VSEPRnet: Physical structure encoding of sequence-based biomolecules for functionality prediction: Case study with peptides

Siddharth Rath[1,2,3], Jonathan Francis-Landau[1,4,&], Ximing Lu[1,5,&], Oliver Nakano-Baker[1],[2] Jacob Rodriguez[1,2], Burak Berk Ustundag[6], Mehmet Sarikaya[1,2,7,8*]

[1]GEMSEC, Genetically Engineered Materials Science and Engineering Center,

[2] Department of Materials Science and Engineering, University of Washington, Seattle, WA, 98195, USA

[3] Department of Nanotechnology and Molecular Engineering, University of Washington, Seattle, WA, 98195, USA

[4] Department of Mathematics, University of Washington, Seattle, WA, 98195, USA

[5] Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, 98195, USA

[6] Faculty of Computer and Informatics Engineering, Istanbul Technical University, Maslak, Istanbul, 34469, Turkey

[7] Department of Chemical Engineering, University of Washington, Seattle, WA, 98195, USA

[8] Department of Oral Health Sciences, University of Washington, Seattle, WA, 98195, USA

[*]Corresponding Author:

E-Mail: sarikaya@uw.edu

[&] These authors contributed equally to this work.

Abstract

Predicting structure-dependent functionalities of biomolecules is crucial for accelerating a wide variety of applications in drug-screening, biosensing, disease-diagnosis, and therapy. Although the commonly used structural “fingerprints” work for biomolecules in traditional informatics implementations, they remain impractical in a wide range of machine learning approaches where the model is restricted to make data-driven decisions. Although peptides, proteins, and oligonucleotides have sequence-related propensities, representing them as sequences of letters, e.g., in bioinformatics studies, causes a loss of most of their structure-related functionalities. Biomolecules lacking sequence, such as polysaccharides, lipids, and their peptide conjugates, cannot be screened with models using the letter-based fingerprints. Here we introduce a new fingerprint derived from valence shell electron pair repulsion structures for small peptides that enables construction of structural feature-maps for a given biomolecule, regardless of the sequence or conformation. The feature-map introduced here uses a simple encoding derived from the molecular graph - atoms, bonds, distances, bond angles, etc., that make up each of the amino acids in the sequence, allowing a Residual Neural network model to take greater advantage of information in molecular structure. We make use of the short peptides binding to Major-Histocompatibility-Class-I protein alleles that are encoded in terms of their extended structures to predict allele-specific binding-affinities of test-peptides. Predictions are consistent, without appreciable loss in accuracy between models for different length sequences, marking an improvement over the current models. Biological processes are heterogeneous interactions, which justifies encoding all biomolecules universally in terms of structures and relating them to their functionality. The

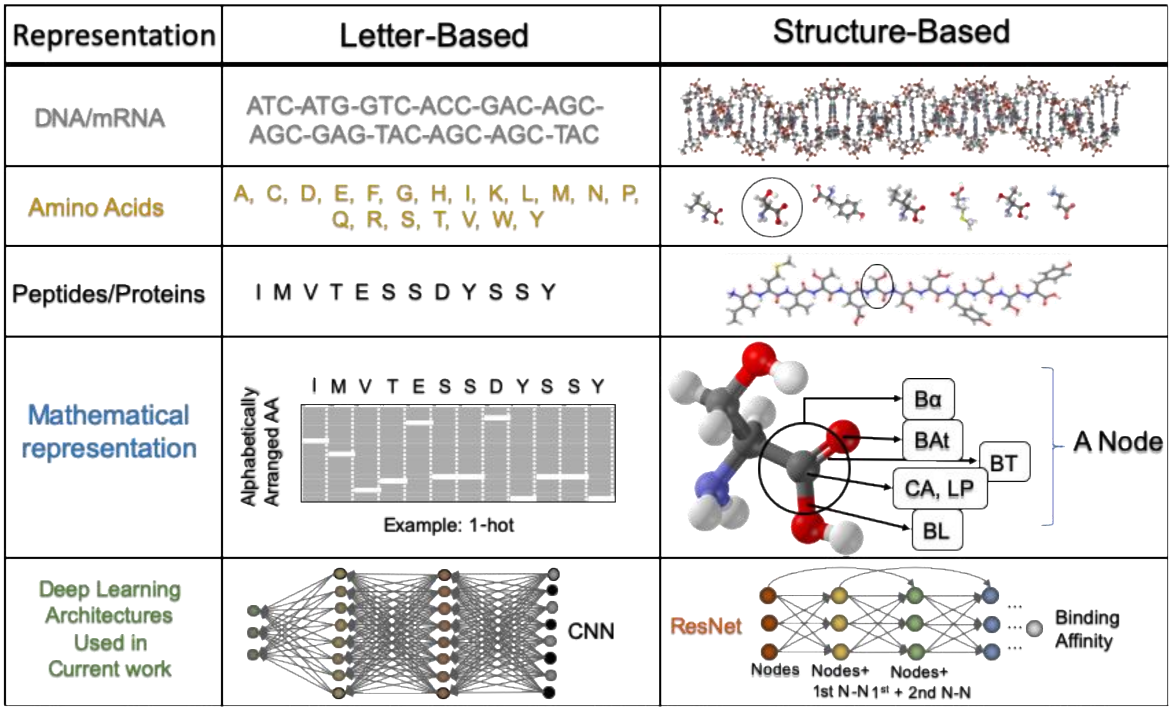
capabilities facilitated by the model expands the paradigm in establishing structure-function correlations among small molecules, short and longer sequences including large biomolecules, and genetic conjugates that may include polypeptides, polynucleotides, RNAs, lipids, peptidoglycans, peptido-lipids, and other biomolecules that could be implemented in a wide range of medical and nanobiotechnological applications in the future.

Introduction

Cheminformatics tools have been used to predict solubility, binding-affinity to receptors, toxicity, and other properties of small-molecules, which, for example, include Extended-Connectivity-Fingerprints (ECFP's) [1], Reduced-Graph representations [2], Simplified-Molecular-Input-Line-Entry-System (SMILES) [3], SMILES-Arbitrary-Target-Specification (SMARTS),[4] and International-Chemical-Identifier (InCHI) string analysis tools [5], Autoencoder implementations [6], Coulomb-matrices [7], Symmetry functions [8] and Graph-Convolutions [9,10]. Success of such tools have stimulated their implementation in bioinformatics. Graph-Convolution-Networks (GCN), where each amino-acid (AA) unit is considered as a node, has been used successfully on polypeptides as a classification tool in prediction of the protein-ligand interface [11]. Tools such as PotentialNet [12] that learn AA-connectivity of ligand binding sites have also been successfully implemented. The focus of such tools, however, has been on the small ligand and not the large biomolecular receptors. Additionally, a comprehensive structural feature-map is unavailable for proteins and peptides as neither the molecular structures nor their conformations are taken into consideration in the current GCNs. GCNs consider atom or AA connectivity for predicting properties of small-molecules. However, conformable biomolecules have connectivity beyond covalent bonds (such as hydrogen bonds) that are susceptible to changes based on the environmental and operational conditions. Tools directly employing three-dimensional coordinates as inputs to Neural Networks (NN) for small-molecule screening with integrated visualization-techniques have been developed [13]. However, the applications to biomacromolecules have been computationally intensive and currently impractical.

Traditional bioinformatics tools do not deal with small-molecules and are mostly concerned with AA sequences in proteins or oligonucleotide sequences in RNA and DNA. Letter-based representations are ubiquitous in addressing complicated functions owing to their simplicity, applicability, and accuracy in finding aligned domains in a sequence [14-17] or within a larger structure [18-20]. Several Machine Learning (ML) models to predict functionality using deep-learning, NNs, feature representation, and pattern analyses such as DeepMHC and NetMHCpan among others [21-23], have been developed by using the data in the Immuno-Epitope Database (IEDB) Analysis resource [24]. This database contains Major-Histocompatibility-Class-I, II (MHC-I, MHC-II) peptide-to-allele binding-affinity data for several species. In a recently developed Convolutional Neural Network (CNN), called DeepSeqPan [25], the authors recognize the importance of structural information in improving prediction accuracy and recommend their model as a supplement to other cumbersome models built with structural-alignment methods.

The traditional methodologies work only on letter-based AA or oligonucleotide labels and their derivations. The underlying physical-meaning, especially molecular structure or conformation is not apparent to the machine agent upon implementing ML algorithms. There is a loss of generalizability to include the molecules which do not have an obviously intrinsic sequence. Tools that work for or incorporate lipids, carbohydrates, and other biomacromolecules in their structures are exceedingly rare. Biological processes, however, are seldom isolated for a specific type of molecule, and commonly incorporate a wide range of biomolecules. Consequently, there is an imperative need for a method capable of encoding diverse biomolecules in a universal and meaningful manner (Fig 1) to study the interfacial phenomena at the molecular level. These processes may involve all biological systems, e.g., peptide and lipid or peptidoglycan [26], and biology/solid soft interfaces relevant to technological and nanomedicine applications [27].



CA: central Atom, LP: lone pair, BL: Bond length, BT: Bond Type, BAt: Bonded Atoms, Bα: Bond Angles, N-N: Nearest-Neighbors

Fig 1. Schematics show the differences between the letter-based and structure-based representation of biomolecules for ML studies in functionality prediction. The central column is the index while the middle column shows the letter-based representation and the rightmost column shows the structure-based representation.

Implementations of such ML tools could broaden the paradigm of drug-design, combating antibiotic resistance, restorative dentistry [28], disease-diagnostics, biocompatible-coatings, lab-on-chip technologies, and biosensors [29]. In this work, we demonstrate a comprehensive feature-map for peptides that can be generalizable to other biomolecules. The immediate goals of the current work have been, (a) To take any AA sequence and convert it to a VSEPR structure-based representation via a reversible transformation; (b) To decide on an NN model that takes neighborhood information and performs consistently well across different length sequences, and (c) To benchmark the model with respect to the model used in DeepMHC. The long-term goal is to establish groundwork for future research in developing an accurate, interpretable and generalizable feature-map that incorporates conformations and multiple biomolecules to study complex phenomena.

The binding-affinity obtained from the current study displays higher prediction accuracy for 10-AA long peptides than the one-hot encoded shallow CNN model from DeepMHC [23], while the reverse is true for 9-AA long peptides. 5-fold Cross-Validation (CV) remains consistent across 9-AA and 10-AA long sequences, a significant improvement compared to DeepMHC where there is an appreciable drop in predictive power between 9-AA and 10-AA sequences. Since the VSEPR implementation consists of a larger feature map in conjunction with a deep residual neural network (ResNet), there is some overfitting and a loss of interpretability. It is noted that including angles in a GCN would be more interpretable. Indeed, such a model is aimed as one of the next steps to be taken towards generation of precise and pan-specific predictive tools, generalizable to other biomolecules of interest in medical and technological applications.

Materials and Methods.

Data Cleaning and Preparation

Data compiled in 2013 from IEDB (www.iedb.org) Analysis Resource [24] are downloaded and cleaned. The binding affinities are measured in terms of Inhibitor Concentration IC_{50} required to reduce binding by half [30]. The values are converted to -ln(IC_{50}), as a normalization step. According to extant standards, any sequence with an IC_{50} less than or equal to 500 is labeled as a binder and the others labeled as non-binder for binary classification. The dataset is then interfaced with a Python script to extract peptide sequences with transformed binding-affinity values to any allele of interest from any species within the dataset. All human alleles with at least 1,000 different corresponding epitope sequences are used in this study. 20% of the sequences are frozen out of the dataset for testing the model. Remaining 80% of sequences for each allele are used as a training set. The peptides in each set are then converted into their VSEPR encoded fingerprints as described below.

VSEPR extended structure feature-map

As a first attempt, Bioluminate [31] is used to obtain the protein data bank (PDB) files for each of the naturally occurring AA. These PDB files contain information for each atom, including the data of atom type (in terms of atomic number) and cartesian coordinates of the given atom in space. The ProDy [32] library in python is used to traverse through the PDB files. Iterating through each of the neighbors of an atom, the bond type of each neighbor bonded to the central atom (CA) is obtained, based on prior knowledge of the AA structures. Euclidean distances are calculated to determine corresponding bond lengths. The number of lone pairs on any given CA is inferred based on the number of bonds and the bond-types that the given atom has, and its electronic structure. To calculate the angles made by pairs of Bonded-Atoms subtended at the CA, angle formula is used, and it is repeated parallelly for all CA and all combinations of Bonded-Atoms pairs per CA. Fig 2 shows the schematic of such a feature-map for Serine in an example peptide sequence.

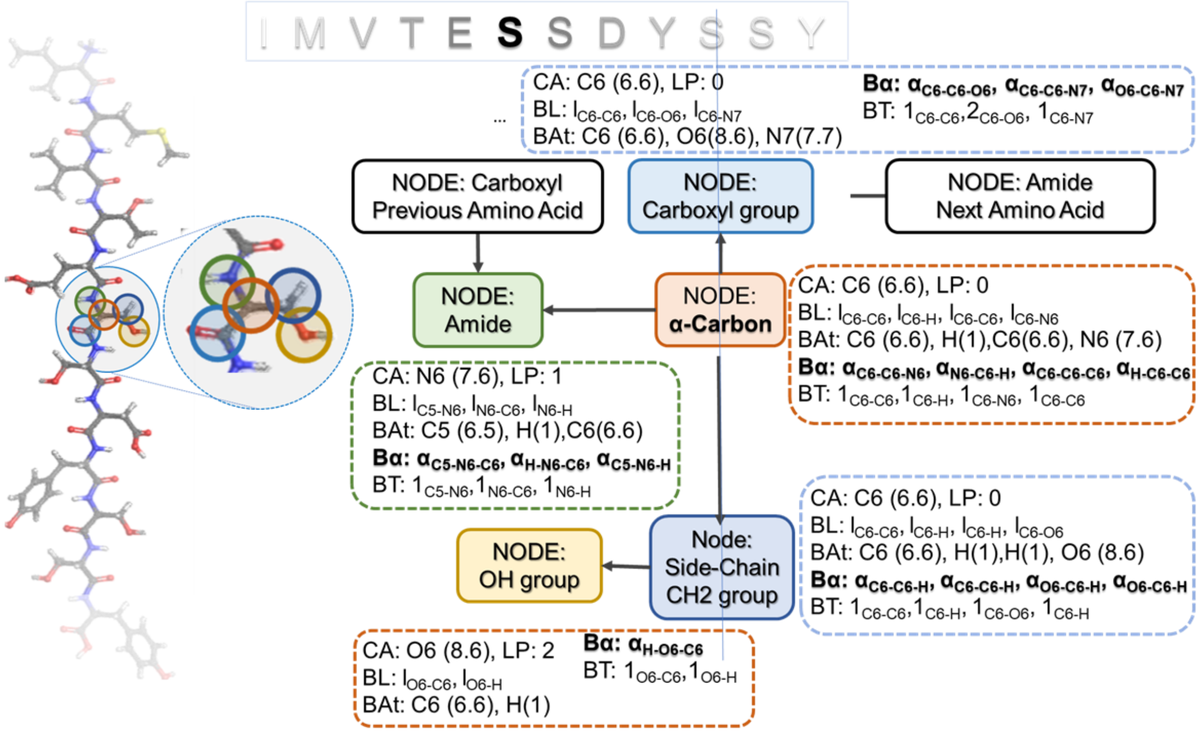

I M V T E S S D Y S S Y
...
CA: C6 (6.6), LP: 0
BL: $l_{C6\text{-}C6}$, $l_{C6\text{-}O6}$, $l_{C6\text{-}N7}$
BAt: C6 (6.6), O6(8.6), N7(7.7)
Bα: $\alpha_{C6\text{-}C6\text{-}O6}$, $\alpha_{C6\text{-}C6\text{-}N7}$, $\alpha_{O6\text{-}C6\text{-}N7}$
BT: $1_{C6\text{-}C6}$, $2_{C6\text{-}O6}$, $1_{C6\text{-}N7}$
NODE: Carboxyl Previous Amino Acid
NODE: Carboxyl group
NODE: Amide Next Amino Acid
NODE: Amide
NODE: α-Carbon
CA: C6 (6.6), LP: 0
BL: $l_{C6\text{-}C6}$, $l_{C6\text{-}H}$, $l_{C6\text{-}C6}$, $l_{C6\text{-}N6}$
BAt: C6 (6.6), H(1),C6(6.6), N6 (7.6)
Bα: $\alpha_{C6\text{-}C6\text{-}N6}$, $\alpha_{N6\text{-}C6\text{-}H}$, $\alpha_{C6\text{-}C6\text{-}C6}$, $\alpha_{H\text{-}C6\text{-}C6}$
BT: $1_{C6\text{-}C6}$, $1_{C6\text{-}H}$, $1_{C6\text{-}N6}$, $1_{C6\text{-}C6}$
CA: N6 (7.6), LP: 1
BL: $l_{C5\text{-}N6}$, $l_{N6\text{-}C6}$, $l_{N6\text{-}H}$
BAt: C5 (6.5), H(1),C6(6.6)
Bα: $\alpha_{C5\text{-}N6\text{-}C6}$, $\alpha_{H\text{-}N6\text{-}C6}$, $\alpha_{C5\text{-}N6\text{-}H}$
BT: $1_{C5\text{-}N6}$, $1_{N6\text{-}C6}$, $1_{N6\text{-}H}$
CA: C6 (6.6), LP: 0
BL: $l_{C6\text{-}C6}$, $l_{C6\text{-}H}$, $l_{C6\text{-}H}$, $l_{C6\text{-}O6}$
BAt: C6 (6.6), H(1),H(1), O6 (8.6)
Bα: $\alpha_{C6\text{-}C6\text{-}H}$, $\alpha_{C6\text{-}C6\text{-}H}$, $\alpha_{O6\text{-}C6\text{-}H}$, $\alpha_{O6\text{-}C6\text{-}H}$
BT: $1_{C6\text{-}C6}$, $1_{C6\text{-}H}$, $1_{C6\text{-}O6}$, $1_{C6\text{-}H}$
NODE: OH group
Node: Side-Chain CH2 group
CA: O6 (8.6), LP: 2
BL: $l_{O6\text{-}C6}$, $l_{O6\text{-}H}$
BAt: C6 (6.6), H(1)
Bα: $\alpha_{H\text{-}O6\text{-}C6}$
BT: $1_{O6\text{-}C6}$, $1_{O6\text{-}H}$

Fig 2. Schematic of Valence Shell Electron Pair Repulsion structural feature-map for bioinformatics studies. Green: N-terminus/Connection from previous Amino-Acid, Orange: Alpha-Carbon, Dark-Blue: Functional groups in side-chain, and Light-Blue: Connection to next Amino-Acid/C-terminus. Each such node contains 5 channels of information: Central Atom (CA) with associated Lone Pairs (LP), Bond lengths (BL), Bonded Atoms (BAt), Bond Types (BT) and Bond Angles (Bα).

The information for a given CA is appended to all successive non-hydrogen CA’s starting from the N-terminus of the peptide and ending at the C-terminus. Each type of parameter obtained from the VSEPR extended structure, is input as a separate channel of data to the neural network for training without overlap. The tenths place-value of the atomic number of the CA is the index of the residue location and the hundredths place-value is the location within the residue. For example, the α-Carbon in the 1st AA at the N-terminus is given a value 6.01, whereas the carbon at the center of the planar carboxyl group bonded to the amine group of the 2nd AA, is given a value of 6.00 (See S1 Appendix for more details).

The symbolic-connectivity reduces dimensionality but increases information bandwidth. It means that there are now two non-linear data bands in terms of power of 10. One band is 10^{-2} and the other is 10^{-1} in this case. Since the bands do not overlap, owing to the channel-splitting, machine learning methods also work as long as there are enough hidden layers to fit the respective non linearity levels. This is a way of multiplexing three separate inputs into one. Future implementations will eliminate this input through analytical transformation that only affects linear part of dominating input parameters. Binary vectorization of the encoding will also be attempted since power of two is more flexible instead of power of 10, in management of information bandwidth. Nevertheless, incorporating conformations as well as using adjacency matrices in a GCN is the clear next step towards making VSEPR methodology more impactful.

Neural Network Architecture

Since behavior of molecular components of peptides depends on their neighborhood, Residual Neural Network (ResNet) was chosen to incorporate such information. The schematic of the process is shown in Fig 3. Such a Neural Network architecture comprises of a convolution block called the Residual Convolutional Unit (RCU) which performs a set of convolutions on the channels and a Fully Connected (FC) block. The RCU is implemented in terms of an Efficient Spatial Pyramid (ESP) [33]. ESP in the RCU allows for an improved gradient flow for training the network and essentially makes each atom 'see' its neighbors.

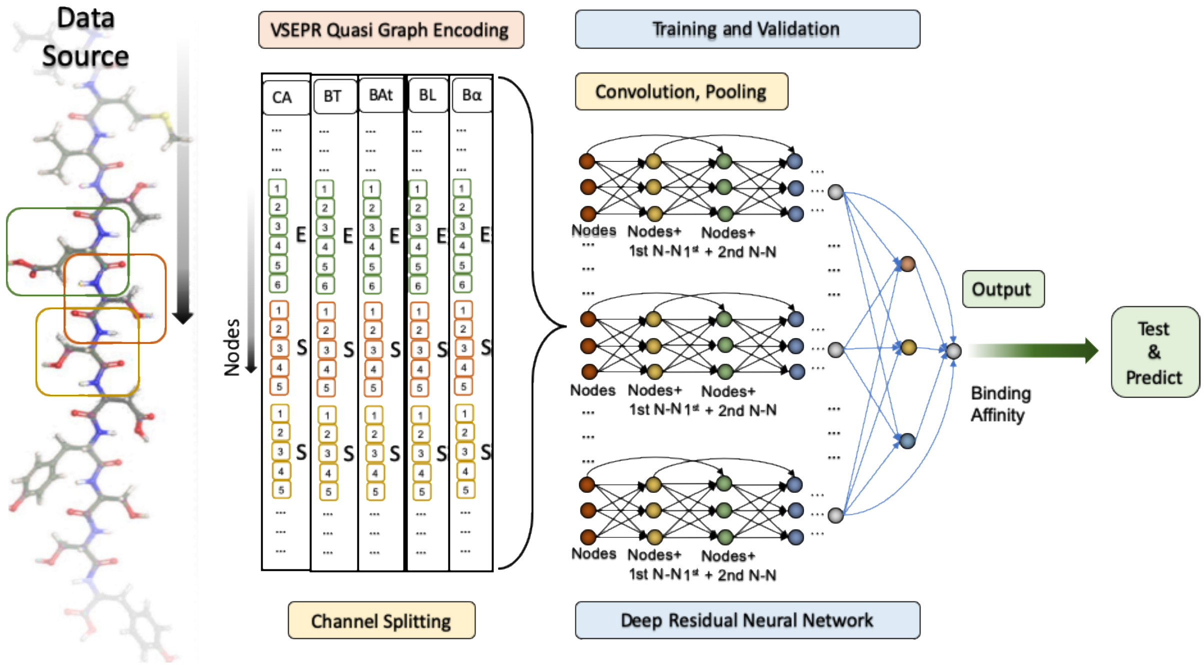


Fig 3. Schematic of the Training and Validation with the ResNet Architecture. In the convolution block, convolution proceeds on all atomic nodes simultaneously, with each successive layer seeing effects from more neighbors. Features thus extracted are sent through a fully connected network for prediction. The prediction can be carried out on any function that can be represented in terms of a numerical value. Here we predict the -ln(IC50) binding affinity.

The outputs of the ESP enhanced RCU block are then passed into the FC block, with a Rectified-Linear-Unit activation on all the layers and SoftMax on the last. Mean-Squared-Error is the loss function to be minimized to output the binding affinity of the peptide to the corresponding MHC-I allele. Batch Normalization is performed after every layer in the network. Randomly initialized weights are then learned in a supervised learning protocol and hyperparameters are tuned following a training process as described below.

Training, Validation and Testing.

Sequences for each allele in the training-set are divided into five equal parts randomly selected, to set the stage for a 5-fold cross-validation as a control against sampling bias. Four out of five such parts are used to train the model and the fifth one is used for testing. Then the model rotates through another set of four such parts as training and fifth one as test set. In each such model training round, per allele, each of the feature-maps are split into 5 channels per input sequence. They are sent in simultaneously in mini-batches of 20 peptides at a time into the ResNet described above, for 5000 epochs. The PyTorch [34] deep-learning library is used for training. The model is labeled 'converged', if validation loss (10% of the training data is used for validation) did not reduce by more than 1% for 100 subsequent epochs.

After the training is completed, hyperparameters are tuned to maximize the 5-fold cross validation resulting in a learning rate of 5e^{-4}. The process is repeated three times to ensure that the cross-validations observed are consistent and not affected by choice of training samples. A similar procedure is followed to train a regular Convolutional neural Network with one-hot encoded peptide sequences as in DeepMHC for one-to-one comparison and evaluation. Meanwhile, the 20% of data frozen before training is then used as a blind test set for evaluating model performance.

Results and Discussions

We compare the allele-specific VSEPRnet model with the state-of-the-art CNN model, DeepMHC that works with letter-based fingerprints. The 5-fold CV results as obtained by the reproduced DeepMHC model versus the current VSEPRnet model is shown in Fig 4A and 4B for 9-AA and 10-AA long sequences respectively. Results show a consistent response across sequence lengths in the VSEPRnet case in contrast to DeepMHC, where there is a fall in prediction accuracy for 10-AA sequences (refer S1 Fig). In the case of DeepMHC, the average 5-fold CV (Fig 4C) across all alleles studied is 0.87 for 9-AA sequences, with a standard-deviation of 0.03. For 10-AA sequences it is 0.65 with a standard deviation of 0.11. For VSEPRnet, the average 5-fold CV for 9-AA long peptides is 0.74 with a standard deviation of 0.06. While for 10-AA long peptides it is 0.69 with a standard-deviation of 0.03. Taking available data and overfitting into consideration, VSEPRnet therefore has a consistency in predictability over sequence lengths. One of the reasons for a marked fall in cross validation for 10-length sequences, as outlined in DeepMHC, is a dependency of the model on distal effects which dominate as lengths increase. We note that because feature-maps and neural-network architectures usually go hand-in-hand, further investigation is mandated to isolate the cause of the flattening response observed in the case of VSEPRnet. However, due to the nature of the ESP convolution block in the ResNet architecture, distal effects in the convolution may not dominate. Moreover, the distinction in input sizes between 9-AA and 10-AA peptides is based on physical rather than sequence length.

Since the VSEPR feature-map contains more information than the one hot encoding, the data required to avoid over-fitting becomes higher. Thus, the lack of requisite data-density lowers the average 5-fold CV from 0.87 (DeepMHC) to 0.74 (VSEPRnet) for the 9-AA long peptides. As discussed previously, there is a role-reversal for the 10-AA case because there is a pronounced distal-effect in the DeepMHC implementation whereas it is negligible for the VSEPRnet implementation (see S2 Appendix for more details). The overall performance of VSEPRnet in terms of 5-fold CV is contingent mostly on the available data-points to train on. Future work could be directed to implement the model on datasets with higher density of data obtained from High Throughput Sequencing techniques [35].

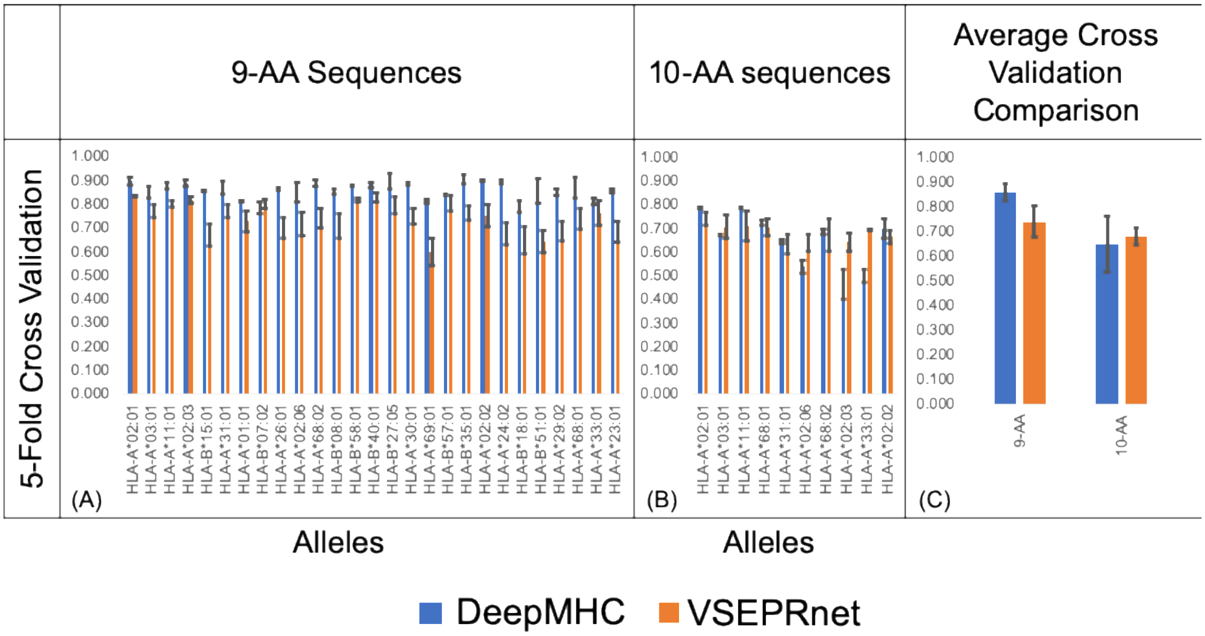


Fig 4. 5-Fold Cross Validation results from VSEPRnet compared with DeepMHC. (A) 5-fold CV for 9-Length peptide sequences and (B) 5-fold CV for 10-Length peptide sequences; (C) Average 5-fold CV for 9 and 10-AA peptides across alleles. Performance of VSEPRnet falls in comparison to DeepMHC in the 9-Length peptides case for most alleles, the most probable reason being overfitting due to increased dimensionality. The performance of VSEPRnet is better than DeepMHC in case of 10-Length peptides on most alleles due to reduced dominance of distal

effects. Overall, VSEPRnet performs consistently across sequence lengths and does not have the drop in accuracy between 9 and 10-AA peptides as is the case with DeepMHC.

Performance Comparison of VSEPRnet and DeepMHC on previously frozen test data, uses Pearson Coefficient (PC), Spearman Rank Correlation Coefficient (SRCC) and Area Under receiver operating Curve (AUC) as metrics to compare the performance of the two models. For the PC metric, VSEPRnet wins on 11 out of 27 alleles in the 9-AA case and 5 out of 10 alleles in the 10-AA case; for the SRCC metric, VSEPRnet wins on 8 out of 27 alleles in the 9-AA case and 4 out of 10 alleles in the 10-AA case; And for the AUC metric, VSEPRnet wins or performs equally on 7 out of 27 alleles in the 9-AA case and wins on 9 out of 10 alleles in the 10-AA case. Across all metrics, the performance of VSEPRnet is within the first standard deviation of DeepMHC for 9-AA peptides, and for 10-AA peptides, VSEPRnet wins on both PC and AUC metrics. The average PC across all alleles for DeepMHC is 0.244 with a standard deviation of 0.112 for 9-AA peptides, and 0.279 with a standard deviation of 0.112 for 10-AA peptides. The average PC for VSEPRnet is 0.235 with a standard deviation of 0.096 for 9-AA peptides and 0.296 with a standard deviation of 0.042 for 10-AA peptides. Similarly, across all tested alleles, the average SRCC of DeepMHC is 0.6 with a standard deviation of 0.140 for 9-AA peptides and 0.514 with a standard deviation of 0.165 for 10-AA peptides, while the average SRCC, across all tested alleles for VSEPRnet is 0.571 with a standard deviation of 0.099 for 9-AA peptides and 0.5 with a standard deviation of 0.046 for 10-AA peptides (See Fig 5 and S2 Fig for more details).

Additionally, across all tested alleles, the average AUC of DeepMHC is 0.795 with a standard deviation of 0.072 for 9-AA peptides and 0.684 with a standard deviation of 0.072 for 10-AA peptides, while the average AUC, across all tested alleles for VSEPRnet is 0.767 with a standard deviation of 0.062 for 9-AA peptides and 0.745 with a standard deviation of 0.025 for 10-AA peptides. A consistent response across alleles is also shown by the VSEPRnet, without being affected by the sequence length of the peptides.

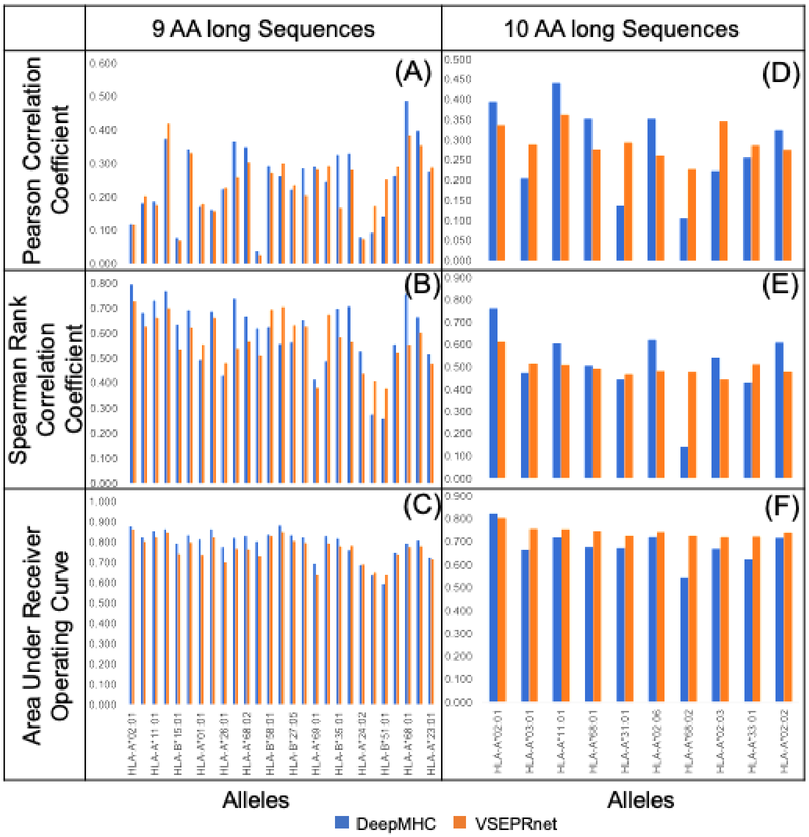


Fig 5. Comparison of DeepMHC and VSEPRnet on test data for 9-AA and 10-AA long peptides. DeepMHC performs consistently better on the (A) PC, (B) SRCC and (C) AUC metric for 9-AA long peptides because of lower probability of overfitting due to low information content of the 1-hot encoding. VSEPRnet PC values are within the mean and spread of the PC values for DeepMHC. For 10-AA, VSEPRnet performs equally as well as DeepMHC in case of (D) PC and (E) SRCC, while performing consistently better for (F) AUC owing to the elimination of distal effects.

Conclusions and Future Work

VSEPRnet is an introductory implementation for extending cheminformatics style feature-maps to bioinformatics studies while maintaining generalizability across lengths and molecule-types. There is a demonstrated consistency in prediction-accuracy of VSEPRnet model across alleles and between 9-AA to 10-AA long peptides binding to MHC-I allele. Therefore, there are advantages of using this implementation as a first step in generalization of feature-maps to include other molecules. There is a need to incorporate conformations and substrate information into the model to make it truly generalizable to DNA, RNA, proteins, peptides, intrinsically-disordered regions, lipids, peptidoglycans, phospholipids, sugars, and smaller biomolecules such as vitamins and co-factors.

Since the VSEPRnet 5-fold CV does not show appreciable dependency on distal-effects, there are available strategies to further improve the displayed generalizability of the model. The strategies are: (a) Binary-vectorizing the input without overlap between the channels; (b) Incorporating angular information into GCN; (c) Implementation on high density datasets; (d) Appending error modulating layers downstream; and (e) Incorporating allele information to generalize the VSEPRnet to a pan-specific model. It is also worthy of noting that because the structures of the functional groups are encoded in VSEPRnet, this is applicable to lipids, peptidoglycans, polynucleotides, small molecules, sugars, etc. As long as the size of the molecule is within the limits of the training set, peptide data may be used to train the model while using a small set of peptidoglycans as the test set, for example.

The data and scripts for all the above steps including model building and training are available on GitHub (https://github.com/Sarikaya-Lab-GEMSEC).

Acknowledgments

We acknowledge the guidance of Kevin Jamieson (Paul G Allen School of Computer Science and Engineering), Marina Meila (Statistics), René Overney (Chemical Engineering and Molecular Engineering and Science Institute) and Deniz T. Yucesoy, (at GEMSEC), all at the University of Washington.

References

1. Rogers D, Hahn M. Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling. 2010;50(5):742-54.
2. Takahashi Y, Sukekawa M, Sasaki S. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. Journal of Chemical Information and Modeling. 1992;32(6):639-43.
3. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. Journal of Chemical Information and Modeling. 1989;29(2):97-101.
4. Proceedings of the 1997 1st Electronic Packaging Technology Conference (Cat. No.97TH8307). Proceedings of the 1997 1st Electronic Packaging Technology Conference (Cat No 97TH8307) EPTC-97: IEEE; 1997.
5. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI - the worldwide chemical structure identifier standard. Journal of Cheminformatics. 2013;5(7); 1-9.

6. Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H. Application of Generative Autoencoder in De Novo Molecular Design. Molecular Informatics. 2017;37(1-2):1700123.

7. Neese F. An improvement of the resolution of the identity approximation for the formation of the Coulomb matrix. Journal of Computational Chemistry. 2003;24(14):1740-7.

8. Behler J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. The Journal of Chemical Physics. 2011;134(7):074106.

9. Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, Duvenaud D, Maclaurin D, Blood-Forsythe MA, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nature Materials. 2016;15(10):1120-7.

10. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. Journal of Computer-Aided Molecular Design. 2016;30(8):595-608.

11. Shariat B, Neumann D, Ben-Hur A. BLRM: A Basic Linear Ranking Model for Protein Interface Prediction. IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2018 Dec 3 (pp. 29-35).

12. Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, et al. PotentialNet for Molecular Property Prediction. ACS Central Science. 2018;4(11):1520-30.

13. An Y, Sherman W, Dixon SL. Kernel-Based Partial Least Squares: Application to Fingerprint-Based QSAR with Model Visualization. Journal of Chemical Information and Modeling. 2013;53(9):2312-21.

14. Dayhoff MO, Schwartz RM, Orcutt BC. 22 a model of evolutionary change in proteins. InAtlas of protein sequence and structure 1978 (Vol. 5, pp. 345-352). National Biomedical Research Foundation Silver Spring.

15. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. 1992;89(22):10915-9.

16. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970;48(3):443-53.

17. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981;147(1):195-7.

18. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology. 1995;247(4):536-40.

19. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Research. 2016;45(D1):D289-D95.

20. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. A database of protein structure families with common folding motifs. Protein Science. 1992;1(12):1691-8.

21. Vang YS, Xie X. HLA class I binding prediction via convolutional neural networks. Bioinformatics. 2017;33(17):2658-65.

22. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics. 1998;14(2):121-30.

23. Hu J, Liu Z. DeepMHC: Deep convolutional neural networks for high-performance peptide-MHC binding affinity prediction. bioRxiv. 2017 Jan 1:239236.

24. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, et al. Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Research. 2008;36(Web Server):W513-W8.

25. 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. Scientific Reports. 2019 Jan 28;9(1):794.

26. Thiam K, Loing E, Verwaerde C, Auriault C, Gras-Masse H. IFN-γ-derived lipopeptides: influence of lipid modification on the conformation and the ability to induce MHC class II expression on murine and human cells. Journal of medicinal chemistry. 1999 Sep 9;42(18):3732-6.

27. Yamankurt G, Berns EJ, Xue A, Lee A, Bagheri N, Mrksich M, Mirkin CA. Exploration of the nanomedicine-design space with high-throughput screening and machine learning. Nature Biomedical Engineering. 2019, 3(1), 318–327.

28. Dogan S, Fong H, Yucesoy DT, Cousin T, Gresswell C, Dag S, Huang G, Sarikaya M. Biomimetic tooth repair: amelogenin-derived peptide enables in

vitro remineralization of human enamel. ACS Biomaterials Science & Engineering. 2018 Mar 9;4(5):1788-96.

29. Hayamizu Y, So CR, Dag S, Page TS, Starkebaum D, Sarikaya M. Bioelectronic interfaces by spontaneously organized peptides on 2D atomic single layer materials. Scientific Reports. 2016 Sep 22;6:33778.

30. Yung-Chi C, Prusoff WH. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochemical pharmacology. 1973 Dec 1;22(23):3099-108.

31. Bhachoo J, Beuming T. Investigating Protein–Peptide Interactions Using the Schrödinger Computational Suite. InModeling Peptide-Protein Interactions 2017 (pp. 235-254). Humana Press, New York, NY.

32. Bakan A, Meireles LM, Bahar I. ProDy: Protein Dynamics Inferred from Theory and Experiments. Bioinformatics. 2011;27(11):1575-7.

33. Mehta S, Rastegari M, Caspi A, Shapiro L, Hajishirzi H. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. Computer Vision – ECCV 2018: Springer International Publishing; 2018. p. 561-80.

34. Mishra P. Introduction to Neural Networks Using PyTorch. InPyTorch Recipes 2019 (pp. 111-126). Apress, Berkeley, CA.

35. Metzker ML. Sequencing technologies—the next generation. Nature reviews genetics. 2010 Jan;11(1):31.

Supporting information

S1 Appendix. Description of Channel Inputs to VSEPRnet. This section describes the information obtained from VSEPR structures of peptides that is sent through each of the 5 channels into the neural network.

S1 Fig. 5-fold CV data across all alleles. The 5-fold CV of training set peptides for 9 and 10-Amino Acid long sequences, and their means and standard deviations are tabulated for DeepMHC and VSEPRnet.

S2 Appendix. Model Comparison of dependency of 5-fold CV on available training data. This section describes the dependency of 5-fold CV's obtained from the VSEPRnet and DeepMHC models on available training data.

S2 Fig. PC, SRCC and AUC metrics from test set. The Pearson Correlations, Spearman Rank Correlation Coefficients, and Area Under the Curve of test peptides for 9 and 10-Amino Acid long sequences, for DeepMHC and VSEPRnet implementations and their means and standard deviations are tabulated.