Machine Learning in T- and B-Cell Epitope Prediction

Subjects: Computer Science, Interdisciplinary Applications | Immunology

Contributor: Syed Nisar Hussain Bukhari

An antigenic determinant (AD) is a portion of an antigen molecule known as an epitope that is recognized by the human immune system, specifically by antibodies or T and B cells. Recognition of epitopes is considered important in EBPV design to contain pandemics, epidemics, and endemics due to the outbreak of infectious diseases. To design an effective and viable EBPV against different strains of a pathogen, it is important to identify the putative T- and B-cell epitopes. Using the wet-lab experimental approach to identify these epitopes is time-consuming and costly because the experimental screening of a vast number of potential epitope candidates is required. Fortunately, various available machine learning (ML)-based prediction methods have reduced the burden related to the epitope mapping process by decreasing the potential epitope candidate list for experimental trials. Moreover, these methods are also cost-effective, scalable, and fast.

machine learning epitopes B-Cell T-Cell antigenic determinant antigen antibody immune-relevant determinants epitope-based peptide vaccine SARS-CoV-2 COVID-19 ensemble model

1. Introduction

An antigenic determinant (AD) is a portion of an antigen molecule known as an epitope that is recognized by the human immune system, specifically by antibodies or T and B cells [1]. Recognition of epitopes is considered important in EBPV design to contain pandemics, epidemics, and endemics due to the outbreak of infectious diseases. The ongoing COVID-19 pandemic due to the SARS-CoV-2 outbreak is the latest among the major pandemics that have occurred in the last decade [1]. COVID-19 can be severe and has caused millions of deaths around the world. It is a respiratory illness and affects people according to the physiology and immune system of the human body. Affected people mostly develop mild to moderate illness and recover without hospitalization [1][2]. While the progress in COVID-19 vaccine design so far is remarkable, successfully vaccinating the worldwide population entails numerous hurdles, from manufacturing to distribution and deployment, and, most crucially, acceptability.

Due to the rate at which SARS-CoV-2 is circulating in the population, thereby causing unprecedented infections, its chances of mutating more and more have increased by now. The variant B.1.617.2, named Delta [3], first identified during a serious wave of COVID-19 infections in India in April and May 2021 [4], was declared a variant of concern (VOC) by the “US Centers for Disease Control and Prevention (CDC)” on 15 June 2021 [5]. Due to its partial resistance to existing vaccines, the infected cases per day increased to over 400,000 [6]. A study conducted by the Chinese Academy of Medical Sciences confirmed that viral loads in Delta infections are approximately 1000 times higher than those in previous SARS-CoV-2 variants [7]. The Mu variant, also known as B.1.621 [3], first identified in January 2021 in Colombia, was declared a “variant of interest” (VOI) on 26 August 2021 by the European Centre for Disease Prevention and Control (ECDC) [8]. On August 30, “the Mu variant was added to the World Health Organization’s (WHO’s) watch list after being found to have a constellation of mutations that indicate potential properties of immune escape” [8]. The most recent variant, B.1.1.529, named Omicron, was first reported to WHO from South Africa on 24 November 2021 [8]. On 26 November 2021, WHO designated the variant B.1.1.529 a VOC on the advice of the Technical Advisory Group on Virus Evolution (TAG-VE) [8]. The hotspot of SARS-CoV-2 mutations is the spike S protein. The spike protein enables the pathogen to infect cells and is the basis for the majority of the vaccines. In [9], it has been reported that “out of 10333 spike protein sequences analyzed, 8155 proteins comprised one or more mutations. A total of 9654 mutations were observed that correspond to 400 distinct mutation sites. The receptor binding domain (RBD) which is involved in the interactions with human angiotensin-converting enzyme-2 (ACE-2) receptor and causes infection leading to the COVID-19 comprised 44 mutations that included residues within 3.2 Å interacting distance from the ACE-2 receptor”.

1.1. Epitopes and Paratopes

An antigen is any substance that causes the immune system to produce antibodies against it. Its molecules are large biological polymers and introduce various molecular attributes that act as interaction sites between antibodies, T_H cells and B cells, and antigen molecules. These interaction sites are called epitopes [10][11][12]. Epitopes are of two types: B-cell epitopes

(BCEs) and T-cell epitopes (TCEs). The fragment of an antigen that is attached to an antibody is called the B-cell epitope [13]. The BCEs are recognized by B cells and comprise a solvent region that is exposed to an antigen. On the other hand, T cells have a receptor on their surface, known as the T-cell receptor (TCR) [13]. When presented on the surfaces of APCs that are linked to MHC molecules, the TCR aids in antigen recognition. TCEs identified by CD8 and CD4 T cells are represented by MHC class I (MHC I) and class II (MHC II) molecules, respectively [13]. **Figure 1** shows an antibody containing two paratopes, indicating that these two paratopes can bind to two pathogens [14][15]. Chemical interactions between epitopes and paratopes that promote antigen–antibody binding are non-covalent [16][17][18].

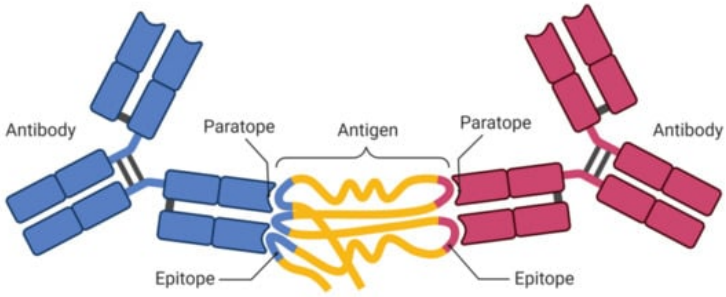

Figure 1. Antigen recognition by antibodies.

1.2. Need for T- and B-Cell Epitope Prediction

The identification of epitopes is of great importance for many reasons, including EBPV design, antibody production, and immunodiagnostic tests. They also play a crucial role in activating the human immune system. Among the reasons listed, EBPV design is important for researchers, biologists, and scientists because there are numerous drawbacks to using whole-organism vaccines, particularly in immunocompromised patients [19][20]. EBPVs can be utilized to overcome the issues associated with heterogeneous and multicomponent vaccines and are seen as an alternative to traditional vaccines. They can act as powerful alternatives to conventional vaccines due to their low production cost, having less reactogenic and allergenic responses. A well-trained ML model of experimentally determined epitopes and non-epitopes can identify potential epitopes as vaccine candidates quickly and efficiently and can reduce the burden related to the epitope mapping process by decreasing the potential epitope candidate list for experimental trials. Using the wet-lab experimental approach to identify these epitopes is time-consuming and costly because the experimental screening of a vast number of potential epitope candidates is required. However, epitope prediction methods based on ML can prove to be cost-effective, scalable, and fast. The most recent vaccine technology is based on RNA vaccines, which have the distinct advantage of being simple to design and manufacture. Epitopes are critical, but often overlooked, for boosting the effectiveness of RNA vaccines. Although RNA vaccines can encode any gene of interest, even the most recent designs commonly encode sequences of original genes from the natural virus. Epitope prediction can be useful in assisting RNA vaccine design by guiding the sequence design and vaccine structure. RNA (mRNA) vaccines, on the other hand, can benefit from epitope-based design approaches, in which both B-cell and T-cell epitopes can be used for vaccine design. The epitope properties determine whether or not the RNA vaccine will elicit an immune response and which types of responses will be elicited.

2. ML-Based Studies for the Prediction of T- and B-Cell Epitopes

ML is concerned with the automated learning of machines that is not explicitly programmed. It focuses on making data-driven predictions and has several applications in bioinformatics [21]. Bioinformatics deals with applying computational techniques to derive knowledge from biological data. It covers the collection, retrieval, storage, manipulation, and data modeling for analysis or prediction using various algorithms and software [21]. Earlier, one had to explicitly program bioinformatics algorithms, which was an extremely laborious task for predicting protein structures [21]. However, with the advent of ML algorithms, such problems have become much easier to solve. In recent years, the exponential growth of T- and B-cell epitope data has become the primary motivation for researchers to develop ML-based methods for the prediction of ADs or IRDs, i.e., B- and T-cell epitopes. ML applied to experimentally determined peptide sequence data of pathogens (virus, bacteria, etc.) opens up new frontiers for areas such as EBPV design, antibody production, and immunodiagnostic tests. The ML-based in silico approach has emerged as a promising field for epitope prediction [22]. Accordingly, various ML-based studies and methods exist that utilize the physicochemical properties of amino acids as features or descriptors for the prediction of epitopes (**Table 1**).

Table 1. Existing studies for T- and B-cell epitope prediction.

Study Conducted	Methodology Adopted	Strengths/Limitations
T. Liu et al. [23]	A feedforward deep neural network-based ensemble of 11 classifiers was created to predict BCEs. IEDB was used to obtain the BCE peptide dataset. On the test set, the model was evaluated using the AUROC metric.	Model reports peptide as an epitope if classified by all 11 classifiers. It would provide the best results if simple majority voting was used for classification.
Fatoba, A. J. et al. [24]	In [24], potential epitope-based vaccine candidates were explored. After retrieving 600 genome sequences of SARS-CoV-2 from the ViPR repository, CD8+ and CD4+ epitopes and B-cell (linear) epitopes were generated and screened for immunogenicity, antigenicity, and non-allergenicity.	The results of [25] reported 19 candidate T-cell epitopes (CD8+), which were found to overlap strongly with 8 B-cell epitopes. The results provide the basis for an experimental design for a suitable peptide vaccine against SARS-CoV-2.
R. Moody et al. [26]	Authors used IEDB prediction tools for predicting B-cell epitopes and those with high scores in terms of prediction were selected as candidate epitopes. The epitopes were then matched to human proteins using NCBI Blast technology.	The findings showed eleven (11) novel B-cell epitopes in the host that were capable of explaining key elements of COVID-19 extrapulmonary disease that previous research had not been able to explain.
Jespersen MC et al. [27]	The authors employed feedforward neural networks (FFNN) with two hidden layers, each with 25 neurons, an activation function (sigmoid) at all neurons, and an ADAM as an optimizing function to predict antibody-specific epitopes (B cell) or epitope targets of provided cognate antibodies. The dataset was obtained from the IEDB database. PCA was used for dimensionality reduction before the model was trained.	It was shown that a simple set of attributes retrieved from the cognate antibody boosted the rate of accuracy in predicting individual epitopes. Furthermore, sophisticated features such as Zernike Moments can improve the model's predictive potential. When compared to DiscoTope 2.0, this model performs better in finding patches overlapping with an actual patch of an epitope in cross-validation and on an independent dataset.
Ling-yun Liu et al. [28]	The authors used PCA and RNN networks. They converted the physicochemical properties into digital vectors, intending to have high-dimensional feature space, and later PCA was applied to process them. The output from PCA was used as an input to the RNN for predicting epitopes.	Prediction results obtained by this process demonstrated that PCA reduced dimensions, but at the same time, original features of the main component were retained, and the rate of prediction was also improved.
Bin Cheng et al. [29]	Authors introduced a novel scale to measure feature importance, called the relevance of amino acid pair (RAAP). RAAP was calculated by decomposing the sequences of amino acids based on their physicochemical properties.	The successful prediction rate was drastically improved here by using LSTM. It does not suffer from gradient instability and is good enough for textual classification sequences. Fivefold cross-validation was used to test and validate the models.
Balachandran Manavalan et al. [30]	Here, a non-redundant dataset was constructed containing 5500 BCEs experimentally validated, and 6893 non-B-cell epitopes were retrieved from IEDB. Then, an ensemble model to predict B-cell epitopes based on ERT (extremely randomized tree) and a classifier called GB (gradient boosting) was developed. The model works based on the physicochemical properties, AA composition, and combination of dipeptides and PCP as the input features.	After performing cross-validation on a benchmark dataset, it was shown that this model performed far better than the individual classifiers such as ERT and GB, with an MCC (Matthews correlation coefficient) of 0.454.
Yuh-Jyh Hu et al. [31]	A cost-sensitive strategy based on bagging MDT was suggested, which integrates two ensemble-based learning algorithms. Without employing the prediction of a pre-trained single predictor, it makes it independent of multiple prediction tools. It can also learn a meta-classification architecture with varied data, without being constrained by a particular hierarchy.	It was demonstrated that the performance of prediction is superior as compared to a single epitope predictor. However, epitope prediction based on meta-learning is purely dependent upon the predictive strength of various other pre-trained linear and conformational epitope prediction tools, which cannot be retained directly by users. Hence, this limits the flexibility and applicability of these meta-classifiers.
Jing Ren et al. [32]	The authors proposed a novel staged heterogeneity-based learning model. The model learns both heterogeneity and characteristics of data in a phased manner to identify residue of antigens of conformational B-cell type epitopes that are heterogeneous, purely based on sequences of antigens. In the first stage, the model is made to learn the generic epitope pattern with propensities, and in the second stage, the same model is made to	It was demonstrated that if heterogeneity was learned well, the transferability of the model improved remarkably in handling new data.It was tested and validated on two different datasets: one with epitopes determined experimentally and another with computationally defined. It showed outstanding performance that was around twice that of existing predictors, including CBTOPE.

Study Conducted	Methodology Adopted	Strengths/Limitations
	learn the complementarity of the propensities used in the first stage, which is heterogeneous but this time on a small dataset of experimentally verified epitopes.	
Georgios A. et al. [33]	A novel method, “SEPIa”, has been proposed here to predict B-cell epitopes from protein sequences and is sufficiently faster, and it can also be applied to large-scale datasets. The model is the combination of two classifiers, random forest and naïve Bayes algorithm.	The average prediction accuracy of SEPIa is limited. The AUC score is 0.65 in both 10-fold cross-validation and on the independent test dataset, which is higher than other approaches tested on the same test dataset.
Gene Sher et al. [25]	Authors proposed a novel, analytically trained DREEP (Deep Ridge Regressed Epitope Predictor) based on string kernels using a deep neural network tailored to predict continuous epitopes.	The model was tested with input as long sequences of proteins from datasets such as AntiJen, Pellequer, and HIV. The results were compared with epitope predictors such as DMNLBE, LBtope, etc. Using the area under the curve (AUC) metric, the model achieved performance improvements over SARS by 13.7%, HIV by 8.9%, and Pellequer by 1.5%.
Wen Zhang et al. [34]	Authors attempted to differentiate immunogenic epitopes from non-immunogenic epitopes based purely on their primary structure. To effectively utilize various features, an ensemble method based on a genetic algorithm was proposed.	The model was tested on two benchmark datasets: IMMA2, PAAQD. The model was compared with methods such as POPI, PAAQD, and POPISK, which are considered state-of-the-art in nature. The model performed better, with an AUC score on IMMA2 of 0.846 and 0.829 on PAAQD.
Wei Zheng et al. [35]	The authors used ensemble learning to improve the prediction of BCEs. Their ensemble method combined twelve SVMs. To handle imbalanced datasets, resampling and AdaBoost methods were used.	The proposed ensemble model achieved an AUC score of 0.642–0.672 on the training dataset with five-fold cross-validation and an AUC score of 0.579–0.604 on the test dataset.
Jian Zhang et al. [36]	To predict antigenic determinants, the authors devised a cost-sensitive ensemble approach, and a spatial clustering-based algorithm was used to identify probable epitopes.	The model performed admirably in terms of prediction. AUC scores of 0.721 and 0.703 were obtained using leave-one-out cross-validation (LOOCV) on two benchmark datasets: bound and unbound.
Kavitha K V et al. [37]	PCA was used to reduce dimensions and to filter out the essential features; for prediction purposes, a random forest algorithm was used.	Experimental results showed that the random forest-based classifier had an improved prediction accuracy rate as compared to BCPred, AAP, etc.
Wen Zhang et al. [38]	The authors used sequence-derived features and developed an ensemble model based on random forest to predict epitopes accurately.	The model was evaluated using the leave-one-out cross-validation procedure, and an AUC score of 0.687 and 0.651 on bound and unbound datasets was obtained.
Ping Chen et al. [39]	Authors reviewed various prediction models for epitopes, such as models based on SVM, neural network, random forest, etc., to defend computational approaches in the prediction of epitopes as in silico methods require a lot of effort and time.	Apart from defending the computational approaches, it was also concluded that there is a limitation to current models as it is impossible to devise an exact model without having complete knowledge of the immune system, and current models are simply best at approximation.
Claus Lundegaard et al. [40]	Here, an artificial neural network was used. The standard feedforward neural network with backpropagation was employed to predict epitopes. The dataset was retrieved from the SYFPEITHI database.	The model efficiently and accurately predicts MHC class I type peptides and outperforms the existing methods.

References

1. Immunology Guidebook|ScienceDirect. Available online: https://www.sciencedirect.com/book/9780121983826/immunology-guidebook (accessed on 25 September 2021).
2. COVID Live Update: 270,426,226 Cases and 5,321,864 Deaths from the Coronavirus—Worldometer. Available online: https://www.worldometers.info/coronavirus/ (accessed on 10 December 2021).
3. Centers for Disease Control and Prevention (CDC). SARS-CoV-2 Variant Classifications and Definitions. Available online: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html (accessed on 7

3. Tools for T- and B-Cell Epitope Prediction

3.1. Tools for T-Cell Epitope Prediction

The primary basis for T-cell epitope prediction is peptide–MHC binding prediction. A number of tools and methodologies for predicting T-cell epitopes have been developed and are freely available online. We hereby provide a categorized review of

these tools based on the methods they use for prediction. The methods used are structure-based (SB), motif matrix (MM), sequence motif (SM), quantitative affinity matrix (QAM), artificial neural network (ANN), support vector machine (SVM), the quantitative structure–activity relationship model (QSAR), and combined (using QAM and ANN). For each tool, we have mentioned the URL and which class of MHC binding prediction is supported (class I or II or both). These tools only assess a peptide's binding capability. It is still difficult for these methods to estimate deterministically whether a given peptide is an epitope or not. CTLpred [41], one of the servers, works in this category; however, it is limited to peptides with a length of up to 9 mers only. However, the benefit of using ML algorithms for epitope prediction for the methods illustrated is that they address two distinct problems: the differentiation of MHC binders from non-binders and the prediction of the binding affinity of a peptide to MHC molecules. The first issue has been addressed by using classifiers such as ANNs, SVMs, decision trees (DT), and Hidden Markov models (HMMs). All of these classifiers have been trained on data containing peptides that have or do not have binding affinity to the MHC molecule. ML classifiers were developed on a dataset of peptides with an affinity to the MHC molecule to solve the second problem, i.e., binding affinity prediction. Here, SVMs and ANNs have been used to first predict affinity for MHC I and then for MHC II molecules. However, when using the MHC binding model to predict T-cell epitopes, difficulty arises due to MHC polymorphism [42]. To address this, pan-MHC-specific models were created by training ANNs on data containing MHC residues [43]. Furthermore, it has been established that combining different approaches and providing a consensus prediction improves peptide–MHC prediction [44].

3.2. Tools for B-Cell Epitope Prediction

The goal of predicting BCEs is to make it easier to identify a BCE for antigen replacement in an antibody production process. BCEs are classified into two types: conformational and linear. As shown in **Figure 2**, linear BCEs are composed of consecutive peptides and residues. Conformational ones, on the other hand, are formed of patches of solvent-exposed atoms from non-sequential residues. As a result, conformational and linear BCEs are also known as discontinuous and continuous BCEs.

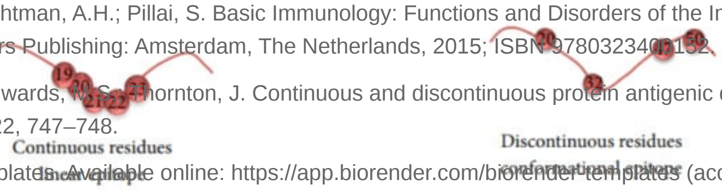

Figure 2. Linear and conformational B-cell epitopes.

Regarding Linear BCEs, although being in the minority, their prediction has received more attention. A few existing bioinformatics-based tools, such as PEOPLE-EP [45] and PREDITOP [46] for BCE prediction, make use of propensity scales. The tool PREDITOP [46] is based on a multi-parametric method using the accessibility, hydrophilicity, and flexibility properties of amino acids. On the other hand, PEOPLE-EP [45] is also based on these parameters but includes the assessment of β-turns. However, in [47], by Blythe and Flower, it has been shown that the amino acid propensity scale is unreliable for predicting epitope location.

The unreliability issue in predicting BCEs due to amino acid scales has been mitigated using ML algorithms. To differentiate BCEs from non-epitopes, ML algorithms have been trained on feature vectors extracted from BCEs. A few methods based on ML include ABCpred [48], BCPREDS [49], LBtope [50], SVMTrip [51], and BepiPred [52]. It has been reported that methods based on ML techniques outperform the techniques based on amino acid scales [53]. Conformational BCEs constitute the majority portion; however, their prediction is lagging behind that of linear types due to two main reasons. Firstly, their prediction necessitates knowledge of the 3D protein structure. Only a limited percentage of proteins have 3D information [54]. Secondly, extracting conformational epitopes for specific antibody synthesis from a protein context is a difficult process that requires the use of appropriate scaffolds for epitope grafting. Therefore, their prediction thus far is of less relevance for EBPV design.

4. Predicting SARS-CoV-2 Epitopes

Coronaviruses belong to the family *Coronaviridae*, the enveloped viruses having a large single-stranded RNA genome whose length ranges from 26 to 32 kilobases [55]. In [56], by Lin and colleagues, it has been found that, among 26 viral proteins of SARS-CoV-2, a few proteins on its surface, such as the spike protein (S), are more variable, while others are more conserved and internal, such as the nucleocapsid protein (N). It has been found that the spike protein (S) is responsible for activating cytotoxic $CD8^+$ T cells and hence is considered an ideal vaccine target.

August 2021).

4. WHO Director-General's opening remarks at the 8th meeting of the IHR Emergency Committee on COVID-19—14 July 2021. Available online: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-8th-meeting-of-the-ihr-emergency-committee-on-covid-19-14-july-2021 (accessed on 10 December 2021).
5. Coronavirus Disease 2019 (COVID-19)|CDC. Available online: https://www.cdc.gov/coronavirus/2019-ncov/index.html (accessed on 7 August 2021).
6. Callaway, E. Delta coronavirus variant: Scientists brace for impact. Nature 2021, 595, 17–18.
7. Li, B.; Deng, A.; Li, K.; Hu, Y.; Li, Z.; Xiong, Q.; Liu, Z.; Guo, Q.; Zou, L.; Zhang, H.; et al. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. MedRxiv 2021.
8. COVID-19: What Is the Mu Variant? United Nations Western Europe. Available online: https://unric.org/en/covid-19-what-is-the-mu-variant/ (accessed on 25 September 2021).
9. Guruprasad, L. Human SARS CoV-2 spike protein mutations. Proteins Struct. Funct. Bioinform. 2021, 89, 569–576.
10. Marshall, J.S.; Warrington, R.; Watson, W.; Kim, H.L. An introduction to immunology and immunopathology. Allergy Asthma Clin. Immunol. 2018, 14, 49.
11. Abbas, A.K.; Lichtman, A.H.; Pillai, S. Cellular and Molecular Immunology; Saunders Elsevier: Philadelphia, PA, USA, 2007; p. 566.
12. Doan, T.; Melvold, R.; Viselli, S. Lippincott's Illustrated Reviews, Immunology, 2nd ed.; Wolter Kluwer: Alphen aan den Rijn, The Netherlands, 2012; ISBN 9781451109375.
13. Abbas, A.K.; Lichtman, A.H.; Pillai, S. Basic Immunology: Functions and Disorders of the Immune System; Elsevier Slanders Publishing: Amsterdam, The Netherlands, 2015; ISBN 9780323400152.
14. Barlow, D.J.; Edwards, M.S.; Thornton, J. Continuous and discontinuous protein antigenic determinants. Nature 1986, 322, 747–748.
15. BioRender Templates. Available online: https://app.biorender.com/biorender-templates (accessed on 26 September 2021).
16. Mix, E.; Goertsches, R.; Zettl, U.K. Immunoglobulins—Basic considerations. J. Neurol. 2006, 253 (Suppl. 5), V9–V17, Erratum in J. Neurol. 2008, 255, 308.
17. A Compact Vocabulary of Paratope-Epitope Interactions Enables Predictability of Antibody-Antigen Binding|Elsevier Enhanced Reader. Available online: https://reader.elsevier.com/reader/sd/pii/S2211124721001704?token=7448F25258D74599D0802A9AFA03C34793008C315DF289599AE40FDBA0AF1A482C4B92C75ADC47372988E9FABB4A34 west-1&originCreation=20210904091233 (accessed on 4 September 2021).
18. Ravetch, J.V.; Bolland, S. IgG Fc Receptors. Annu. Rev. Immunol. 2001, 19, 275–290.
19. Janeway, C.A., Jr.; Travers, P.; Walport, M.; Shlomchik, M.J. Immunobiology: The Immune System in Health and Disease, 5th ed.; Garland Science: New York, NY, USA, 2001. Available online: https://www.ncbi.nlm.nih.gov/books/NBK10757/ (accessed on 12 October 2021).
20. Al Qaraghuli, M.M.; Kubiak-Ossowska, K.; Ferro, V.A.; Mulheran, P.A. Antibody-protein binding and conformational changes: Identifying allosteric signalling pathways to engineer a better effector response. Sci. Rep. 2020, 10, 13696.
21. Introduction to Antigen-Antibody Reactions. Available online: https://microbenotes.com/introduction-to-antigen-antibody-reactions/ (accessed on 4 September 2021).
22. An Introduction to Antibodies: Antibody-Antigen Interaction. Available online: https://www.sigmaaldrich.com/IN/en/technical-documents/technical-article/protein-biology/elisa/antibody-antigen-interaction (accessed on 4 September 2021).
23. Roper, R.L.; Rehm, K.E. SARS vaccines: Where are we? Expert Rev. Vaccines 2009, 8, 887–898.
24. Shang, W.; Yang, Y.; Rao, Y.; Rao, X. The outbreak of SARS-CoV-2 pneumonia calls for viral vaccines. NPJ Vaccines 2020, 5, 18.

The infection caused by SARS-CoV-2 elicits both adaptive and innate arms of immunity [57]. In general, antigen-presenting cells recognize viruses, leading to T-cell activation; CD4+ T cells mainly differentiate into effector cells, which produce cytokines and chemokines; cytotoxic CD8+ T cells, on the other hand, are key players in the immune response to viral infection, as they participate directly in viral clearance [58]. It has been demonstrated that T cells, apart from targeting the structural proteins of coronaviruses, are also responsible for lung immunopathological damage due to SARS-CoV and MERS-CoV [59][60]. Thus, in the case of SARS-CoV-2, the major focus has been on identifying viral T-cell epitopes presented on human leukocyte antigens (HLA) [61][62] (**Table 2**).

Table 2. Existing ML methods used in SARS-CoV-2 epitope prediction.

Sr. No.	Method Name	Usage
01	NetMHC [63]	To predict HLA I class or CD8+ SARS-CoV-2 T-cell epitopes
02	NetMHCpan [64]	
03	NetCTLpan_1.1 [65]	
04	NetMHC_4.0 [66]	
05	HLAthena [67]	
06	MHCflurry [68]	
07	NetHMCII_2.3 [69]	To predict HLA II class or CD4+ SARS-CoV-2 T-cell epitopes
08	NetMHCIIpan_3.0 [70]	
09	NetMHCIIpan_4.0 [71]	
10	NeonMHC2 [72]	
11	MARIA [73]	

A few techniques listed in **Table 2** have "pan" as a suffix, which indicates an ability to predict the binding of HLA peptides for a huge collection of the alleles inside a particular HLA type, including those not present in the training dataset [72]. A few studies have also used algorithms specific to HLA-I, namely NetChop [74] and NetCTL [75], where extra- and intracellular variables responsible for the presentation of HLA antigens were integrated to improve the prediction accuracy of the binding of peptide HLA. The methods NetCTL-1.2 [75] and NetChop [74] have also been utilized in a few studies, where extra- and intracellular variables have been integrated, which are responsible for presenting HLA antigens. It is essential to mention here that almost all modern T-cell epitope prediction systems use ANNs. A few early ones (such as RANKPEP [76] and CTLPred [41]) used a different ML approach, support vector machines (SVM). The spike proteins in the original virus bind to the ACE2 receptor on human cells. It has been reported in [77] that the D614G mutation alters the genetic code of the spike protein of SARS-CoV-2, where a change in a single amino acid takes place, and most of the COVID-19 vaccines are based on this spike protein. Due to this mutation, the virus spreads faster and the spikes become more stable than those in the original virus. As a result, more functional spikes are available to bind to ACE2 receptors, making the virus more infectious. Crooke et al. [78] developed a computational model using various open-source algorithms and web-based tools to analyze the SARS-CoV-2 proteome so as to identify antigenic and putative T-cell and B-cell epitopes as potential vaccine targets. After using a set of stringent selection criteria to filter out the peptide epitopes, the study discovered 41 T-cell epitopes (5 HLA class I, 36 HLA class II) and six B-cell epitopes that have the potential to serve as primary targets for epitope-based peptide vaccine development against SARS-CoV-2.

5. Future Research Directions in T- and B-Cell Epitope Prediction

By now, it is clear that the key to designing an EBPV is the identification of BCEs and TCEs [79][80]. Several studies have been performed to predict BCEs and TCEs, as illustrated in **Table 1**. The methods used to predict SARS-CoV-2 epitopes are listed in **Table 2**; again, these predict only the peptide-binding capacity. This is a limitation with these methods; instead of predicting the binding capability of a peptide, predicting epitopes deterministically is desired. Because viruses continue to mutate, as with SARS-CoV-2, existing vaccines may prove to be somewhat less effective against new variants. Either the vaccine's composition has to be changed or a new vaccine needs to be developed to protect against these variants [81]. Time being the critical factor, EBPVs can be a great solution. Based on the research conducted, EBPVs are highly recommended vaccines and should be considered in the quest for the rapid development of protective vaccines. Below, we mention the future research directions for epitope prediction as predicting epitopes is a sensitive task and needs due attention in order to improve it.

25. Manavalan, B.; Govindaraj, R.G.; Shin, T.H.; Kim, M.O.; Lee, G. iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction. Front. Immunol. 2018, 9, 1695.

26. Larrañaga, P.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; Inza, I.; Lozano, J.A.; Armañanzas, R.; Santafé, G.; Pérez, A.; et al. Machine learning in bioinformatics. Brief. Bioinform. 2006, 7, 86–112.

27. Cunha-Neto, E.; Rosa, D.; Harris, P.; Olson, T.; Morrow, A.; Ciotlos, S.; Herst, C.V.; Rubsamen, R.M. An Approach for a Synthetic CTL Vaccine Design against Zika Flavivirus Using Class I and Class II Epitopes Identified by Computer Modeling. Front. Immunol. 2017, 8, 640.

28. Liu, T.; Shi, K.; Li, W. Deep learning methods improve linear B-cell epitope prediction. BioData Min. 2020, 13, 1.

29. [illegible] ...cs ...s based

30. [illegible] ...dicted B ...m. 2021,

31. [illegible] ...e ...19, 10,

32. [illegible] ...work. In ...Biology

33. [illegible] ...hort ...s and Computational Biology, Chengdu, China, 12–14 March 2018; pp. 55–59.

34. Hu, Y.-J.; You, S.-N.; Ko, C.-L. Computational Ensemble Approach for Immune System Study: Conformational B-cell Epitope Prediction. Eur. J. Biomed. Inform. 2017, 14, 4–15.

35. Ren, J.; Song, J.; Ellis, J.; Li, J. Staged heterogeneity learning to identify conformational B-cell epitopes from antigen sequences. BMC Genom. 2017, 18, 113.

36. Georgios, A.; Rooman, D.M. SEPIa, a knowledge-driven algorithm for predicting conformational B-cell epitopes from the amino acid sequence. BMC Bioinform. 2017, 18, 95.

37. Shen, G.; Zhi, D.; Zhang, S. DRREP: Deep ridge regressed epitope predictor. BMC Genom. 2017, 18, 676.

38. Zhang, W.; Niu, Y.; Zou, H.; Luo, L.; Liu, Q.; Wu, W. Accurate Prediction of Immunogenic T-Cell Epitopes from Epitope Sequences Using the Genetic Algorithm-Based Ensemble Learning. PLoS ONE 2015, 10, e0128194.

39. Zheng, W.; Zhang, C.; Hanlon, M.; Ruan, J.; Gao, J. An ensemble method for prediction of conformational B-cell epitopes from antigen sequences. Comput. Biol. Chem. 2014, 49, 51–58.

40. Zhang, J.; Zhao, X.; Sun, P.; Gao, B.; Ma, Z. Conformational B-Cell Epitopes Prediction from Sequences Using Cost-Sensitive Ensemble Classifiers and Spatial Clustering. BioMed Res. Int. 2014, 2014, 689219.

41. Larsen, M.V.; Lundegaard, C.; Lamberth, K.; Buus, S.; Brunak, S.; Lund, O.; Nielsen, M. An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur. J. Immunol. 2005, 35, 2295–2303.

42. Reche, P.; Reinherz, E.L. Sequence Variability Analysis of Human Class I and Class II MHC Molecules: Functional and Structural Correlates of Amino Acid Polymorphisms. J. Mol. Biol. 2003, 331, 623–641.

43. Zhang, L.; Udaka, K.; Mamitsuka, H.; Zhu, S. Toward more accurate pan-specific MHC-peptide binding prediction: A review of current methods and tools. Brief. Bioinform. 2012, 13, 350–364.

44. Wang, P.; Sidney, J.; Dow, C.; Mothé, B.; Sette, A.; Peters, B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput. Biol. 2008, 4, e1000048.

45. Alix, A.J.P. Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine 1999, 18, 311–314.
46. Pellequer, J.-L.; Westhof, E. PREDITOP: A program for antigenicity prediction. J. Mol. Graph. 1993, 11, 204–210.
47. Blythe, M.J.; Flower, D.R. Benchmarking B cell epitope prediction: Underperformance of existing methods. Protein Sci. 2005, 14, 246–248.
48. Saha, S.; Raghava, G.P.S. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins Struct. Funct. Bioinform. 2006, 65, 40–48.
49. El-Manzalawy, Y.; Dobbs, D.; Honavar, V. Predicting linear B-cell epitopes using string kernels. J. Mol. Recognit. 2008, 21, 243–255.
50. Singh, H.; Ansari, H.R.; Raghava, G.P.S. Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence. PLoS ONE 2013, 8, e62216.
51. Yao, B.; Zhang, L.; Liang, S.; Zhang, C. SVMTriP: A Method to Predict Antigenic Epitopes Using Support Vector Machine to Integrate Tri-Peptide Similarity and Propensity. PLoS ONE 2012, 7, e45152.
52. Jespersen, M.C.; Peters, B.; Nielsen, M.; Marcatili, P. BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017, 45, W24–W29.
53. Greenbaum, J.A.; Andersen, P.H.; Blythe, M.; Bui, H.-H.; Cachau, R.E.; Crowe, J.; Davies, M.; Kolaskar, A.S.; Lund, O.; Morrison, S.; et al. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. J. Mol. Recognit. 2007, 20, 75–82.
54. Levitt, M. Nature of the protein universe. Proc. Natl. Acad. Sci. USA 2009, 106, 11079–11084.
55. Su, S.; Wong, G.; Shi, W.; Liu, J.; Lai, A.C.K.; Zhou, J.; Liu, W.; Bi, Y.; Gao, G.F. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol. 2016, 24, 490–502.
56. Lineburg, K.E.; Grant, E.J.; Swaminathan, S.; Chatzileontiadou, D.S.; Szeto, C.; Sloane, H.; Panikkar, A.; Raju, J.; Crooks, P.; Rehan, S.; et al. CD8+ T cells specific for an immunodominant SARS CoV-2 nucleocapsid epitope cross-react with selective seasonal coronaviruses. Immunity 2021, 54, 1055–1065.e5.
57. Zhang, X.; Tan, Y.; Ling, Y.; Lu, G.; Liu, F.; Yi, Z.; Jia, X.; Wu, M.; Shi, B.; Xu, S.; et al. Viral and host factors related to the clinical outcome of COVID-19. Nature 2020, 583, 437–440.
58. Schmidt, M.E.; Varga, S.M. The CD8 T Cell Response to Respiratory Virus Infections. Front. Immunol. 2018, 9, 678.
59. Ng, O.-W.; Chia, A.; Tan, A.T.; Jadi, R.S.; Leong, H.N.; Bertoletti, A.; Tan, Y.-J. Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection. Vaccine 2016, 34, 2008–2014.
60. Channappanavar, R.; Perlman, S. Pathogenic human coronavirus infections: Causes and consequences of cytokine storm and immunopathology. Semin. Immunopathol. 2017, 39, 529–539.
61. Huber, S.E.; Beek, J.E.; de Jonge, J.; Luytjes, W.; Baarle, D.E. T Cell Responses to Viral Infections—Opportunities for Peptide Vaccination. Front. Immunol. 2014, 5, 171.
62. Seder, R.A.; Darrah, P.A.; Roederer, M. T-cell quality in memory and protection: Implications for vaccine design. Nat. Rev. Immunol. 2008, 8, 247–258.
63. Zhang, G.L.; DeLuca, D.S.; Keskin, D.B.; Chitkushev, L.; Zlateva, T.; Lund, O.; Reinherz, E.L.; Brusic, V. MULTIPRED2: A computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles. J. Immunol. Methods 2011, 374, 53–61.
64. Nielsen, M.; Lundegaard, C.; Worning, P.; Lauemoller, S.L.; Lamberth, K.; Buus, S.; Brunak, S.; Lund, O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003, 12, 1007–1017.
65. Stranzl, T.; Larsen, M.V.; Lundegaard, C.; Nielsen, M. NetCTLpan: Pan-specific MHC class I pathway epitope predictions. Immunogenetics 2010, 62, 357–368.
66. Paul, S.; Croft, N.P.; Purcell, A.W.; Tscharke, D.C.; Sette, A.; Nielsen, M.; Peters, B. Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. PLoS Comput. Biol. 2020, 16, e1007757.

6. Conclusions

Prediction of T- and B-cell epitopes can play a game-changing role in the EBPV design process, as well as in disease diagnosis. In this study, a review of various existing studies for epitope prediction has been provided. Moreover, a review has been provided for the state-of-the-art ML-based tools that are available online and free to use for researchers working in

67. Abelin, J.G.; Keskin, D.B.; Sarkizova, S.; Hartigan, C.R.; Zhang, W.; Sidney, J.; Stevens, J.; Lane, W.; Zhang, G.L.; Eisenhaure, T.M.; et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity 2017, 46, 315–326.

68. O'Donnell, T.J.; Rubinsteyn, A.; Bonsack, M.; Riemer, A.B.; Laserson, U.; Hammerbacher, J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst. 2018, 7, 129–132.e4.

69. Jensen, K.K.; Andreatta, M.; Marcatili, P.; Buus, S.; Greenbaum, J.A.; Yan, Z.; Sette, A.; Peters, B.; Nielsen, M. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology 2018, 154, 394–406.

70. Karosiene, E.; Rasmussen, M.; Blicher, T.; Lund, O.; Buus, S.; Nielsen, M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics 2013, 65, 711–724.

71. Reynisson, B.; Alvarez, B.; Paul, S.; Peters, B.; Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020, 48, W449–W454.

72. Abelin, J.G.; Harjanto, D.; Malloy, M.; Suri, P.; Colson, T.; Goulding, S.P.; Creech, A.L.; Serrano, L.R.; Nasar, G.; Nasrullah, Y.; et al. Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction. Immunity 2019, 51, 766–779.e17.

73. Chen, B.; Khodadoust, M.S.; Olsson, N.; Wagar, L.; Fast, E.; Liu, C.L.; Muftuoglu, Y.; Sworder, B.; Diehn, M.; Levy, R.; et al. Predicting HLA class II antigen presentation through integrated deep learning. Nat. Biotechnol. 2019, 37, 1332–1343.

74. Nielsen, M.; Lundegaard, C.; Lund, O.; Keşmir, C. The role of the proteasome in generating cytotoxic T-cell epitopes: Insights obtained from improved predictions of proteasomal cleavage. Immunogenetics 2005, 57, 33–41.

75. Larsen, M.V.; Lundegaard, C.; Lamberth, K.; Buus, S.; Lund, O.; Nielsen, M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinform. 2007, 8, 424.

76. Reche, P.A.; Glutting, J.P.; Reinherz, E.L. Prediction of MHC class I binding peptides using profile motifs. Hum. Immunol. 2002, 63, 701–709.

77. Zhang, J.; Xiao, T.; Cai, Y.; Chen, B. Structure of SARS-CoV-2 spike protein. Curr. Opin. Virol. 2021, 50, 173–182.

78. Crooke, S.; Ovsyannikova, I.G.; Kennedy, R.B.; Poland, G.A. Immunoinformatic identification of B cell and T cell epitopes in the SARS-CoV-2 proteome. Sci. Rep. 2020, 10, 14179.

79. Bukhari, S.N.H.; Jain, A.; Haq, E. A Novel Ensemble Machine Learning Model for Prediction of Zika Virus T-Cell Epitopes. In Lecture Notes on Data Engineering and Communications Technologies; Gupta, D., Polkowski, Z., Khanna, A., Bhattacharyya, S., Castillo, O., Eds.; Springer: Singapore, 2022; Volume 91, pp. 275–292.

80. Zhou, P.; Yang, X.L.; Wang, X.G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.R.; Zhu, Y.; Li, B.; Huang, C.L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020, 579, 270–273.

81. The Effects of Virus Variants on COVID-19 Vaccines. Available online: https://www.who.int/news-room/feature-stories/detail/the-effects-of-virus-variants-on-covid-19-vaccines (accessed on 7 August 2021).

82. Wee, L.J.; Simarmata, D.; Kam, Y.W.; Ng, L.F.; Tong, J.C. SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction. BMC Genom. 2010, 11, S21.

83. Nisar, S.; Bukhari, H.; Dar, M.A. Using Random Forest to Predict T -Cell Epitopes of Dengue Virus. Dengue Virus 2021, 20, 2543–2547.

84. Artificial Neural Network Disadvantages. Retrieved 4 September 2021. Available online: https://www.datascienceexamples.com/artificial-neural-network-disadvantages/ (accessed on 10 December 2021).

85. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 2021, 23, 18.

86. Gagniuc, P.A.; Ionescu-Tirgoviste, C.; Gagniuc, E.; Militaru, M.; Nwabudike, L.C.; Pavaloiu, B.I.; Vasilăţeanu, A.; Goga, N.; Drăgoi, G.; Popescu, I.; et al. Spectral forecast: A general purpose prediction model as an alternative to classical neural networks. Chaos Interdiscip. J. Nonlinear Sci. 2020, 30, 033119.
87. Bukhari, S.N.H.; Jain, A.; Haq, E.; Khder, M.A.; Neware, R.; Bhola, J.; Najafi, M.L. Machine Learning-Based Ensemble Model for Zika Virus T-Cell Epitope Prediction. J. Health Eng. 2021, 2021, 9591670.
88. Huang, F.; Xie, G.; Xiao, R. Research on Ensemble Learning. In Proceedings of the 2009 International Conference on Artificial Intelligence and Computational Intelligence, Shanghai, China, 7–8 November 2009; Volume 3, pp. 249–252.
89. A Gentle Introduction to Ensemble Learning Algorithms. Available online: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms (accessed on 8 September 2021).
90. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. Front. Comput. Sci. 2020, 14, 241–258.
91. Why Use Ensemble Learning? Available online: https://machinelearningmastery.com/why-use-ensemble-learning/ (accessed on 10 July 2021).
92. Osorio, D.; Rondón-Villarreal, P.; Torres, R.T.R. Peptides: A Package for Data Mining of Antimicrobial Peptides. Small 2015, 12, 44–444.
93. Hofmann, H.; Hare, E.; GGobi Foundation. Peptider: Evaluation of Diversity in Nucleotide Libraries. R Package Version 0.2.2. 2015. Available online: https://CRAN.R-project.org/package=peptider (accessed on 10 September 2021).
94. Jain, P.; Chawla, P. A Novel Smart Healthcare System Design for Internet of Health Things. In Proceedings of the 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 24–25 September 2021; pp. 1–8.
95. Bukhari, S.N.H.; Jain, A.; Haq, E.; Mehbodniya, A.; Webber, J. Ensemble Machine Learning Model to Predict SARS-CoV-2 T-Cell Epitopes as Potential Vaccine Targets. Diagnostics 2021, 11, 1990.