AUC	0.754	0.789	0.819	0.790	0.761

Table 4: Evaluation of models on the benchmark data after removing the overlapping peptides

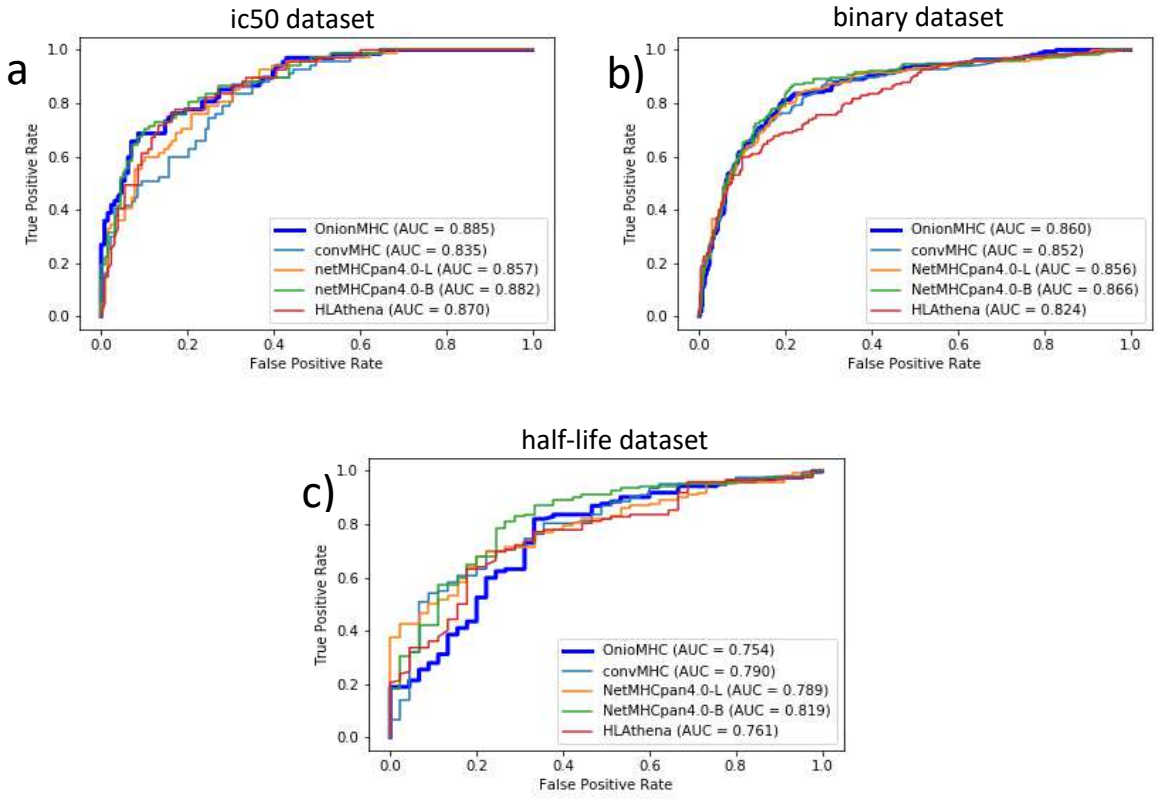


Figure 6: Area under the receiver operating curves for the non-overlapping peptides a) ic50 dataset (195 peptides) b) binary dataset (694 peptides) c) half-life dataset (167 peptides)

Model Ablation

OnionMHC has both structure as well as sequence modules. To understand the importance of each module in predicting the binding affinity of a peptide, we performed model ablation analysis on our model. In this analysis, either of the modules is removed and the rest of the model is trained similarly to the original model and evaluated for its performance on the IEDB benchmark dataset. As shown in Figure 7, the sequence module is the main contributor towards the binding affinity prediction, however, the

addition of the structure module further improves the predictive power of the sequence module. The structure information was obtained mostly from modeling instead of experimentally determined PDB structures. The inaccuracy or the noise in the structural information may damage the quality of the structural models. This could be the reason why the sequence module performs better than the structure module in the prediction of binding affinity.

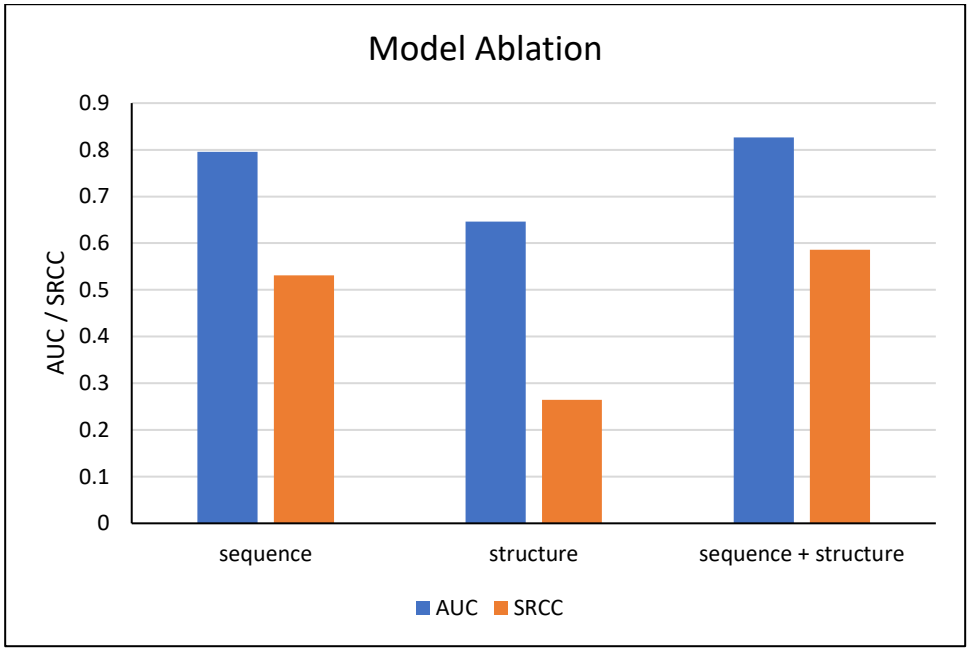


Figure 7: Model Ablation analysis of OnionMHC showing the individual performance of sequence and structure module on IEDB benchmark dataset

Demonstration by Molecular Dynamics Simulations

To demonstrate the stability of the predicted binding peptides as well as the non-binding peptides in the complex with the receptor, we performed molecular dynamics simulations on the receptor-peptide system. These peptides in the IEDB benchmark data have already been validated experimentally[35]. In total we randomly selected 6 peptides, 3 of which are predicted to be binders with the binding affinity less than 500 nM (FLDEFMEGV, LLIDLTSFL, NLIDFNFKL) and the remaining 3 were non-binders with the binding affinity greater than 500 nM (ANRAVKPTG, DIKVQFQSG, TPTAYRPPN). We calculated the residue-wise root mean square fluctuation (RMSF) of the peptide backbone (Figure 8). The peptide backbone,

especially those at the termini, shows larger RMSF for the non-binding peptides compared to that of the binding peptides. The peptide binds to class I MHC primarily through the invariant peptidic termini[36, 37]. The second residue and the last residue (residue 9) of the peptide play an important role in its binding affinity to the HLA-A*02:01 receptors [38-40]. Higher RMSF of peptide backbone at the amino terminus indicates that the peptide is highly flexible at the termini thus, loosely anchored to the receptor. The low RMSF for the binding peptides at the amino terminus indicates the strong binding of the peptide to the HLA receptor thus, which is consistent with the predictions made by OnionMHC. A similar idea of using the fluctuations from MD trajectories to distinguish the true binding from decoys has also proved to be quite successful[41].

We also performed the binding energy calculations using g_mmpbsa [42] available at https://rashmikumari.github.io/g_mmpbsa/. The molecular mechanics van der Waals potential energy is lower for binders than non-binders which is consistent with the predictions made by our model. However, the total binding energy that includes both molecular mechanics potential energy and free energy of solvation is unable to capture the same trend (Table S2, Additional File 3). Thus, the MMPBSA method is unable to distinguish the binders from non-binders.

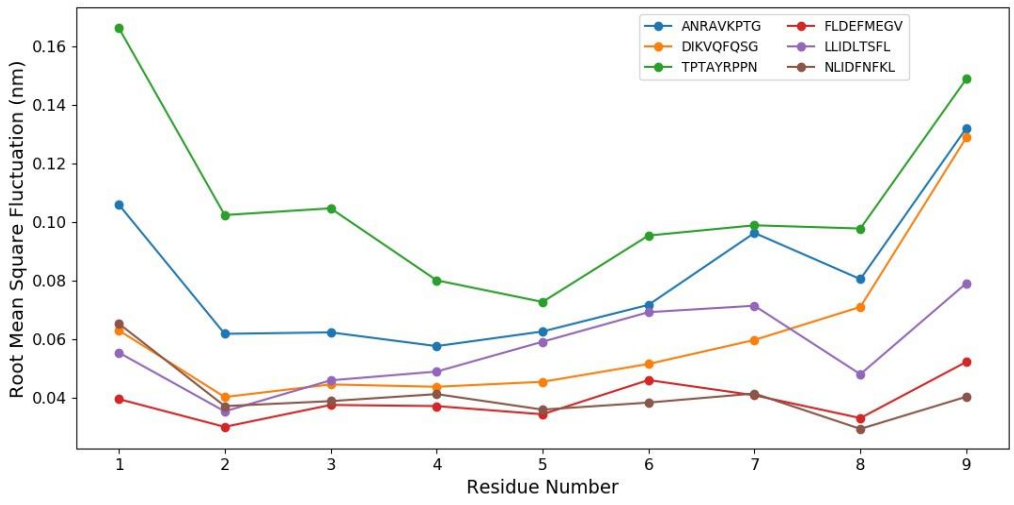


Figure 8: Root Mean Square Fluctuation (nm) of the residues in the peptide in the Molecular Dynamics simulation.

Experimental Validation

We also tested our model on the peptides that were obtained from neoantigen analysis on the patient exome sequencing samples from breast cancer patients. Our model OnionMHC performed equally well as NetMHCpan4-L and convMHC with correctly predicting 8 out of 11 peptides (Table 5). Two peptides LLAYRAYEL and NLWPVPHLL were predicted as false positives by all the models whereas RLPIFRLSI was predicted as false positive by NetMHCpan4-L and convMHC only. Of the 11 peptides tested, NetMHCpan4-L and convMHC predicted 3 false positives (LLAYRAYEL, NLWPVPHLL, and RLPIFRLSI), NetMHCpan4-B predicted 2 false positives (LLAYRAYEL, NLWPVPHLL) whereas OnionMHC predicted 2 false positives (LLAYRAYEL and NLWPVPHLL) and 1 false negative (GLIEEGEKL). With 2 false positives, OnionMHC and NetMHCpan4-B showed a better positive predictive value (PPV) of 0.67 as compared to the value of 0.57 by NetMHCpan4-L and convMHC.

S.No.	Peptide Sequence	convMHC		NetMHCpan4.0-L		NetMHCpan4.0-B		OnionMHC		Experimental results	
		Score	Binding	Score	Binding	Score	Binding	Score	Binding	MFI Score	Binding
1	FLPPPECPV	0.8989	YES	0.676588	YES	0.856453	YES	13.31661	YES	3.13	YES
2	FLLPPECPV	0.9211	YES	0.799927	YES	0.759096	YES	6.828626	YES	2.51	YES
3	RLPIFSRLS	0.096	NO	0.003027	NO	0.406621	NO	9656.127	NO	-0.052	NO
4	RLPIFRLSI	0.8179	YES	0.170836	YES	0.13418	NO	610.4254	NO	-0.043	NO
5	LLAYRAYEL	0.8889	YES	0.495571	YES	0.68252	YES	18.51016	YES	-0.066	NO
6	LLACRAYEL	0.8559	YES	0.31138	YES	0.750783	YES	40.86188	YES	0.63	YES
7	GLMEEGEKL	0.8416	YES	0.936163	YES	0.68583	YES	175.2056	YES	2.33	YES
8	GLIEEGEKL	0.8378	YES	0.886652	YES	0.593803	YES	682.3957	NO	0.66	YES
9	NLWPVPHLL	0.9146	YES	0.970742	YES	0.809898	YES	13.40361	YES	0.03	NO
10	MLWRKTLKN	0.0944	NO	0.001155	NO	0.129977	NO	11879.35	NO	-0.026	NO
11	MLWRKTSKN	0.106	NO	0.000408	NO	0.087518	NO	13751.64	NO	0.03	NO

Table 5: Experimental validation of peptides derived from exome sequencing analysis of breast cancer patients.

Conclusion

In this study, we have developed a deep learning-based model to predict the binding affinity of a peptide with the HLA-A*02:01 receptors. The model employs both structure as well as sequence feature sets to predict binding affinity. This approach is quite different from the previous structure only or sequence only based approaches. Each of the methods has its advantage, however, combining features from both domains can provide a better prediction of the binding affinity of the peptide with MHC. Our model not only uses the information about the residue and their position in the peptide but also about the interactions taking place at the atomic scale. Since, in the structure-based features, the whole peptide is treated as a ligand, the residue-wise contribution towards peptide binding is lost. Thus, explicitly adding the sequence-based features with the structure-based features would allow the network to learn from the atomic interactions between the peptide-MHC as well as the different peptide residues contributing towards binding. We tested 2 different architectures in the sequence module with 3 different sequence feature sets. Although, there was not a very significant difference yet blosum62 encoding of the peptide sequences showed the best performance in LSTM architecture of the sequence module along with the structure module on the weekly benchmark dataset from IEDB. The combination of both structure and sequence features increases the performance compared to those models with either of the features alone as shown in model ablation analysis. Our model performed better than convMHC and HLAthena and is comparable with the current state-of-the-art model, NetMHCpan4. Furthermore, the predictions made by our model were consistent with the results from the molecular dynamics simulations and our model performs well on the peptides from breast cancer patients validated experimentally. Due to the involvement of intensive computation in structure modeling, the model was currently trained and tested on HLA-A*02:01 only. In the near future, the model can be further extended to predict the binding affinities of peptides to different alleles, a pan-alleles model.

References:

1. Castle JC, Kreiter S, Diekmann J, Lower M, van de Roemer N, de Graaf J, Selmi A, Diken M, Boegel S, Paret C *et al*: **Exploiting the mutanome for tumor vaccination**. *Cancer Res* 2012, **72**(5):1081-1091.
2. Banchereau J, Palucka K: **Immunotherapy: Cancer vaccines on the move**. *Nat Rev Clin Oncol* 2018, **15**(1):9-10.
3. Yewdell JW, Bennink JR: **Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses**. *Annu Rev Immunol* 1999, **17**:51-88.
4. Massimo Andreatta MN: **Gapped sequence alignment using artificial neural networks: application to the MHC class I system**. *Bioinformatics* 2016:511-517.
5. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M: **NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data**. *J Immunol* 2017, **199**(9):3360-3368.
6. Vang YS, Xie X: **HLA class I binding prediction via convolutional neural networks**. *Bioinformatics* 2017, **33**(17):2658-2665.
7. Sumanta Mukherjee CB, Nagasuma Chandra **HLaffy: estimating peptide affinities for Class-1 HLA molecules by learning position-specific pair potentials**. *Bioinformatics* 2016, **32**(15):2297-2305.
8. Bordner AJ, Abagyan R: **Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes**. *Proteins* 2006, **63**(3):512-526.
9. Han Y, Kim D: **Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction**. *BMC Bioinformatics* 2017, **18**(1):585.
10. Jianjun Hu ZL: **DeepMHC: Deep Convolutional Neural Networks forHigh-performance peptide-MHC Binding Affinity Prediction**. 2017.
11. Yan Hu ZW, Hailin Hu, Fangping Wan, Lin Chen, Yuanpeng Xiong, Xiaoxia Wang, Dan Zhao, Weiren Huang, Jianyang Zeng: **ACME: Pan-Specific peptide-MHC Class I Binding Prediction Through Attention-Based Deep Neural Networks** *Bioinformatics* 2019:4946-4954.
12. Thammakorn Saethang OH, Ingorn Kimkong, Vu Anh Tran, Xuan Tho Dang, Lan Anh T Nguyen, Tu Kien T Le, Mamoru Kubo, Yoichi Yamada, Kenji Satou: **PAAQD: Predicting immunogenicity of MHC class I binding peptides using amino acid pairwise contact potentials and quantum topological molecular similarity descriptors.** *Journal of Immunological Methods* 2013, **387**(1-2):293-302.
13. Michelle P. Aranha CS, Omar Demerdash, Bogdan Czejdo, Jeremy C. Smith, Julie C. Mitchell: **Prediction of peptide binding to MHC using machine learning with sequence and structure-based feature sets**. *Biochimica et Biophysica Acta (BBA) - General Subjects* 2020, **1864**(4).
14. Zheng L, Fan J, Mu Y: **OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein-Ligand Binding Affinity Prediction**. *ACS Omega* 2019, **4**(14):15956-15965.
15. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, Hartigan CR, Zhang W, Braun DA, Ligon KL *et al*: **A large peptidome dataset improves HLA class I epitope prediction across most of the human population**. *Nat Biotechnol* 2020, **38**(2):199-209.
16. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B: **The Immune Epitope Database (IEDB): 2018 update**. *Nucleic Acids Res* 2019, **47**(D1):D339-D343.

17. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, Peters B: **Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions**. *BMC Bioinformatics* 2014, **15**:241.
18. Liu Z, Cui Y, Xiong Z, Nasiri A, Zhang A, Hu J: **DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction**. *Sci Rep* 2019, **9**(1):794.
19. Wu J, Wang W, Zhang J, Zhou B, Zhao W, Su Z, Gu X, Wu J, Zhou Z, Chen S: **DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity**. *Front Immunol* 2019, **10**:2559.
20. Helen M. Berman JW, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne: **The Protein Data Bank** *Nucleic Acids Research* 2000, **28**(1):235-242.
21. Krivov GG, Shapovalov MV, Dunbrack RL, Jr.: **Improved prediction of protein side-chain conformations with SCWRL4**. *Proteins* 2009, **77**(4):778-795.
22. Knapp B, Omasits U, Schreiner W: **Side chain substitution benchmark for peptide/MHC interaction**. *Protein Sci* 2008, **17**(6):977-982.
23. Mark James Abraham TM, Roland Schulz, Szilárd Pálla, Jeremy C.Smith, Berk Hessa Erik Lindahl: **GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers**. *SoftwareX* 2015, **1-2**:19-25.
24. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci U S A* 1992, **89**(22):10915-10919.
25. Vinyals CZaSBaMHaBRaO: **Understanding deep learning requires rethinking generalization**. *arXiv:161103530* 2017.
26. Hertz AKaJA: **A Simple Weight Decay Can Improve Generalization**. *Advances in Neural Information Processing Systems 4* 1992:950-957.
27. Moody JE: **Note on generalization, regularization and architecture selection in nonlinear learning systems**. *Neural Networks for Signal Processing Proceedings of the 1991 IEEE Workshop* 1991:1-10.
28. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J: **MHCflurry: Open-Source Class I MHC Binding Affinity Prediction**. *Cell Syst* 2018, **7**(1):129-132 e124.
29. Phloyphisut P, Pornputtapong N, Sriswasdi S, Chuangsuwanich E: **MHCSeqNet: a deep neural network model for universal MHC binding prediction**. *BMC Bioinformatics* 2019, **20**(1):270.
30. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE: **Improved side-chain torsion potentials for the Amber ff99SB protein force field**. *Proteins* 2010, **78**(8):1950-1958.
31. Berk Hess HB, Herman J. C. Berendsen, Johannes G. E. M. Fraaije: **Lincs: a linear constraint solver for molecular simulations**. *Journal of Computational Chemistry* 1998, **18**:1463-1472.
32. Berendsen WFVGaHJC: **A Leap-frog Algorithm for Stochastic Dynamics**. *Molecular Simulation* 1988, **1**:173-185.
33. Harvey MJ, De Fabritiis G: **An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware**. *J Chem Theory Comput* 2009, **5**(9):2371-2377.
34. Hansen T, Myers N: **Peptide induction of surface expression of class I MHC**. *Curr Protoc Immunol* 2003, **Chapter 18**:Unit 18 11.
35. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, Sette A, Peters B, Nielsen M: **Automated benchmarking of peptide-MHC class I binding predictions**. *Bioinformatics* 2015, **31**(13):2174-2181.

36. Khan AR, Baker BM, Ghosh P, Biddison WE, Wiley DC: **The structure and stability of an HLA-A*0201/octameric tax peptide complex with an empty conserved peptide-N-terminal binding site**. *J Immunol* 2000, **164**(12):6398-6405.
37. Bouvier M, Wiley DC: **Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules**. *Science* 1994, **265**(5170):398-402.
38. Sharma AK, Kuhns JJ, Yan S, Friedline RH, Long B, Tisch R, Collins EJ: **Class I major histocompatibility complex anchor substitutions alter the conformation of T cell receptor contacts**. *J Biol Chem* 2001, **276**(24):21443-21449.
39. Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A: **Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules**. *Cell* 1993, **74**(5):929-937.
40. Robert A. Colbert SLR-J, Andrew J. McMichael, Jeffrey A. Frelinger: **Differences in peptide presentation between B27 subtypes: The importance of the P1 side chain in maintaining high affinity peptide binding to B★2703**. *Immunity* 1994, **1**(2):121-130.
41. Guterres H, Im W: **Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations**. *J Chem Inf Model* 2020, **60**(4):2189-2198.
42. Rashmi Kumari RK, Open Source Drug Discovery Consortium and Andrew Lynn: **g_mmpbsa—A GROMACS Tool for High-Throughput MM-PBSA Calculations**. *Journal of Chemical Information and Modelling* 2014.

Declarations:

Acknowledgments:

We thank Liangzhen Zheng for providing his valuable inputs that helped in optimizing the deep learning models developed in this study.

Funding:

This research is supported by MOE Tier 1 Grant RG146/17 awarded to Yuguang Mu; National Research Foundation (NRF) Singapore through an NRF Fellowship awarded to M.J.F (NRF-NRFF2012-054); NTU start-up funds awarded to M.J.F.; RNA Biology Center at the Cancer Science Institute of Singapore, NUS, as part of funding under the Singapore Ministry of Education Academic Research Fund Tier 3 awarded to Daniel Tenen (MOE2014-T3-1-006); National Research Foundation Singapore and the Singapore Ministry

of Education under its Research Centres of Excellence initiative. The computing facility was supported by the National Supercomputing Centre, Singapore (https://www.nscc.sg).

Availability of data and its materials:

OnionMHC model and its source code, together with a usage guide, the dataset used for training as well as weekly benchmark data used for independent evaluation are freely available on GitHub at https://github.com/shikhar249/OnionMHC. The binding affinity data used for training is also provided in Additional File 4.

Authors' Affiliations:

Cancer Science Institute, National University of Singapore, 14 Medical Drive, Singapore, 117599
Sambhavi Animesh and Melissa Fullwood

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551
Shikhar Saxena, Melissa Fullwood, and Yuguang Mu

Authors' Contributions:

SS and YM designed the study. SS developed and analyzed the model. SA performed the experimental validation of peptides. YM and MF provided overall supervision. SS and YM wrote the manuscript with assistance from other authors. All authors approved the final manuscript.

Corresponding author:

Correspondence to Yuguang Mu

Ethics approval and consent to participate:

Not applicable

Consent for publication:

Not applicable

Competing interests:

The authors declare no competing interests.

Figures

a)

		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
R	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
F	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
T	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
D	7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

b)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
F	1	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
V	2	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
R	3	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
F	4	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
S	5	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	6	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
D	7	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
K	8	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2

c)

A:1, C:2, D:3, E:4, F:5 . . . V:18, W:19, Y:20

1	5	18	15	5	16	17	3	9

Embedding Vector

Embedding Matrix

Figure 1

Illustration of encoding of peptide AFVRFSTDK with three different encoding techniques used in this study a) In one-hot encoding, each position is represented as a 20-dimensional vector with the value of corresponding residue kept as 1 and remaining 19 residues as 0 b) In BLOSUM62 encoding, each position is represented by the vector corresponding to the residue from the BLOSUM62 matrix c) In embedding vector, each residue is assigned an integer value and the peptide is represented as embedding vector formed out of those integer values.

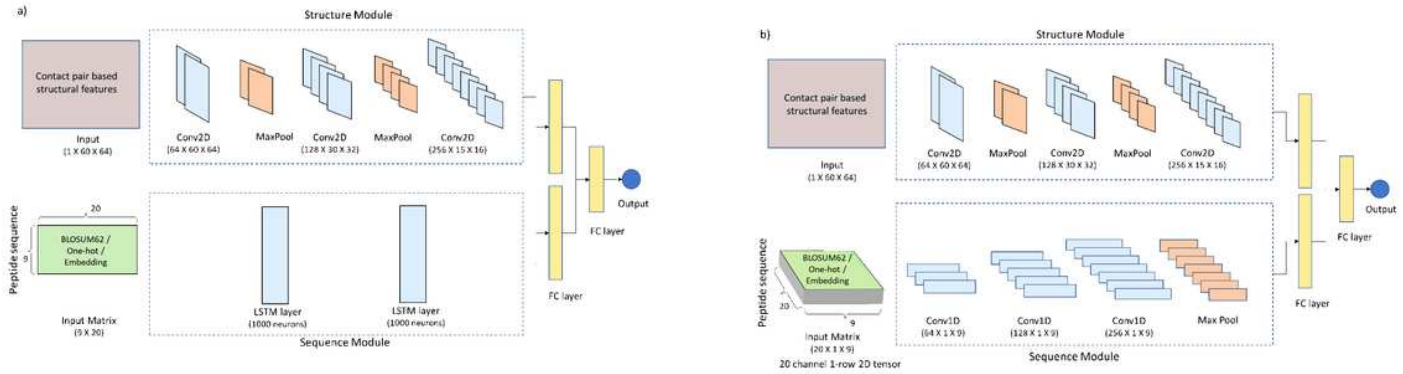


Figure 2

Architectures tested in this study with structure module based on Convolutional Neural Network: a) Long Short-Term Memory based neural network was used in sequence module with two lstm layers of 1000 neurons each and its output is combined with that of structure module using Fully Connected (FC) layer b) Convolutional Neural Network-based neural network was used in sequence module with three 1-D convolutional layers (of filter size 64, 128, 256), followed by maximum pooling and its output is also combined with that of structure module using Fully Connected (FC) layer

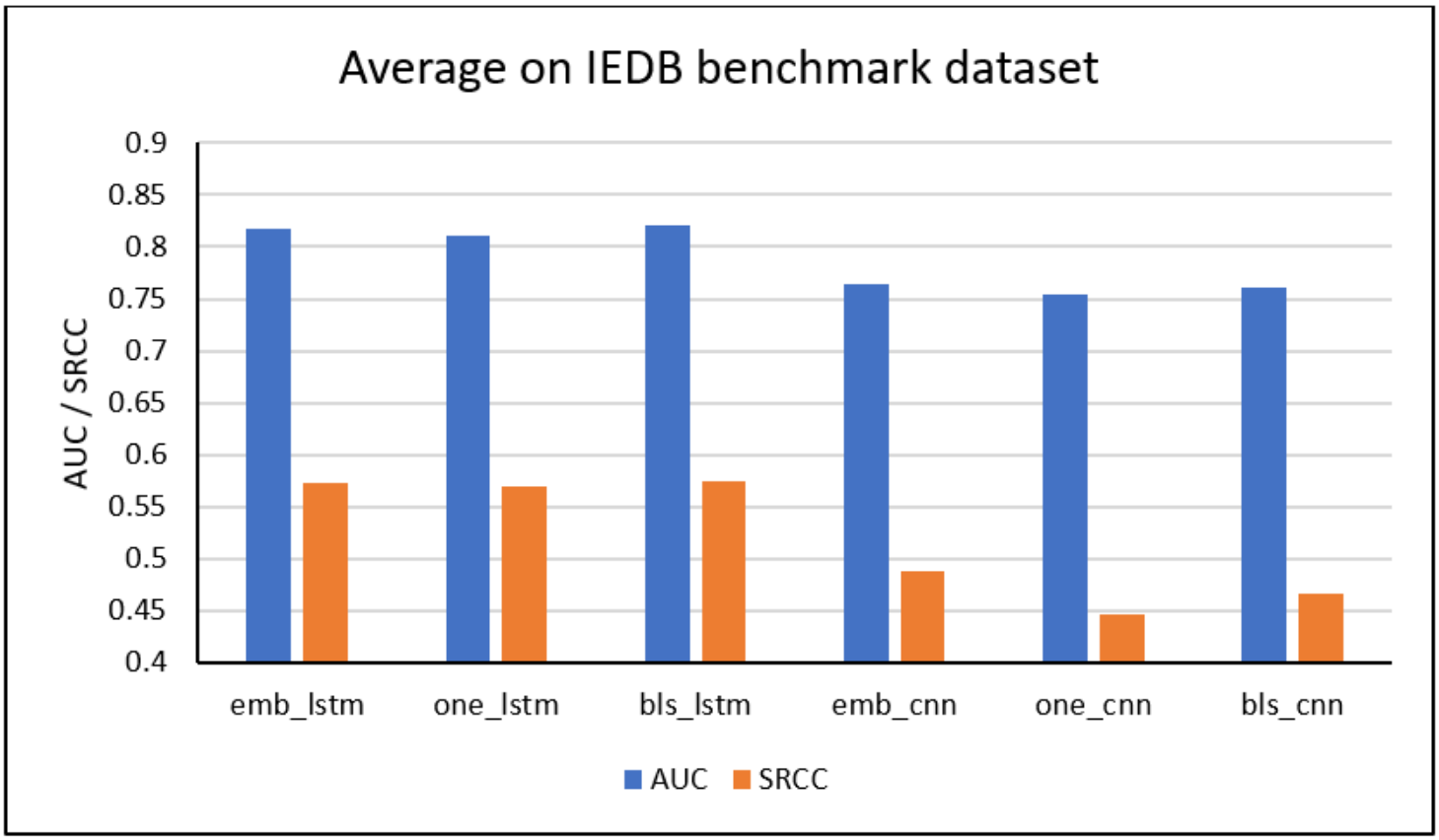


Figure 3

Performance evaluation of the six generated models on the IEDB benchmark dataset

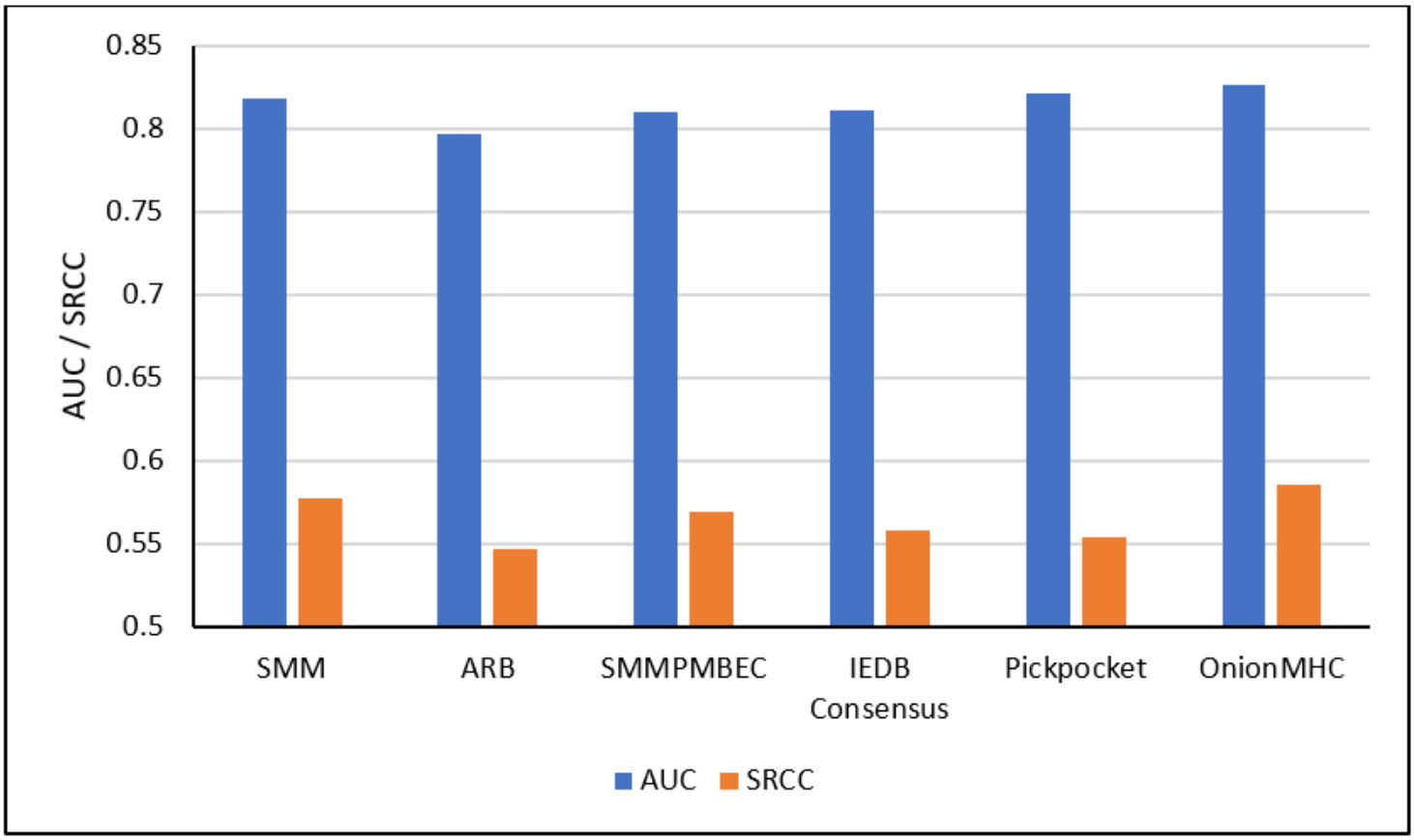


Figure 4

Performance comparison of OnionMHC with different prediction algorithms (Trolle et. al 2015) on the most recent IEDB benchmark dataset, measured in terms of Spearman Rank Correlation Coefficient (SRCC) and area under receiver operating curve (AUC)

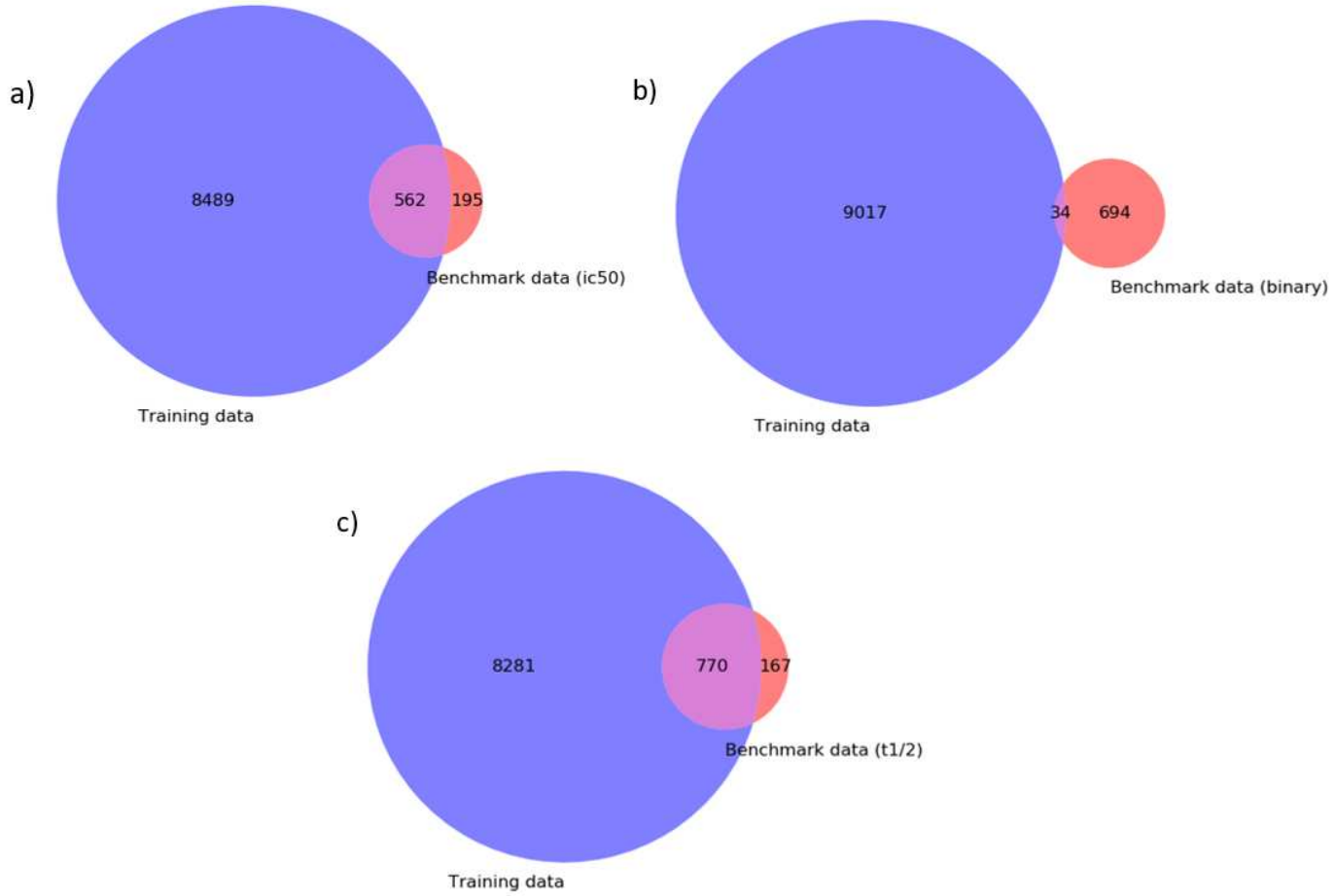


Figure 5

Overlapping peptides between the training data and benchmark data with measurement value: a) binary; b) ic50; c) half-life (t1/2)

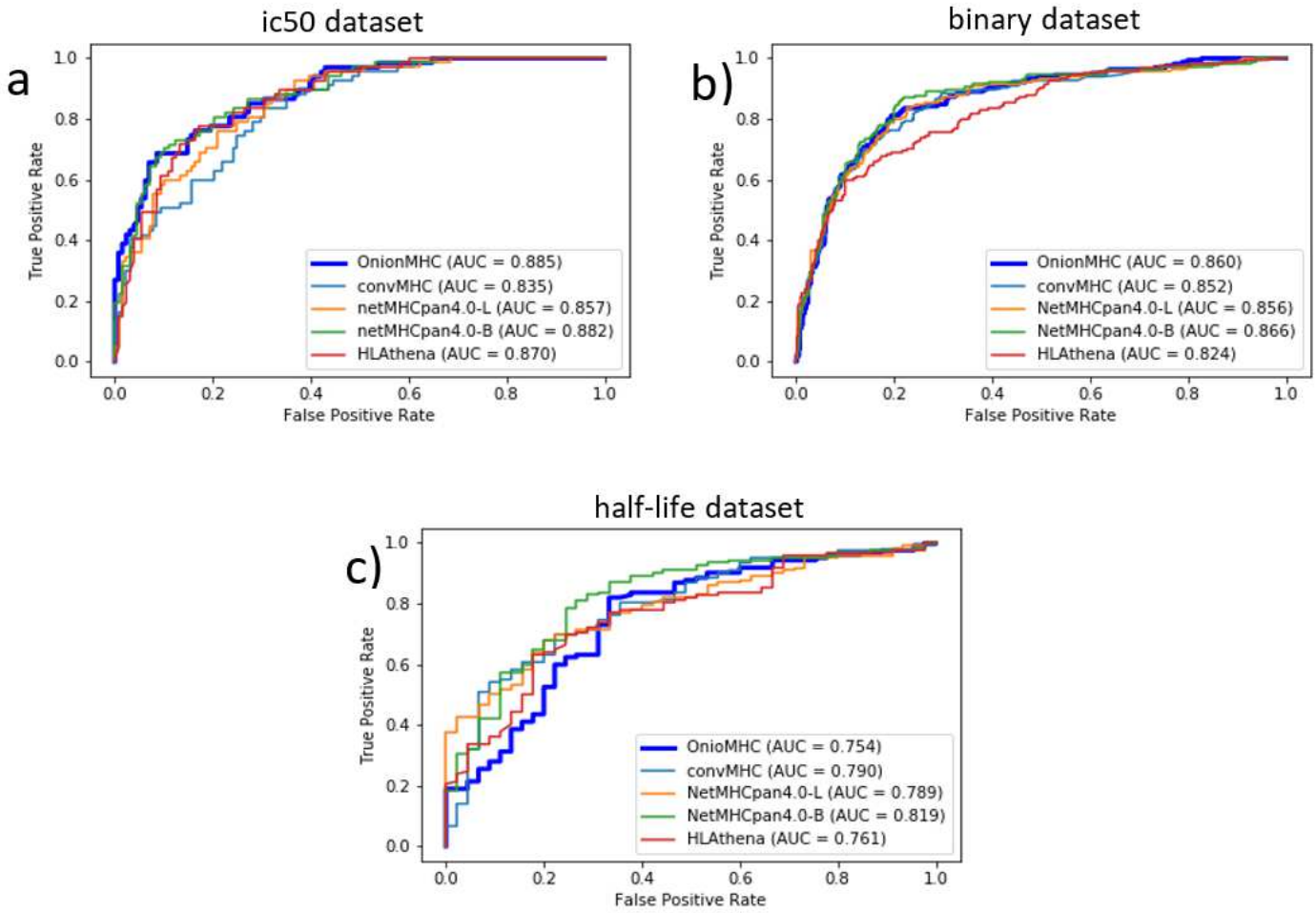


Figure 6

Area under the receiver operating curves for the non-overlapping peptides a) ic50 dataset (195 peptides) b) binary dataset (694 peptides) c) half-life dataset (167 peptides)

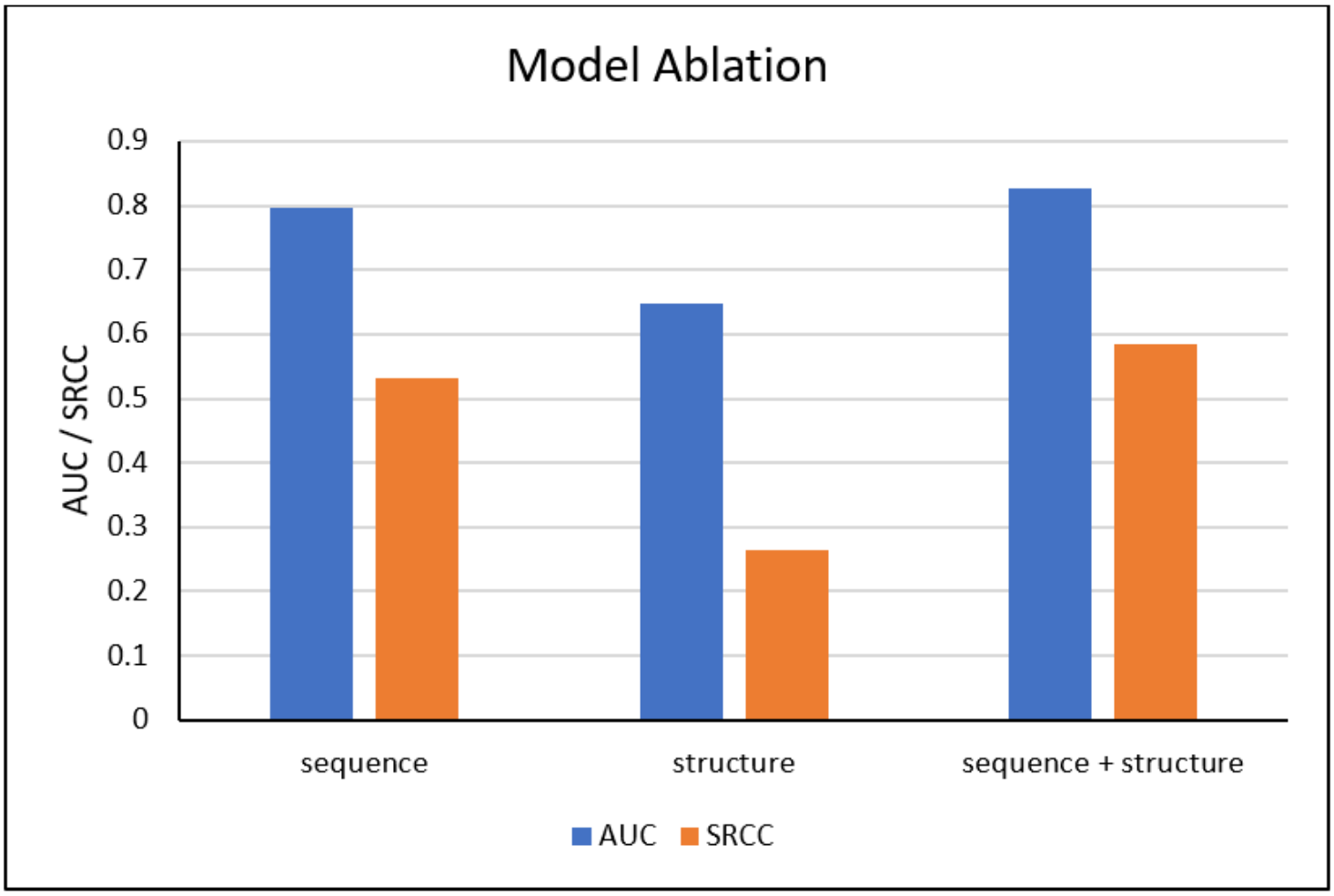


Figure 7

Model Ablation analysis of OnionMHC showing the individual performance of sequence and structure module on IEDB benchmark dataset

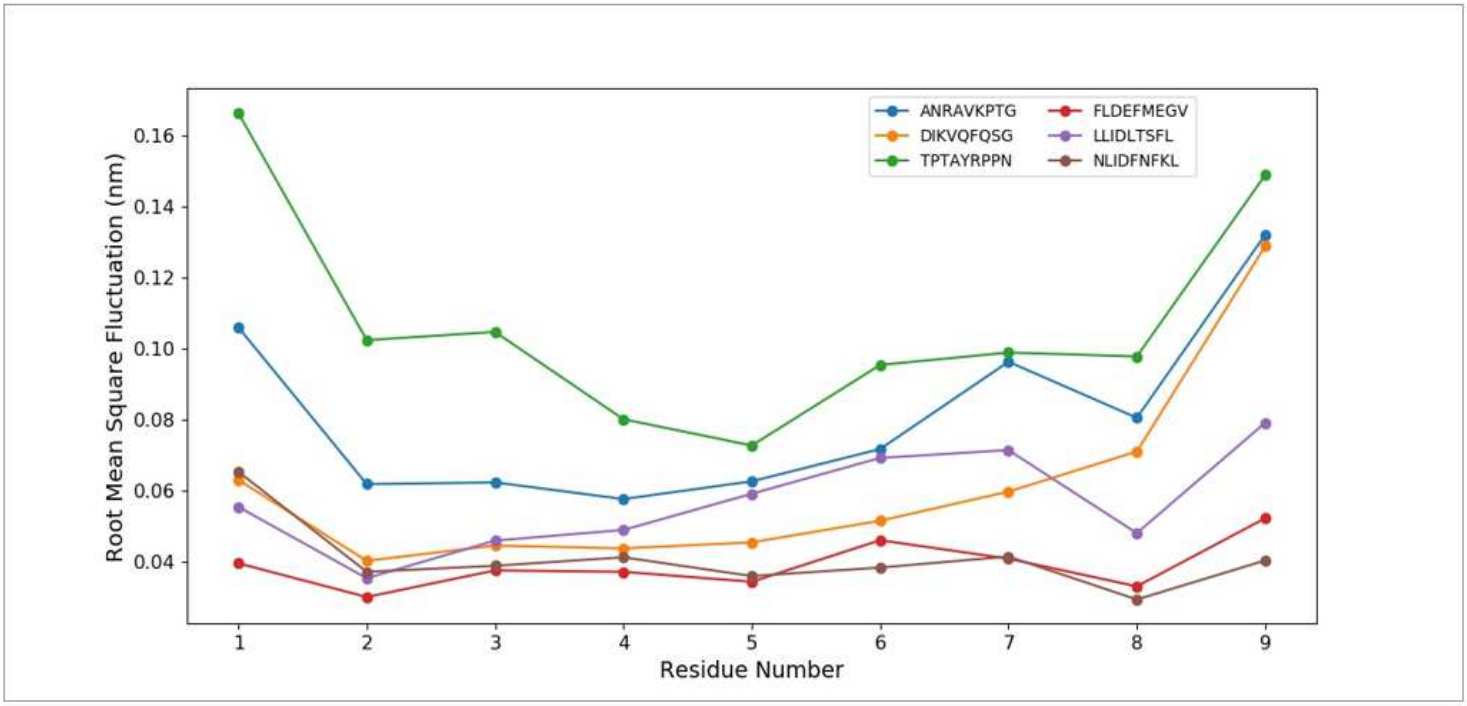


Figure 8

Root Mean Square Fluctuation (nm) of the residues in the peptide in the Molecular Dynamics simulation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.docx
- AdditionalFile2.xlsx
- AdditionalFile3.xlsx
- AdditionalFile4.xlsx