S1. Table: The division of amino acid groups based on physicochemical properties and amino acid indices.

| Attributes | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Hydrophobicity | Polar<br>Q, E, R, K, D, N | Neutral<br>G,P, H, A, S, T, Y | Hydrophobic<br>C, V, F, L, I, M, W |
| Polarizability | 0-1.08<br>S, D, G, A, T | 0.128-0.186<br>C, Q, I, P, N, V, E, L | 0.219- 0.409<br>Y, M, K, R, H, F, W |
| Normalized Van der Waals | 0-2.78<br>S, C, G, A, T, P, D | 2.95-4.0<br>E, Q, N, V, I, L | 4.0-8.1<br>K, F, M, H, R, Y, W |
| Polarity | 4.9-6.2<br>W, C, L, I, F, M, V, Y | 8.0-9.2<br>T, G, P, A, S | 10.4-13.0<br>K, N, H, Q, R, E, D |
| Solvent Accessibility | Buried<br>A, I, F, C, G, L, V, W | Exposed<br>R, K, Q, E, N, D | Intermediate<br>M, S, P, T, H, Y |
| Secondary Structure | Helix<br>E, A, L, M, Q, K, R, H | Strand<br>V, I, Y, C, W, F, T | Coil<br>G, N, P, S, D |
| Charge | Positive<br>K, R | Neutral<br>A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V | Negative<br>D, E |

S2. The parameter setting of traditional machine learning methods and DNN models for epitope prediction of human-adapted viruses.

For traditional machine learning methods, the parameters are optimized, and the values are shown below.

Random Forest (n_estimators=100, criterion='gini', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', min_impurity_decrease=0.0)

Neural Network (hidden_layer_sizes=(50,50,50), activation='relu', solver='adam', alpha=0.001, batch_size='auto',learning_rate='adaptive',learning_rate_init=0.001,power_t=0.5,max_iter=200)

Gaussian Naive Bayes (var_smoothing=1e-09)

K-nearest neighbors (n_neighbors=8, weights='distance', algorithm='auto', leaf_size=30, p=1, metric='minkowski')

Support Vector Machine (C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, max_iter=- 1, decision_function_shape='ovr')

XGBoost (base_score=0.5, colsample_bytree=0.85, gamma=0.2, importance_type='gain', learning_rate=0.05, max_delta_step=0, max_depth=9, min_child_weight=3, n_estimators=180, num_parallel_tree=1, random_state=283, reg_alpha=0.05, reg_lambda=1, scale_pos_weight=1, subsample=0.75, verbosity=1)

For VGG and AlexNet, we experimented on the batch size = 2, learning rate = 0.0001, and dropout = 0.1, epoch = 100. For SqueezeNet, we experimented on the batch size = 8, learning rate = 0.00005, and dropout = 0.1, epoch = 100. For RNN-based models, we experimented on the batch size = 2, learning rate = 0.0005, and dropout = 0.1, epoch = 100.

S3 Table: The detailed description of compared methods for viral epitope prediction.

| Method | Description | Length of epitope | Reference |
|---|---|---|---|
| BepiPred 2.0 | BepiPred 2.0 predicts b-cell epitope using Random Forest algorithm trained on epitope and non-epitope amino acids determined from crystal structures. The residual is predicted as part of epitope when the score is above the threshold. | >=10 | [6] |
| LBtope | LBtope predicts variable length b-cell epitope using SVM algorithm by constructing dipeptide vectors. | All length | [7] |
| Parker Hydrophilicity prediction | This method uses hydrophilic scale, which is based on peptide retention times during high-performance liquid chromatography (HPLC) on a reversed-phase column. A window of seven residues was used for analyzing epitope region. | >7 | [5] |
| Chou and Fasman beta turn prediction | According to [9], it is reasonable to predict antibody epitopes by predicting turns. In this case we used the Chou and Fasman scale in order to implement the turn scale of that paper which has some non-standard properties | >7 | [1][9] |
| Emini surface accessibility Prediction | This method is based on surface accessibility scale, obtained from $S_n = 0.37^{-6} \prod_{i=1}^{6} \delta_{n+4-i}$, where Sn is the surface probability, $\delta$ x is the fractional surface probability for the amino acid at position x. | >7 | [2] |
| Kolaskar and Tongaonkar antigenicity Prediction | This is a semi-empirical method based on physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes to predict antigenic determinants on proteins | >7 | [4] |
| Karplus and Schulz Flexibility Prediction | This method uses flexibility scale, based on the temperature factors, i.e., B values, of the C$\alpha$ atoms. | >7 | [3] |
| AAP | AAP antigenicity scale was based on the ratio of occurrence frequency of amino acid pairs in the positive set compared to the negative set. | 15 | [8] |
| DPC | Dipeptide composition (DPC) is represented by a vector specifying the percentage of composition of all dipeptide type | All length | [10] |

## Reference


[1] Chou, P Y, and G D Fasman. “Prediction of the secondary structure of proteins from their amino acid sequence.” Advances in enzymology and related areas of molecular biology vol. 47 (1978): 45-148. doi:10.1002/9780470122921.ch2

[2] Emini, E A et al. “Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide.” Journal of virology vol. 55,3 (1985): 836-9. doi:10.1128/JVI.55.3.836-839.1985

[3] Karplus, P. Andrew and Georg E. Schulz. “Prediction of chain flexibility in proteins.” Naturwissenschaften 72 (2005): 212-213.

[4] Kolaskar, A S, and P C Tongaonkar. “A semi-empirical method for prediction of antigenic determinants on protein antigens.” FEBS letters vol. 276,1-2 (1990): 172-4. doi:10.1016/0014-5793(90)80535-q

[5] Parker, J M et al. “New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites.” Biochemistry vol. 25,19 (1986): 5425-32. doi:10.1021/bi00367a013

[6] Jespersen, Martin Closter et al. “BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes.” Nucleic acids research vol. 45,W1 (2017): W24-W29. doi:10.1093/nar/gkx346

[7] Harinder Singh, Hifzur Rahman Ansari, and Gajendra P. S. Raghava. Improved methodfor linear b-cell epitope prediction using antigen’s primary sequence.PLOS ONE, 8(5):1–8,05 2013

[8] Chen, J et al. “Prediction of linear B-cell epitopes using amino acid pair antigenicity scale.” Amino acids vol. 33,3 (2007): 423-8. doi:10.1007/s00726-006-0485-9

[9] Pellequer, J L et al. “Correlation between the location of antigenic sites and the prediction of turns in proteins.” Immunology letters vol. 36,1 (1993): 83-99. doi:10.1016/0165-2478(93)90072-a

[10] Yu, Chin-Sheng, Chih-Jen Lin, and Jenn-Kang Hwang. "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions." Protein science 13.5 (2004): 1402-1406.