Computational and Structural Biotechnology Journal 23 (2024) 2695–2707

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method article

Deep learning of antibody epitopes using positional permutation vectors

Ioannis Vardaxis [a,*], Boris Simovski [a], Irantzu Anzar [a], Richard Stratford [a], Trevor Clancy [a,b,**]

[a] NEC OncoImmunity AS, Oslo Cancer Cluster, Ullernchausseen 64/66, Oslo 0379, Norway
[b] Department of Vaccine Informatics, Institute for Tropical Medicine, Nagasaki University, Japan

ARTICLE INFO

Keywords:
B cell epitope prediction
Antibody epitope prediction
Artificial intelligence
Immune informatics
Epitope discovery

ABSTRACT

Background: The accurate computational prediction of B cell epitopes can vastly reduce the cost and time required for identifying potential epitope candidates for the design of vaccines and immunodiagnostics. However, current computational tools for B cell epitope prediction perform poorly and are not fit-for-purpose, and there remains enormous room for improvement and the need for superior prediction strategies.
Results: Here we propose a novel approach that improves B cell epitope prediction by encoding epitopes as binary positional permutation vectors that represent the position and structural properties of the amino acids within a protein antigen sequence that interact with an antibody. This approach supersedes the traditional method of defining epitopes as scores per amino acid on a protein sequence, where each score reflects each amino acids predicted probability of partaking in a B cell epitope antibody interaction. In addition to defining epitopes as binary positional permutation vectors, the approach also uses the 3D macrostructure features of the unbound protein structures, and in turn uses these features to train another deep learning model on the corresponding antibody-bound protein 3D structures. This enables the algorithm to learn the key structural and physiochemical features of the unbound protein and embedded epitope that initiate the antibody binding process helping to eliminate "induced fit" biases in the training data. We demonstrate that the strategy predicts B cell epitopes with improved accuracy compared to the existing tools. Additionally, we show that this approach reliably identifies the majority of experimentally verified epitopes on the spike protein of SARS-CoV-2 not seen by the model during training and generalizes in a very robust manner on dissimilar data not seen by the model during training.
Conclusions: With the approach described herein, a primary protein sequence and a query positional permutation vector encoding a putative epitope is sufficient to predict B cell epitopes in a reliable manner, potentially advancing the use of computational prediction of B cell epitopes in biomedical research applications.

1. Introduction

B-cell epitopes (BCEs) are clusters of surface accessible amino acids on a protein antigen, recognized by B cell secreted antibodies or B cell receptors (BCR) [1]. The B cell molecular recognition of BCEs by BCRs elicits humoral and cellular immune responses that are key in the fight against pathogenic threats. Knowledge of the precise coordinates of antibody epitope contact points in an antigen upon binding to an antibody can be of tremendous value. Such BCE information can offer crucial guidance in vaccine design [2,3], therapeutic antibody engineering [2], in the streamlining of numerous diagnostic [4,5], and therapeutic applications in molecular medicine [4]. Hence, a variety of BCE mapping strategies have been developed to identify such clusters of BCE coordinates on antigens. Many of these are wet lab-based methods such as X-ray co-crystallography, cryogenic electron microscopy (cryo-EM) and numerous other assays [6]. However, there are innumerable possible BCEs embedded on any given protein antigen sequence, and the experimental approaches to capture these are extremely time consuming, laborious, and expensive, and therefore not amenable to be applied on a large-scale for comprehensive BCE mapping and screening. The ability to accurately predict BCEs computationally would greatly facilitate the comprehensive mapping of complex antigens helping to speed up the development of monoclonal antibody based therapies, vaccines and immune-based diagnostics [6,7]. In recent years, numerous computational prediction algorithms have been developed to attempt the *in-silico* BCE mapping of protein antigens [7,8]. However, the accurate computational prediction of BCEs remains a daunting challenge as most of these prediction tools perform poorly in

* Corresponding author.
** Corresponding author at: NEC OncoImmunity AS, Oslo Cancer Cluster, Ullernchausseen 64/66, Oslo 0379, Norway.
E-mail addresses: Ioannis@oncoimmunity.com (I. Vardaxis), trevor@oncoimmunity.com (T. Clancy).

https://doi.org/10.1016/j.csbj.2024.06.005
Received 2 April 2024; Received in revised form 4 June 2024; Accepted 4 June 2024
Available online 15 June 2024

comparative benchmarking studies [7,9,10] and are consequently not fit for purpose. Discotope3.0, for instance, uses surface exposure to identify potential epitopes, but has shown limited sensitivity in complex test cases [11]. CBtope predicts conformational epitopes based on sequence features alone, which can miss critical structural nuances [12]. Graphbepi incorporates graph-based methods, but often fails to account for the dynamic nature of protein-antibody interactions [13]. These and other limitations underline the need for more sophisticated approaches that can better capture the intricacies of epitope-paratope interactions. Methods that address and improve the current limitations in the

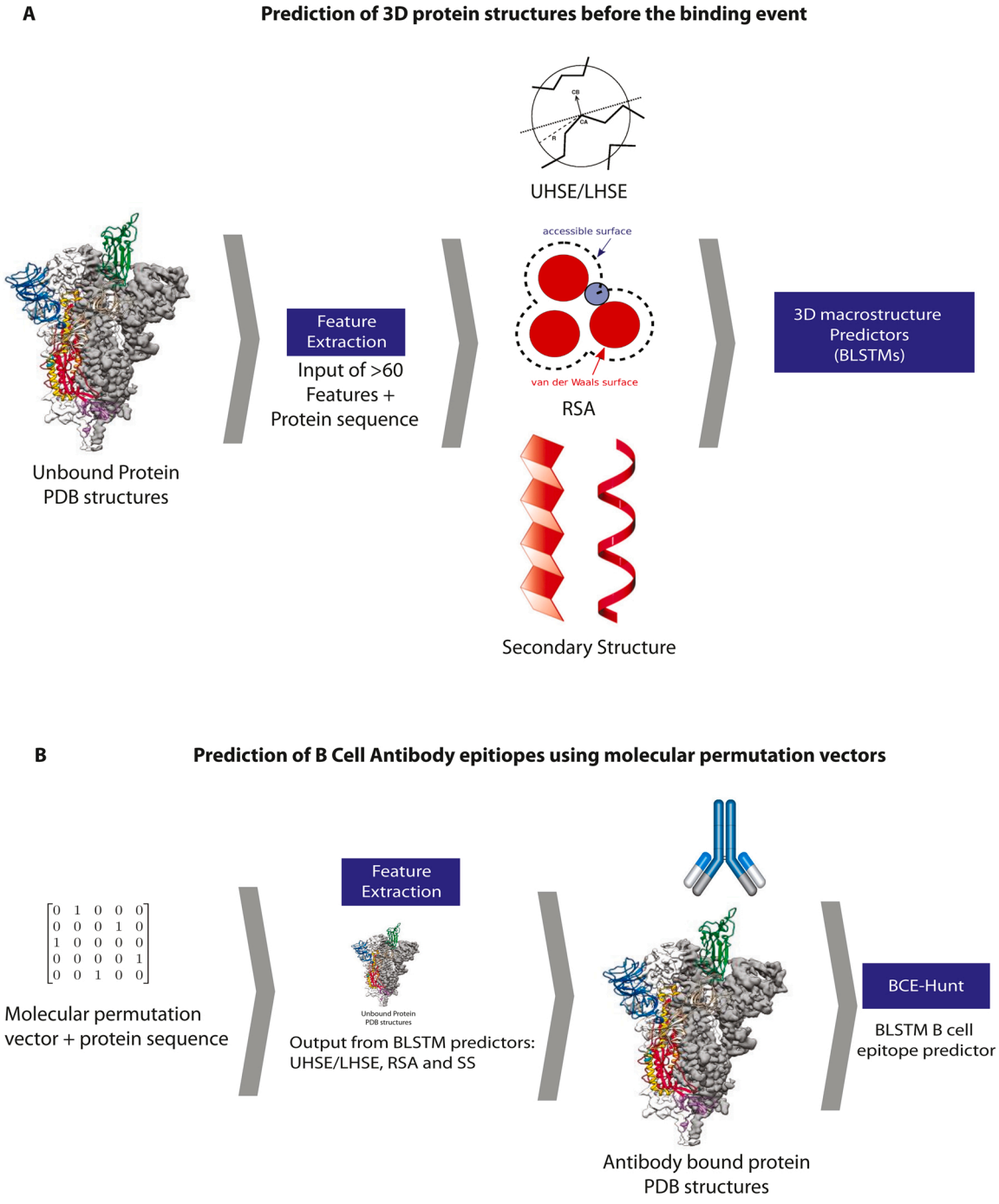


Fig. 1. A top-level outline of the two-step innovative pipeline underlying BCE-Hunt. In step 1, depicted in Fig. 1*A* a BLSTM based deep learning predictor is trained based on all single 3D protein structures that are unbound (i.e., 3D protein structures not bound to any antibody or other ligand) in the PDB database. This predictor learns the 3D macrostructure determinants of 3D protein folding. Namely, RSA, UHSE and LHSE. The features used to train these 3D macrostructure models are derived from the unbound 3D protein structure sequences in addition to numerous other amino acid physiochemical and 3D protein structure sequences (see Table 1). In step 2, depicted in Fig. 1*B* a BLSTM based deep learning predictor is trained based on all of the relevant antibody-bound 3D protein structure complexes in the PDB. The features used in training are in essence the protein sequences of the antibody-bound protein structure and the predicted 3D macrostructure protein features for the unbound protein structure, predicted as described in Fig. 1A. The input at prediction for the main BCE model here is the target antigen protein sequence, and the query positional permutation vector (PPV) that defines the antibody BCE (sequence vector of 1′s and 0′s represents the protein sequence, where 1 represents an amino acid contact point with the antibody and 0 represent not contact with the antibody). Full details are described in methods and the complete ML training pipeline outlined in Supplementary Figure 1.

computational prediction of BCEs have the potential to revolutionize vaccine and immunodiagnostics development [14].

Here we propose a novel approach to BCE prediction that significantly improves the predictive performance compared to the current state-of-the-art tools. Our approach encompasses several novel and distinguishing innovations that more accurately model the underlying biology of BCE/antibody recognition. These include, [1] the development of a set of Bidirectional Long Short-Term Memory (BLSTM) models that predict relevant 3D features of the protein antigen (we term here as "3D macrostructure" features) from its primary sequence, thereby circumventing the need for experimentally derived or computationally predicted 3D protein structures, [2] predicting the aforementioned 3D macrostructure features based on the "unbound" rather than the antibody "bound" protein antigen structure, and [3] defining or encoding the epitopes used for training as binary positional permutation vectors (PPVs) that represent the 3D physical interaction of the BCE with an antibody.

Proteins often undergo conformational changes upon binding to other molecules, a phenomenon known as induced fit [15]. The induced fit model is also applicable in explaining the 3D conformational changes in a protein after antibody binding [16,17], and can alter the 3D structure significantly, potentially leading to inaccurate representations of the native epitope when using antibody bound structures. Therefore, we developed the BLSTM models that predict relevant 3D features of the unbound protein antigen. By leveraging unbound protein structures, our method avoids the inaccuracies associated with induced fit after the binding event, and captures the true native state of the protein prior to antibody binding. This leads to a significant improvement in the prediction accuracy of antibody epitopes.

We propose a novel approach that improves B cell epitope prediction by encoding epitopes as binary positional permutation vectors (PPVs), which represent the position and structural properties of the amino acids within a protein antigen sequence that interact with an antibody. In addition to defining epitopes as binary PPVs, our approach also uses the 3D macrostructure features of the unbound protein structures and in turn uses these features to train another deep learning model on the corresponding antibody-bound protein 3D structures, helping to correct for "induced fit" biases and inaccuracies in the training data. We demonstrate that this strategy, leveraging these novel approaches, predicts B cell epitopes with improved accuracy compared to existing tools. Additionally, this approach reliably identifies the majority of experimentally verified epitopes on the spike protein of SARS-CoV-2 not seen by the model during training and generalizes robustly on dissimilar data.

2. Results

2.1. Outline for a high performing B cell epitope predictor

A top-level overview of the BCE-Hunt approach to epitope prediction is illustrated in Fig. 1. The high performance we report in the subsequent sections below, delivered by BCE-Hunt, is achieved by the two main innovations that distinguish this approach compared to the state-of-the-art BCE predictors. Firstly, the approach circumvents the need for experimentally derived 3D protein structures and/or computationally predicted 3D protein structures (which are not yet fit for purpose[18]), by using protein sequence features and numerous other physiochemical features to predict 3D macrostructure properties (see Table 1 for the complete list of features and their sources). The 3D macrostructure properties are defined here as the key determinants of protein surface exposure of the protein sequence regions, namely, relative solvent accessibility (RSA), upper half-sphere exposure (HSE), lower half-sphere exposure (LHSE), and secondary structure (SS), as outline in Fig. 1A. Critically, in this first innovation, the 3D macrostructure features are learned from the unbound protein structures (Fig. 1A). The pipeline trains distinct BLSTM deep learning models for each of the 3D macrostructure properties to predict these properties from all the relevant unbound single 3D protein structure features in the Protein Data Bank (PDB) [19]. The output at prediction time for each of these BLSTM models is a score for each amino acid of the input protein sequence for each of the 3D macrostructure properties, thereby negating the need for knowing or predicting the exact coordinates of each atom in each amino acid relative to each other offered by complete and accurate experimental 3D protein structure determination (see Fig. 1A). Secondly, the proposed strategy redefines the definition of an epitope, encoding each epitope as a binary PPV, whereby 1′s represents direct amino acid contact points on the 3D protein sequence with the antibody, and 0′s represents amino acids on the protein sequence that are not interacting directly with the antibody (see Fig. 1B). This differs from the existing state-of-the-art predictors which primarily score each amino acid (AA) on a per AA basis for its potential contribution to the BCE interaction with the antibody. In the PPV definition of an epitope the entire BCE interaction structure is defined and predicted. Critically, in this second innovation, the BCE interaction structure is learned from the antibody-bound 3D protein complex structures in the PDB (Fig. 1B). These two unique innovations form the basis of the main BCE-Hunt predictor. As outlined in Fig. 1*B,* similar to the 3D macrostructure predictors, the main BCE predictor model is also trained using BLSTMs. However, in this step the training data is based on all the antibody-bound protein structure complexes from the PDB using the prediction output of the 3D macrostructure features in Fig. 1*A* in addition to the proteins sequence features (additional features are also used, see Table 1). At prediction time the input to the main BCE-Hunt

Table 1
Features used to train the 3D macrostructure predictors – Upper half sphere exposure (UHSE), Lower half sphere exposure (LHSE), Relative Solvent Accessibility (RSA), and secondary structure (SS). In addition to the features used to train the final main BCE prediction model, BCE-Hunt. The source information for each feature is referenced where relevant.

UHSE, LHSE & RSA	SS	BCE-Hunt
Polarity[45]	Conformational parameter for coil[46]	Bulkiness[47]
Free energy of transfer from inside to outside of a globular protein [48]	Average surrounding hydrophobicity[49]	Polarity[45]
Hydrophobicity[50]	Conformational preference for total beta strand (antiparallel and parallel)[51]	Molar fraction of buried residues[48]
Membrane buried helix parameter[52]	Normalized frequency for beta-sheet[53]	Conformational parameter for beta-turn[54]
Hydrophobicity[55]	Conformational parameter for alpha helix [46]	Average flexibility index [56]
Transmembrane tendency[57]	Side chain classes[58]	Normalized frequency for beta-sheet[53]
Proportion of residues buried[59]	Normalized frequency for beta-turn[53]	Normalized frequency for alpha helix[53]
Mean fractional area loss [60]	Binary representation of protein sequence	Side chain classes[58]
Conformational parameter for beta-sheet[54]		Known Number of codon (s) coding for each amino-acid in universal genetic code[58]
Conformational preference for total beta strand[51]		Side chain polarity[58]
Side chain classes[58]		Binary representation of protein sequence
Surface accessibility[61]		Secondary Structure (internal algorithm)
Side chain polarity[58]		Relative Solvent Accessibility (internal algorithm)
Binary representation of protein sequence		Positional Permutation Vectors

predictor model requires only the sequence, and the query PPV. The main output for BCE-Hunt is a score ranging from 0 to 1.0, representing the probability that the PPV for the primary protein sequence being queried is a true positive BCE. The BCE-Hunt complete pipeline workflow for the prediction of both linear and conformational BCE by BCE-Hunt is outlined in Supplementary Figure 1.

2.2. Evaluation metrics of the main BCE-Hunt BLSTM model: cross validation and independent tests

For both the 3D macrostructure BLSTM models and the main BCE BLSTM model we used~ 80 % of the training data on 5-fold cross-validation (CV) to assess the performance of the models, and the remaining 20 % was used as an independent test against the trained models. All the models performed well, without overfitting, and demonstrated high performance for all of the selected evaluation metrics. Moreover, all the models demonstrated high stability, with only a small variation observed between different CV runs. In particular, the model performance for the main BCE predictor is illustrated in Fig. 2. The model performed with a high precision-recall (PR) AUC of 0.8 compared the no skill value of 0.03 (robust across all CVs in Fig. 2A-2E) and comparable to the independent test in Fig. 2F. This BLSTM model predicts conformational BCEs trained on the antibody-bound protein 3D structures, with input features from the 3D macrostructure models of the corresponding unbound protein 3D structures (depicted in Fig. 1A and Supplementary Figure 1B), with the epitope encoded as PPVs. Although all the BLSTM models had promising and robust CV runs, and model training and independent tests exhibited high and robust performance, we next proceeded to evaluate the models against existing state-of-the-art BCE predictors.

2.3. Benchmark comparison of BCE-Hunt against state-of-the-art existing tools

Although the BLSTM and pipeline architecture of BCE-Hunt is setup for both linear and conformational BCEs, the current version is trained on conformational BCEs only. Based on the relevance for conformational BCE predictions, and operational availability for comparisons, three existing state-of-the-art tools were used to compare against, namely; Graphbepi [13], Discotope3.0 [11] and CBtope [12]. Each of these three different tools have different scoring and predictions systems and the scores per AA was interpreted individually for each tool. The data set used in these comparisons was an independent test dataset, not used in the training for BCE-Hunt. For this independent test we kept aside 6 % of the most recent antibody-bound protein 3D structures in the PDB for subsequent testing, and used the remaining (older) 94 % of the PDB data for training (to avoid test overlaps with the other algorithms).

Benchmarking against existing state-of-the-art conformational BCE tools is challenging for several reasons; [1] the tools use different approaches to define the epitopes i.e. they use different Armstrong (Å) distances between the epitope and the paratope contact points on the antibody to "identify" the participating amino acids. and [2] the other state-of-the-art methods use a per amino acid (AA) BCE contribution prediction approach, compared to our PPV method. That is, in the existing conformational BCE prediction tools each AA in the query protein sequence is scored according to its potential to contribute to a BCE antibody interaction. However, our tool BCE-Hunt conceptually defines an epitope based on PPVs representing the protein sequence, whereby we predict the entire epitope's direct contact points represented as a permutation vector for each individual query protein sequence. In BCE-Hunt a probability for each protein sequence and single PPV is then outputted per query at prediction time. Therefore, for BCE-Hunt, the metrics are defined as predicting the entire epitope's direct contact points and non-contact points with the antibody, whereby the existing tools assign a score for each AA on a protein sequence representing its potential contribution to participating in the BCE. This conceptual difference makes it challenging to directly compare BCE-Hunt against existing tools. However, to adjust for these conceptual differences, we made alterations to the architecture of BCE-Hunt such that in addition to outputting the probability scores of the PPV, the model also outputs a probability score per AA (to make comparisons against the existing tools possible).

In a first evaluation against the existing tools an experimental framework was devised that used a PPV/BCE-Hunt-like approach for defining the epitopes for the assessing performance (Fig. 3 A). Here, the entire epitope needs to be predicted positive for it to count in any of the tools being analyzed (meaning each AA that has been shown to contribute to the epitope experimentally has to be predicted correctly by the tool). For example, let us take a case where a true epitope in the test set is represented by the following permutation vector [**1**,0,0,**1**,0,**1**] (where the 1′s represent experimentally verified direct contact points with the antibody at a predefined Å distance, and 0′s represent no direct contact with the antibody in the experimentally verified 3D structure). In this example, if Discotope3.0, which defines an AA as positive if it has a score > 0.9 [11], generated the following set of scores [0.95,0.02,0.3, 0.98,0.2,0.91], this output would be assigned as a successful TP prediction for the AUC calculations.

. For this evaluation we allowed the existing tools to count a successful TP when some of the 0′s in the epitope defined by the permutation vector above were also counted as positives. So, for example, based on the same > 0.9 threshold for Discotope3.0, the following set of outputted scores [**0.95**,0.99,0.91,**0.98**,0.93,**0.91**] would also be assigned as a successful TP prediction for the AUC calculations. In contrast, we used a much stricter approach for evaluating BCE-Hunt, whereby the failure to correctly predict a non-contact point i.e. a 0, led to the output being assigned as a false negative for the AUC calculations. Since we only have true positives (TP) in the independent test of experimentally verified 3D antibody-bound protein structures, we used accuracy as the percentage of correct classifications (ACC) that each model in the comparison makes, as the evaluation metric ($\frac{TP}{P}$). An explanation for how this evaluation metric is derived is outlined in the methods section. The outcome of this first evaluation is summarized in Fig. 3 A, where it is clearly demonstrated that BCE-Hunt significantly outperforms the existing tools, based on an evaluation framework that defines an epitope using a PPV/BCE-Hunt-like approach.

In a second evaluation against the existing state-of-the-art tools an experimental framework was devised that used a "probability per AA-like" scoring approach for defining the epitopes for the assessing performance, which is conceptually similar to the approach adopted by the existing tools.

That is, if one of the existing tools identifies an AA on the experimentally verified epitope, then that is assigned as a TP positive.

For example, let us take a case where a true epitope in the test set is represented by the following permutation vector [1,0,0,1,0,1]. In this example, if Discotope3.0, which defines an AA as positive if it has a score > 0.9 [11] outputted the following set of scores [0.6,0.02,0.3,0.98,0.2, 0.7], this output would be assigned as a successful TP prediction for the AUC calculations, even though it has predicted only one of the positive AAs correctly in the entire epitope.

In contrast, we used a much stricter approach for evaluating BCE-Hunt, whereby all of the positive AAs experimentally verified to be physically participating in the interaction had to be be predicted correctly into be assigned as a true positive for the AUC calculations.

. It was interesting to observe in Figs. 3B and 3 C, that although the relaxed criterion (applied to the second type evaluation) was more favorable to the per AA epitope predictors, BCE-Hunt continues to significantly outperform the existing tools based on both PR and ROC AUC metrics.

For all the tools tested the threshold score for an AA was taken to be the default score as described in the respective published study for that algorithm [11–13]. The threshold for a true positive hit for an AA in an

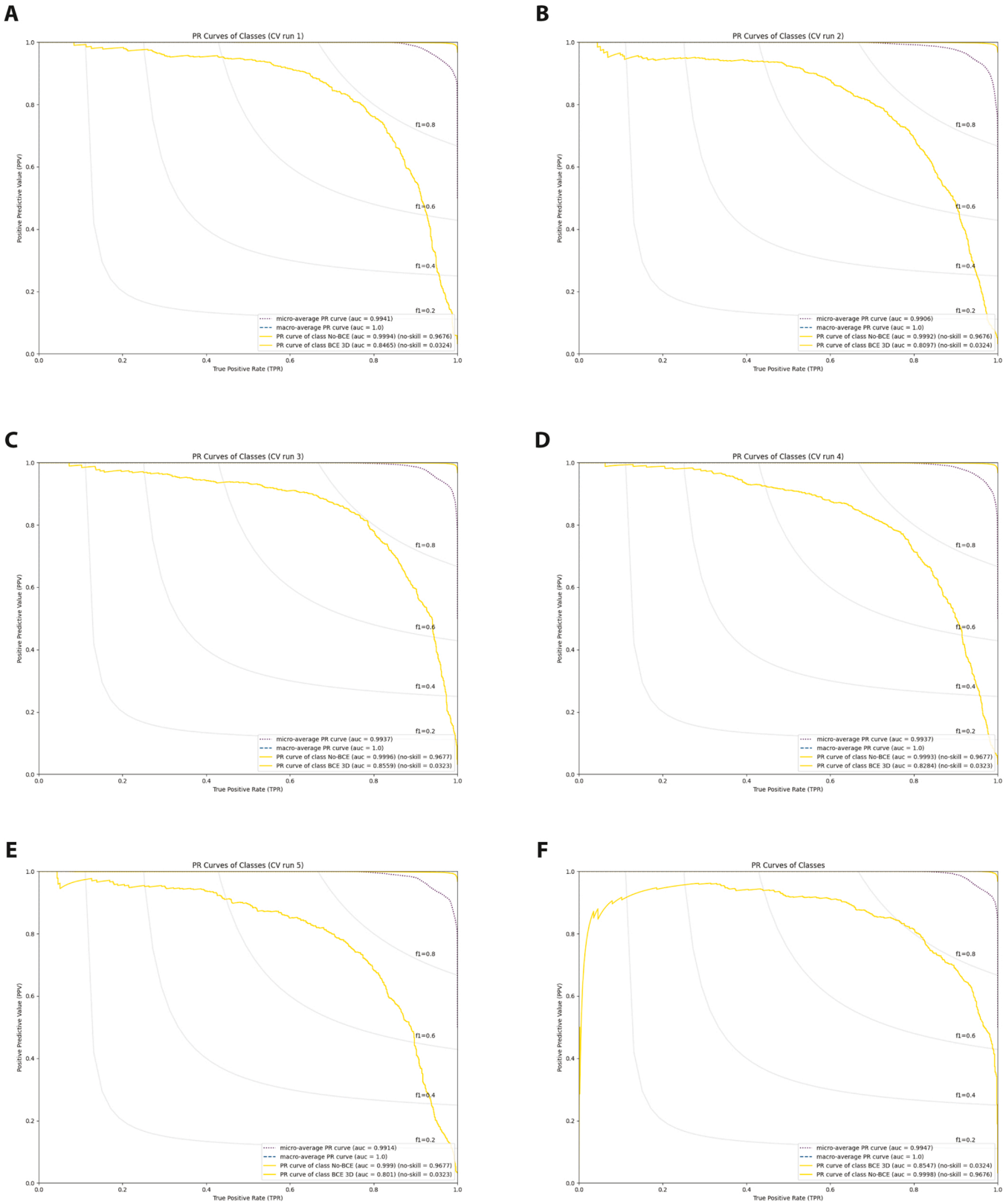
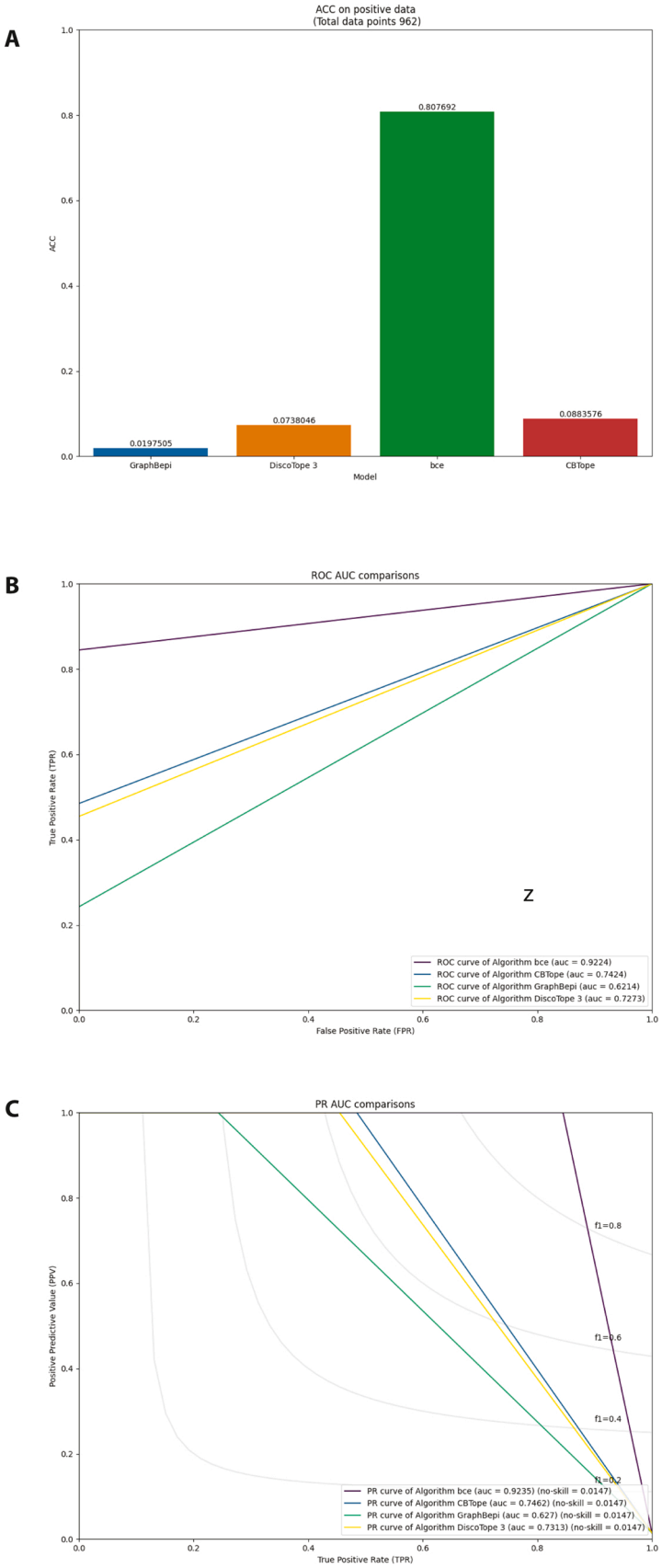


Fig. 2. Panels A-E on this figure shows results for the precision-recall (PR) metric for cross validation runs 1–5, respectively. PR was chosen for this evaluation due to the unbalanced nature of the dataset in terms of positives and negatives in the training data (see methods). The BCE-Hunt model exhibited a very high performance when using the no-skill PR metric. The average PR AUC for each of the CV runs 1–5 (panel A-E) was 0.83 compared to an average no skill PR of 0.03. The CV models (A-E) were trained for 459 epochs and independent test (F) for 583. In panel F we demonstrate the performance of the BCE-Hunt model on an independent, left out and non-redundant, test data set of PDB antibody protein binding structures. The performance in panel F illustrates that the BCE-Hunt model performs well on data not seen during training of the BLSTM model, the PR AUC curve depicts a robust and comparable performance compared to the CV results shown in A-E. "No-BCE" on the graphs are models trained on random positive and negative data.



(caption on next page)

Fig. 3. 3A highlights the superior performance of BCE-Hunt in comparison to existing conformational BCE predictors, using a criterion for success conceptually similar as to how an epitope is defined as PPVs in BCE-Hunt. 3B and 3 C highlights the significantly improved performance of BCE-Hunt compared to existing tools when using the criterion for success as a measure of the existing tools capability to predict at least on AA in the epitopes, for ROC and PR AUCs respectively. In both types of analyses (in 3A and 3B/C) BCE-Hunt must, strictly, predict the entire epitope (each AA in the epitope must be predicted correctly for directly physically interacting with the antibody and not directly interacting).

epitope for BCE-Hunt was very strict in these comparisons (for both types of evaluation); a probability score of being an epitope of > 0.96 was required to be deemed positive ($s = 1 - \frac{1}{31}$), given the 30X ratio of positive to negatives in the training data in BCE-Hunt.

2.4. Validation based on experimentally verified SARS-CoV-2 spike protein antibody epitopes

We next proceeded to assess the performance of the BCEP-Hunt model under a useful case example whereby artificial intelligence (AI) guided B cell antibody epitope mapping could potentially hold promising biomedical benefits. Specifically, we assess the ability of the model to identify *bona fide* epitopes on the spike protein of the SARS-CoV-2 virus. This was a highly relevant validation-case example due to the recent surge of studies that have characterized the neutralizing antibody interactions with the receptor binding domain of the spike protein since the declaration of the COVID-19 pandemic in March 2020. We extracted 177 spike protein-antibody complexes from experimental sources (such as X-ray crystallography or cryo-electron microscopy from the PDB, see Supplementary Table 1 for the list of PDB Ids tested). None of the PDB complex structures used in this test were present in the training data for BCE-Hunt. In turn we then extracted 312 epitopes from

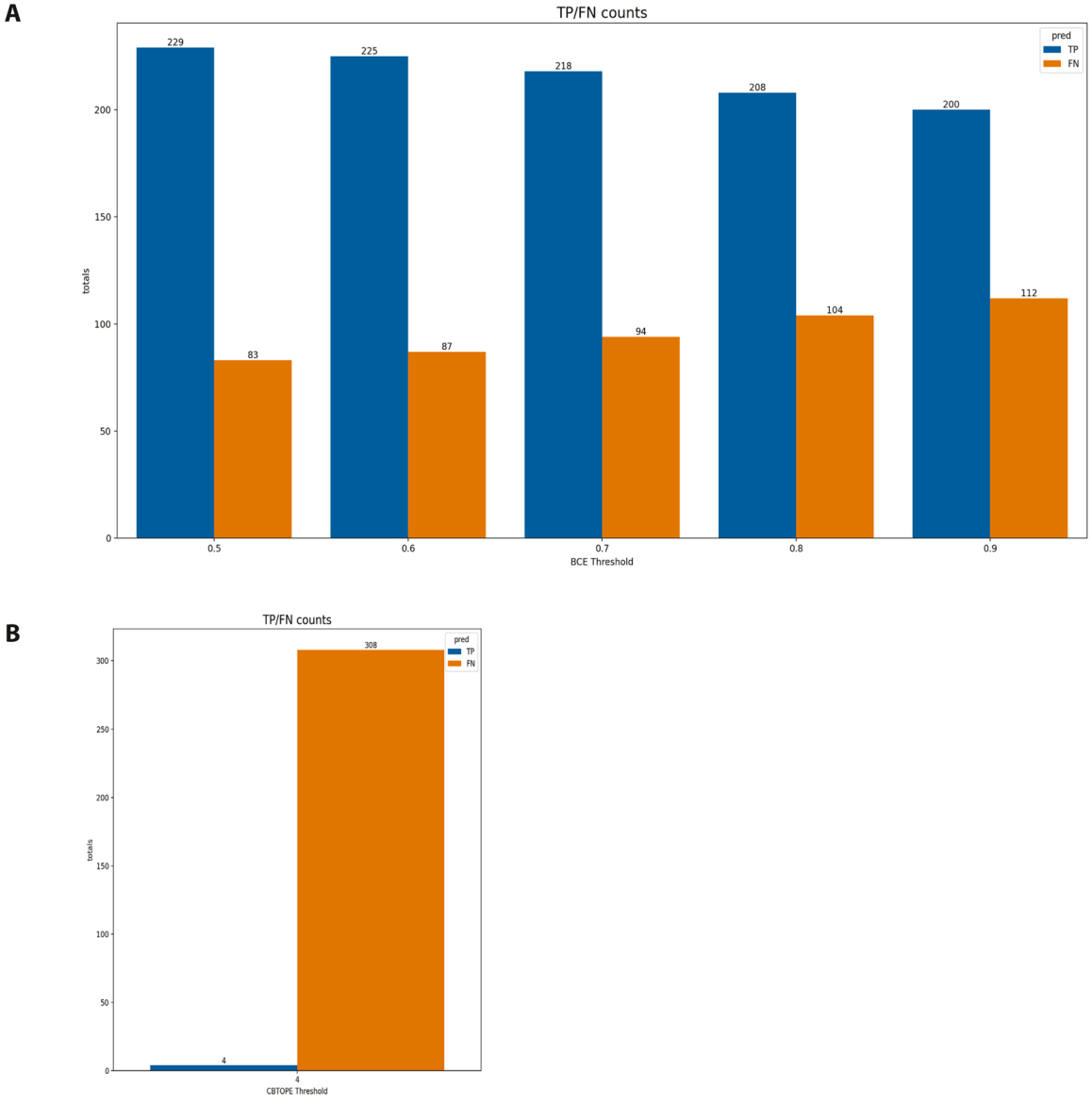


Fig. 4. (A) BCE-Hunt was successfully capable of recovering between 64 % to 75 % of bona fide epitopes experimentally validated antibody bound spike protein complexes (ranging from BCE-Hunt positive hit thresholds from 0.5 to 0.9). None of the antibody-spike protein complex in this test were included in the training data for BCE-Hunt, and the model needed to predict the epitope with a perfect match in the structural properties to be counted as a positive hit. (B) Using the same antibody bound SARS-CoV-2 spike test from the best performing state-of-the-art tools in Fig. 3, CBtope, it was clearly demonstrated that the current approaches to predict BCEs are not fit for purpose compared to the strategy outlined here.

these complexes. In Fig. 4 we demonstrate the ability of BCE-Hunt to correctly predict the majority of these 312 epitopes. At a very strict threshold score of 0.9 for BCE-Hunt we were able to successfully predict 64 % of these epitopes, and 73 % at the more relaxed score of 0.5 (Fig. 4 A). To gauge how the state-of-the-art existing tools would perform on the same data, we chose the best performing tool from the independent benchmarking in Fig. 3, CBTope [12], and demonstrated that it could only predict a mere 0.01 % of the BCEs (see Fig. 4B). In each epitope test, a strict criterion was forced on BCE-Hunt in that it had to predict the entire epitope as a perfect match.

2.5. Validation based on dissimilar experimentally verified antibody-bound protein 3D structures

Whilst we believe that the significant improvement in performance demonstrated by BCE-Hunt compared to the current state-of-the-art

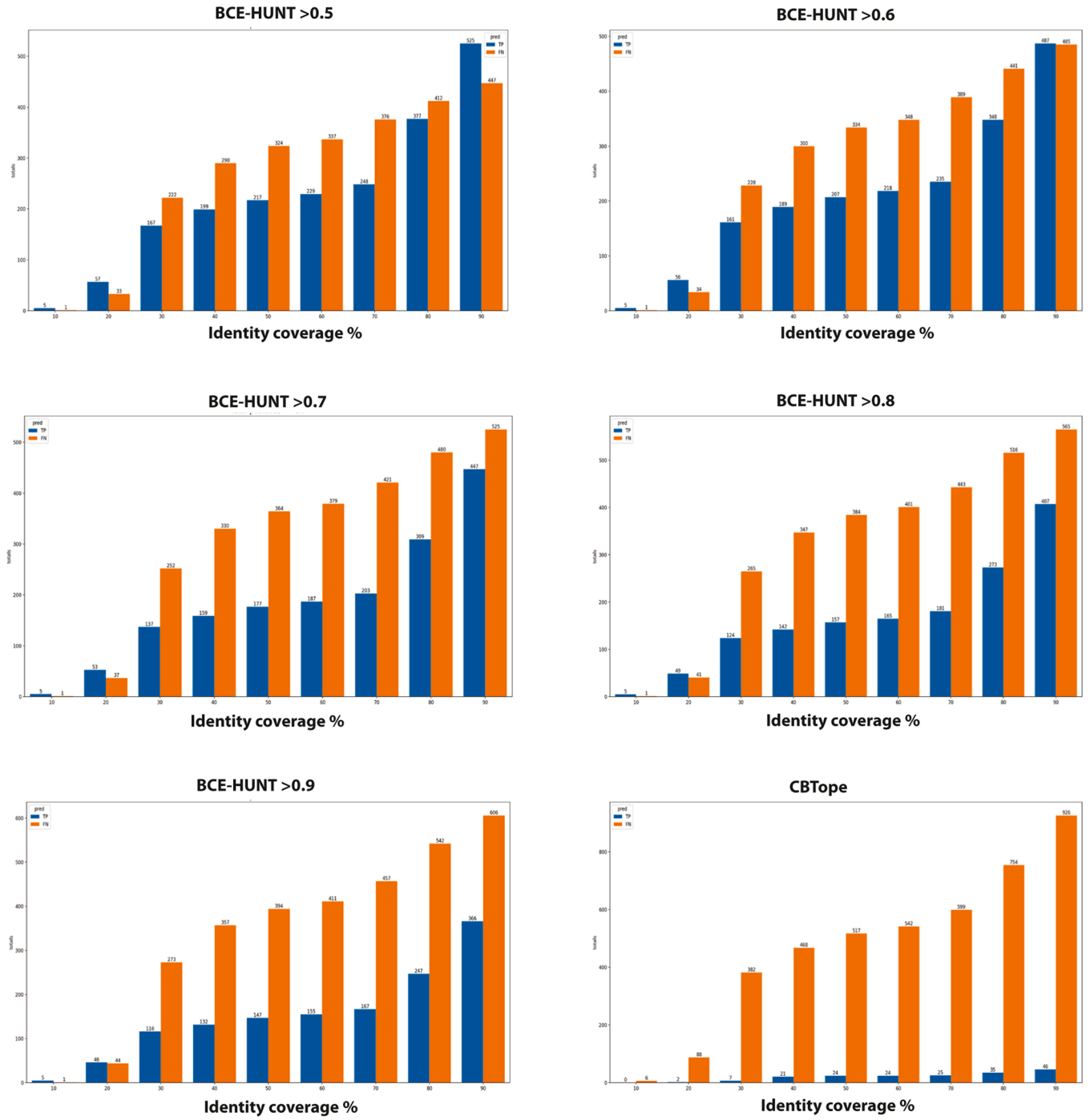


Fig. 5. Each plot shows the true positive (TP) epitopes that are identified as such from the predictor, on the given probability threshold (BCE-Hunt positives, ranging from positive probabilities of 0.5 to 0.9), and the false negative (FN) epitopes that are not identified as such from the predictor, on the given probability threshold. The X-axis depicts the identify coverage % threshold cutoff of the test data. Each threshold on the X-axis represents the population of BCEs on experimentally verified antibody-bound protein structures that are below the given threshold, and both unseen by the model and dissimilar to all the protein sequences in the training data. This analysis demonstrates the ability of BCE-Hunt to generalize and successfully predict on unseen data in a robust manner. This performance was far superior to the best state-of-the-art tool analyzed in Fig. 3, CBTope.

tools demonstrated in the above study reflects the unique and distinguishing innovations underpinning the approach, which more accurately model the underlying biology of BCE/antibody recognition. There is a counter-argument that the high performance could potentially be driven by the presence of "similar" antibody-bound spike protein structures from non-SARS-CoV-2 coronaviruses present in the BCE-Hunt training data (albeit, also likely for the existing state-of-the-art tools). Therefore, to investigate this potential issue we next assessed if BCE-Hunt can successfully predict experimentally validated BCEs that are not only unseen previously by the model as demonstrated in Fig. 4, but also dissimilar to the protein sequences that exist in the training or previous test data. To perform this strict test, we fetched all data from the PDB that was not present in any of our previously used training or test data sets. Global alignments on these sequences were performed using Needle-EMBOSS [20] to identify dissimilar proteins at various cutoffs of sequence coverage identity. A protein sequence was kept if its sequence identity was equivalent to or lower than at least one of the protein sequences in all the training data for the given threshold % identity coverage (see Supplementary Table 2 for the final list of antibody-bound 3D protein structures used for this test). Fig. 5 illustrates that BCE-Hunt can generalize and successfully predict on unseen and dissimilar data in a robust manner. For example, at the strictest threshold for the probability of being a positive BCE (>0.9), and at the strictest threshold of identity coverage (20 %); BCE-Hunt successfully predicted 51 % of the true positive epitopes (Fig. 5). The best performing existing state-of-the-art tool, CBTope, predicted only 2 % of the true positive epitopes at 20 % threshold for sequence identity.

3. Discussion and conclusions

A high performing machine learning (ML) model that accurately predicts BCEs can facilitate the identification of vaccine and diagnostic candidates for further experimental validation and development much faster and more efficiently than wet-lab based experimental approaches alone. Such models can vastly reduce the cost and time related to the BCE mapping process and streamline the identification of potential epitope candidates for further clinical investigations.

Many approaches to predict BCEs have been attempted [7,10, 21–25], but their performance is poor and there is significant room from improvement. Despite the development of several BCE predictors in recent years, computational BCE predictions have a long way to progress toward their practical use to guide the development of vaccines, antibody-based therapeutics and diagnostics. This was notable during the recent COVID-19 pandemic, whereby a wealth of additional ML training data became available due to the large burst of experimental studies on the spike protein from the SARS-COV-2 virus [24]. This increased volume of spike glycoprotein structures, including antibody-bound structures, together with the more than 20 years of development of BCE predictors, could have potentially contributed significantly to the AI accelerated design of vaccines or monoclonal antibody-based therapeutics, however, in the absence of reliable fit-for purpose tools the field reverted to 3D cryo-electron microscopy to guide the design of antibody-based therapeutics [26].

Clearly, there is ample scope for improvement of BCE predictors, and here we present a novel approach that encompasses several unique and distinguishing innovations that more accurately model the underlying biology of BCE/antibody recognition and improve performance. The first innovation is the use of a more structural and precise definition of the BCE which captures the entire protein antigen sequence and embedded epitope in a binary vector as positional PPVs. In a PPV, 1's represent direct amino acid contact points on the 3D antigen structure with the antibody, and 0's represent amino acids that do not interact directly with the antibody. The second innovation is the use of the unbound 3D antigen structure (rather than the antibody-bound structure) as a source of features to capture the pertinent 3D structural properties of the antibody-BCE binding interaction (before the antibody binding event occurs) termed here as the 3D macrostructures. Proteins undergo conformational changes upon binding which are important to their biological function [15,17], therefore by focusing on unbound structures, we provide the model with more accurate information regarding BCE/antibody recognition, consequently improving epitope prediction.

A similar concept underlying the second innovation, i.e. the use of the unbound protein antigen 3D structure to capture features before the binding event, has been reported previously in the literature [27–32]. However, these previous antibody bound/unbound approaches used features captured from the unbound 3D protein antigen structure to directly train their respective BCE predictor models, and their definition of BCEs was on a per amino acid basis (not using PPV-based epitope definitions). In the approach proposed here, the unbound 3D protein antigen structures are used to train several distinct ML models that capture the 3D macrostructure of 3D folded proteins (e.g., RSA, HSE, and SS). Using the target antigen protein sequence alone, the outputs from the trained 3D macrostructure models are in turn used as features to train our main BCE predictor model. When combined, both innovations outlined in the proposed approach, synergize to offer significant performance improvements compared to current state-of-the-art approaches. One of the limitations of the proposed approach is the fact that the user must query an almost unlimited number of PPVs to fully map the complete BCE potential in any sufficiently large antigen structure. This problem is currently handled by using a brute force approach where a random selection of hundreds of thousands of candidate PPVs, guided by priors from empirical BCE data are used, which are hopefully relatively representative of the broader BCE potential. However, this approach will undoubtedly miss good candidates and in future versions of this tool, we intend to integrate deep reinforcement learning (DRL) strategies to identify an even more representative and optimized set of candiate PPVs to better map the complete BCE potential. In recent years, BLSTMs have been used to capture the global properties that may define the landscape of BCE in a protein [13, 33]. The use of BLSTMs have a key advantage as they allow us to capture global relationships between distant amino acids which might constitute a conformational BCE. In contrast, other algorithms used in the other tools described earlier, segment the protein sequences and consequently risk losing these relationships. However, the previous BLSTM approaches, and most other approaches limit the identification of BCEs on a protein to outputting a score on a per amino acid basis. The proposed advantage of the BCE as a binary PPV is that the structural properties of the epitope-paratope interactions can be captured with improved fidelity. Even though, BCE-Hunt can also be adjusted to perform predictions on a per amino acid basis (as performed in this study for benchmarking purposes), the PPV definition of a BCE is preferable as it encodes a high-resolution, more representative, definition of each amino acid that constitutes the epitope, leading to significantly higher performance as reported here.

Another, advantage of the approach we describe herein is that it circumvents the need for an experimentally validated 3D structure to be available at the input stage, which is required by other approaches described earlier such as the DiscoTope family of BCE prediction tools [11,34,35]. A recent development [11] in the BCE prediction field to address the limitation of requiring the PDB 3D protein structure as input, is to integrate the input query with the input 3D protein structure obtained from databases such as the Alphafold protein structure database [36]. Although the Alphafold method [37] is indeed a ground-breaking advancement in 3D protein structure prediction, and serves as an important hypothesis generator, 3D protein structure predictions are arguably not yet a direct placement for experimental structure determination [18]. Therefore, strategies such as BCE-Hunt described herein, are advantageous in that they only require the protein sequence to perform reliable BCE predictions. As mentioned above, BCE-Hunt bypasses they need for the experimentally derived 3D protein structure or potentially unreliable predicted 3D protein structure at prediction-time, by capturing the key 3D macrostructure properties of the protein from

the unbound protein antigen during ML training in the pipeline. We demonstrated in this study that the precise coordinates of each atom in each amino acid relative to each other the 3D protein structure is not necessary to be known with fidelity to capture accurate BCE antibody interactions. The determinants of the 3D protein structure captured by the 3D macrostructure predictors from the unbound protein antigen are sufficient.

Given the failure of the existing tools to reliably predict BCEs on the spike protein of the SARS-Cov-2 virus [7], the validation case study we report herein on unseen SARS-Cov-2 antibody bound spike protein structures suggests that the approach proposed in this study offers an important advancement in the field and may help pave the way towards a future where computational BCE prediction is routinely used in a wide range biomedical research applications and to help design future vaccines and immunodiagnostics.

4. Methods

4.1. Pretraining techniques

For ensuring reproducible results and avoiding the need of random seed in the networks used in our BLSTM models, we used auto-encoders (AEs). AEs are types of neural networks which mirror their inputs. More specifically, an AE takes an input at layer number i = 0.

and processes it through an arbitrary number of layers, say i = 1, …, N, which constitute the encoder part. It then processes it back through a mirrored structure of the encoder part, called the decoder part. Finally, it returns the output which is the same as.

the input. The intermediate layers might reduce or expand the dimensions of the input, as the main use of AEs was indeed dimensionality reduction. Firstly, we randomly generated 10000 protein sequences of length between 50 to 1000 amino acids. The actual amino acids used was random as well. We then pre-trained each layer of the neural network as an individual AE on the whole generated data. The pre-training was done for maximum 1000 epochs for each AE, with the possibility of early stopping if the loss did not decrease after 10 epochs. The weights of the outer layers were copied to deeper AEs after pre-training and were also kept constant during the pre-training of those. The last layer of the model (output) was not pre-trained at all. This procedure was done once, and the same pre-trained weights were used on all the validation procedures. This procedure was faster, it used much more data to pre-train and the same initial weights were used for all downstream analyses.

4.2. BLSTMs

Bidirectional long short-term memory networks (BLSTMs) consist of two LSTMs, each one scanning the time-steps sequence from either direction. That is, one LSTM scans the forward and one the backward sequence of time-steps. This allows the network to capture relationships between past and future time-steps at once. For predicting conformational BCEs, we use BLSTMs models. This allowed us to model the whole protein sequences as one observation, without the need of segmenting it. Therefore, distant amino-acid relationships should be able to get captured by the model. Moreover, the use of BLSTMs allows us to train different protein lengths simultaneously.

4.3. Data preparation for the 3D macrostructure features (unbound 3D protein structure data preparation)

Proteins are dynamic molecules that often undergo significant conformational changes upon binding to other molecules, a process known as induced fit [17]. This phenomenon is fundamental to protein function and influences many biological processes. By extracting features from unbound protein structures in the machine learning training steps, we captured the native state of the protein, providing a more accurate representation of potential antibody binding sites. Learning from the 3D structure of proteins after the binding event might result in capturing artifacts of the induced fit, leading to potentially incorrect information. Therefore, our BLSTM models to learn 3D macrostructures were trained on unbound structures to ensure high fidelity in epitope prediction.

To accurately predict the 3D macrostructure features (SS, RSA, UHSE and LHSE) from a native protein sequence, we used 3D protein structures from the PDB [19] from all available organisms. The goal of those models was to predict the surface and structural characteristics of proteins that are not affected by any other protein, including antibodies (Abs). Therefore, we kept only structures that are not bound by any other structure or molecule. Structures containing more than one copy of the same molecule, but slightly different conformation, are kept in the data. We filtered the structures and kept only those with ≤ 3°A resolution, ensuring that every atom of each amino-acid is mapped with coordinates, and with protein chains longer than 200 amino-acids. After filtering, the database consisted of 41592 total structures (per 20/12/2019). Those structures contained 70489 protein sequences which resulted in 53524 unique sequences. Subsequences of longer sequences were kept as different data points. The reason for that is that their conformational characteristics might be different because of their shorter length.

The DSSP algorithm was used to compute the SS and the RSA for each molecule in each structure file [38]. DSSP computes the following secondary structure classes for a protein sequence; α-helix, 310-helix, π-helix, isolated β-bridge, β-strand, turn, bend and coil. We merged those classes into three super-classes; Helices (α-helix, 310-helix and π-helix), Strands (isolated β-bridge and β-strand) and Coils (turn, bend and coil). Finally, the BioPython package [39] was used to compute the UHSE and LHSE. At the end of the filtering, each amino acid in the data base was assigned a value for the RSA, UHSE, LHSE and a class for the SS. To create a unique data base, the mean per amino-acid was taken for RSA, UHSE, LHSE among identical sequences. Finally, amino-acids of identical sequences but with different SS classes, were assigned to the coil class.

4.4. BLSTM prediction model for RSA and HSE (3D macrostructure features)

We chose to create a single BLSTM model for predicting all surface features. More specifically, the model predicts RSA, UHSE and LHSE from a primary protein sequence. Although LHSE might not give useful information about the surface position of an.

amino acid, it might help predicting UHSE more accurately, since both together are actually forming a prob around each amino-acid. This is a BLSTM model which takes as inputs a batch of features, each computed per amino-acid from each input sequence, and.

predicts a three-way output. For each protein sequence given as input, a value for each RSA, UHSE and LHSE are predicted per amino-acid. For all the three outputs the individual losses were the mean square error (MSE). The global model loss was the weighted sum of the individual losses with weights: 50, 100 and 125 for RSA, UHSE and LHSE, respectively. The weights were decided by first training the model without them and see the differences in the magnitude of the three losses. The weights contribute in such a way that the three losses give the same contribution to the global model loss. Supplementary Figure 2 A shows the network architecture of the model, and Table 1 outlines the features used to the train the BLSTM model.

4.5. BLSTM prediction model for secondary structure (3D macrostructure features)

We also chose to create a BLSTM model for predicting a three class output for SS.The model uses the categorical cross-entropy loss in order to assign one of the following classes to each amino-acid in an input

sequence; Helices, Strands and Coils. Supplementary Figure 2B shows the network architecture of the model, and Table 1 outlines the features used to the train the BLSTM model.

4.6. Data preparation for the conformational BCE predictor (BCE-Hunt)

For modelling conformational BCEs (CBCE) we downloaded non-obsolete protein complexes from the PDB[19]. We allowed structures of any resolution and organisms, if they had at least three different protein chains. The reason for this is that two of the chains might be the variable fragment heavy (VH) and variable fragment light (VL) chains of an antibody (Ab), and the third chain might be an antigen (Ag). Of course, there might be multiple Abs or Ags in one PDB structure. We created a local database using all immunoglobulin V-, D- and J-region genes from the international ImMunoGeneTics information system (IMGT)[40]. Those genes were downloaded for both VH and VL chains, from all the available organisms, that is, human, mouse, rhesus monkey, rabbit, and rat. The IMGT genes not only provide information about the VH and VL chains of an Ab, but its paratope regions as well, that is, the complementarity-determining regions (CDR) and framework regions (FR) of each chain. To identify the Abs, we used IgBlast [41] for protein sequences on the IMGT database. For identifying the paratope in a chain we used the Kabat system that the IgBlast provides [42]. We considered VH and VL chains as valid, only if at least their CDR1 and CDR2 were found by IgBlast. Since CDR3 is more difficult to map [43], we allowed it to be missing. We blasted all the protein chains of each structure to that database. Structures were filtered out if they did not have at least one chain mapped as valid VH and one as valid VL. Any other chain that was not mapped as VH or VL was assumed to be an Ag chain.

We considered as valid Ag chains those chains that were longer than 100 amino acids and were not bound by any other chain other than a VH or VL chain. Finally, only structures including at least one VH, VL and Ag chains were taken to further analyses. We paired VH and VL chains in each structure to recreate valid Abs. In case there were multiple VH and VL chains in one single structure, we measured the mean distance from every atom on each VH to every atom on each VL chain. VH and VL chains with the minimum mean atom distance were assigned as pairs and assumed to belong to one single Ab. Stand-alone VH or.

VL chains were not considered on the downstream analysis as they could not define a complete Ab. A paratope analysis provided information about the contacts formed with each Ag. Two amino acids were assumed to be in contact if any of their atoms were located within a certain probe distance from each other. We computed the total number of Ag amino acids that form contacts with each paratope's parts, using probe distances of 4, 6 and 8°A. Most of the contacts are made with CDR1 and CDR2 of the VHs and CDR1 and CDR3 of the VLs. The absence of CDR3 on VHs might be due to the difficulty of identifying it using IgBlast. Moreover, the FR regions do not seem to form as many contacts as the CDR regions, as expected. Additionally, the standard deviation of the total contacts per amino acid and paratope part was relatively low, indicating that similar number of contacts are made between different Ags and Abs. This could be an indication of the possibility of predicting CBCEs without any further information about the actual Abs. For each VH and VL chain pair we defined a single CBCE. This was done by first identifying atoms on any Ag whose distance from the CDRs regions of any VH and VL chain pair was ≤ 4°A. The amino acids that those atoms belonged to were defined as contacts between the Ag and the Ab, that is, they defined the CBCE. Multiple CBCEs from different Abs could be defined on the.

same Ag. Finally, structures that did not define any CBCE within the 4°A distance were discarded. Observed CBCEs were also mapped to similar Ag. Undiscovered CBCEs could increase the false negative rate of the prediction models. To decrease the possibility of assigning undiscovered CBCEs as negative data, we copied observed CBCEs to.

similar Ags. First, we identified clusters of similar Ags using BlastPlus [44] with $>$ 90 % similarity and at most two gaps. CBCEs were copied from an Ag in a cluster to all the other Ags in the same cluster if their corresponding position and distancing was the same based on the mapping from BlastPlus. Lastly, we created a unique Ags database. Duplicated Ags were removed from the data. Second, Ag sequences that were sub-sequences of longer Ags were kept in the data and treated as different observations. The resulting database consisted of 1003 Ag sequences in fasta format and 12968 CBCEs from which 6986 were unique. Each Ag sequence was associated with at least one CBCE.

4.7. B cell epitope definition as a positional permutation vector

A positional permutation vector (PPV) is defined here as a binary vector used to encode the interaction between a protein antigen and an antibody. Each position in the vector corresponds to an amino acid in the protein sequence. A value of '1' in the PPV indicates that the amino acid at that position directly contacts the antibody, while a '0' indicates that the amino acid does not interact directly with the antibody. This encoding captures the spatial arrangement of contact points, allowing the model to learn and predict the interaction sites accurately.

4.8. BLSTM prediction model for the conformational BCE predictor (BCE-Hunt)

The models utilize input features that include, among others, a permutation vector, also known as a positional permutation vector (PPV). Every sequence in the data is associated with at least one true CBCE, those CBCEs are turned into binary 2D vectors and are given as input to the models. The model's primary output is a probability, which essentially addresses the query: "Does this particular permutation vector within the given specific sequence accurately constitute a true CBCE?". The second output of the model is the permutation vector itself. This part of the model works as an AE, where it returns a probability per amino acid. This output can be seen as a contribution of each amino acid in the sequence to the specific CBCE in question. The goal was to predict CBCEs before the binding event took place. The dataset used comprises linear protein sequences and the true CBCEs. Therefore, it lacks information concerning the structural or surface characteristics of these protein sequences. However, an understanding of the pre-binding secondary structure and surface of each protein is crucial for our analysis. Therefore, we used our prediction models for RSA, UHSE, LHSE and SS (3D macrostructure features) to predict those characteristics in every protein in the dataset. Supplementary Figure 2 C shows the network architecture of the BCE-Hunt model, and Table 1 outlines the features used to the train the BLSTM model.

The BCE AA output is the actual permutation output, where binary cross-entropy loss was used. The BCE perm output is a probability vector indicating if the input permutation is a true CBCE (second position on vector) or not (first position on vector). The input features were computed per amino acid as they are, not averaged by windows. The BCE perm output is computed from amino acid values. The last layer of the output is a Dense layer with sigmoid activation. Such that, for each protein sequence, a value in $\in [0,1]$ is returned per amino-acid. Then a 2-class probability vector is computed for that sequence as.

$$1 - \frac{\sum_{i=0}^{N} X_i}{N}, \quad \frac{\sum_{i=0}^{N} X_i}{N}$$

where N is the length of the sequence and X_i is the dense layer output on amino-acid at position i on the sequence. This 2-class probability vector is then used in the binary cross-entropy loss.

4.9. Negative data generation for the conformational BCE predictor (BCE-Hunt)

We applied three different negative data generating methods. For each sequence in the training and validation data we generated 30 completely random permutations, 10 from each method, and assumed that they are not true CBCEs. In the first method we generated

completely random permutations. For each sequence in the training and validation data we generated 10 completely random permutations and assumed that they were not true CBCEs. Both the total number of amin -acids and the placement of those in the given sequence were random. New permutations were generated on each epoch. The advantage of this method is that because of the complete randomness, it is quite unlikely that any generated CBCE permutation will be false negative. However, the disadvantage is that the randomly generated CBCE permutations might be extremely different than the true CBCE permutations. In that case, the algorithm might learn to separate only based on the permutation input. The second and third methods correct for this disadvantage, which we also applied on every true CBCE of the given protein sequence and generated on each epoch. The second method kept the first and last amino acid of a given true CBCE at the correct position, while it randomly shuffled the internal CBCE amino acids indicators inside the region. The third method kept the total amount of amino acids of a true CBCE, as well their linear distance, constant. It then randomly shifted the whole true CBCE on other parts of the protein. The advantage of this method is that the randomly generated CBCE permutations cover both extremely similar and extremely different true CBCE permutations. This is likely to result in an algorithm that is more robust to both positive and negative data. Conversely, a drawback of the method is the potential for a substantial rise in false negatives, which could lead to diminished performance when applied to novel data sets.

4.10. Evaluation metrics

For the evaluations against the existing state of the art tools in Fig. 3A, ACC was used as the metric, which was derived here as the true positive rate (TPR) $\frac{TP}{P}$. This was derived due to $ACC = \frac{TP+TN}{P+N}$. For those comparisons, we only used scientific verified antibody binding epitopes from the PDB. For that reason, and because we did not want to assume negative epitopes and risk FP rates, both the TN and N are zero. The ACC therefore in effect becomes the true positive rate (TPR), $TPR = \frac{TP}{P} = ACC$.

For AUC calculations we assume that the 0′s representing AAs in the PPV are negative which allowed us to capture the TN rate in addition to the TP rate to perform the evaluations illustrated in Fig. 3B and C.

Author statement

We hereby declare that all authors have read and approved the revised version of the manuscript.

CRediT authorship contribution statement

Richard Stratford: Writing – review & editing, Project administration, Investigation, Formal analysis. **Trevor Clancy:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Irantzu Anzar:** Writing – review & editing, Formal analysis. **Boris Simovski:** Writing – review & editing, Methodology, Formal analysis. **Ioannis Vardaxis:** Writing – review & editing, Writing – original draft, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

All authors are employees at the company NEC OncoImmunity AS.

Acknowledgements

The study was funded by the Research Council of Norway (Grant Number: 282216).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.06.005.

References

[1] Getzoff ED, Tainer JA, Lerner RA, Geysen HM. The chemistry and mechanism of antibody binding to protein antigens. Adv Immunol 1988;43:1–98.
[2] Van Regenmortel MH. Immunoinformatics may lead to a reappraisal of the nature of B cell epitopes and of the feasibility of synthetic peptide vaccines. J Mol Recognit 2006;19(3):183–7.
[3] Dudek NL, Perlmutter P, Aguilar MI, Croft NP, Purcell AW. Epitope discovery and their use in peptide based vaccines. Curr Pharm Des 2010;16(28):3149–57.
[4] Ahmad TA, Eweida AE, Sheweita SA. B-cell epitope mapping for the design of vaccines and effective diagnostics. Trials Vaccinol 2016;5:71–83.
[5] Leinikki P, Lehtinen M, Hyoty H, Parkkonen P, Kantanen ML, Hakulinen J. Synthetic peptides as diagnostic tools in virology. Adv Virus Res 1993;42:149–86.
[6] Potocnakova L, Bhide M, Pulzova LB. An introduction to B-cell epitope mapping and in silico epitope prediction. J Immunol Res 2016;2016:6760830.
[7] Cia G, Pucci F, Rooman M. Critical review of conformational B-cell epitope prediction methods. Brief Bioinform 2023;24(1).
[8] Zheng D, Liang S, Zhang C. B-cell epitope predictions using computational methods. Methods Mol Biol 2023;2552:239–54.
[9] Blythe MJ, Flower DR. Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci 2005;14(1):246–8.
[10] Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and methods for T- and B-cell epitope prediction. J Immunol Res 2017;2017:2680160.
[11] Hoie MH, Gade FS, Johansen JM, Wurtzen C, Winther O, Nielsen M, et al. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. Front Immunol 2024;15:1322712.
[12] Ansari HR, Raghava GP. Identification of conformational B-cell epitopes in an antigen from its primary sequence. Immunome Res 2010;6:6.
[13] Zeng Y, Wei Z, Yuan Q, Chen S, Yu W, Lu Y, et al. Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model. Bioinformatics 2023;39(4).
[14] Caoili SEC. Comprehending B-cell epitope prediction to develop vaccines and immunodiagnostics. Front Immunol 2022;13:908459.
[15] Koshland DE. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci U S A 1958;44(2):98–104.
[16] Keskin O. Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: a case study of antibodies. BMC Struct Biol 2007;7:31.
[17] Rini JM, Schulze-Gahmen U, Wilson IA. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. Science 1992;255(5047):959–65.
[18] Terwilliger TC, Liebschner D, Croll TI, Williams CJ, McCoy AJ, Poon BK, et al. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nat Methods 2024;21(1):110–6.
[19] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28(1):235–42.
[20] Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res 2022;50(W1):W276–9.
[21] Backert L, Kohlbacher O. Immunoinformatics and epitope prediction in the age of genomic medicine. Genome Med 2015;7:119.
[22] Sun P, Guo S, Sun J, Tan L, Lu C, Ma Z. Advances in In-silico B-cell epitope prediction. Curr Top Med Chem 2019;19(2):105–15.
[23] Lundegaard C, Lund O, Kesmir C, Brunak S, Nielsen M. Modeling the adaptive immune system: predictions and simulations. Bioinformatics 2007;23(24):3265–75.
[24] Bukhari SNH, Jain A, Haq E, Mehbodniya A, Webber J. Machine learning techniques for the prediction of B-cell and T-cell epitopes as potential vaccine targets with a specific focus on SARS-CoV-2 pathogen: a review. Pathogens 2022;11(2).
[25] Sun P, Ju H, Liu Z, Ning Q, Zhang J, Zhao X, et al. Bioinformatics resources and tools for conformational B-cell epitope prediction. Comput Math Methods Med 2013;2013:943636.
[26] Taylor PC, Adams AC, Hufford MM, de la Torre I, Winthrop K, Gottlieb RL. Neutralizing monoclonal antibodies for treatment of COVID-19. Nat Rev Immunol 2021;21(6):382–93.
[27] da Silva BM, Myung Y, Ascher DB, Pires DEV. epitope3D: a machine learning method for conformational B-cell epitope prediction. Brief Bioinform 2022;23(1).
[28] Ren J, Liu Q, Ellis J, Li J. Positive-unlabeled learning for the prediction of conformational B-cell epitopes. BMC Bioinforma 2015;16(Suppl 18):S12 (Suppl 18).
[29] Dalkas GA, Rooman M. SEPIa, a knowledge-driven algorithm for predicting conformational B-cell epitopes from the amino acid sequence. BMC Bioinforma 2017;18(1):95.
[30] 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(5):e0128194.
[31] 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.

[32] Zhang W, Xiong Y, Zhao M, Zou H, Ye X, Liu J. Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. BMC Bioinforma 2011;12:341.
[33] Lu S, Li Y, Ma Q, Nan X, Zhang S. A structure-based B-cell epitope prediction model through combing local and global features. Front Immunol 2022;13:890943.
[34] Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci 2006;15(11):2558–67.
[35] Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol 2012;8(12):e1002829.
[36] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50(D1):D439–44.
[37] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.
[38] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637.
[39] Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25(11):1422–3.
[40] Manso T, Folch G, Giudicelli V, Jabado-Michaloud J, Kushwaha A, Nguefack Ngoune V, et al. IMGT(R) databases, related tools and web resources through three main axes of research and development. Nucleic Acids Res 2022;50(D1):D1262–72.
[41] Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 2013;41(Web Server issue):W34–40.
[42] Johnson G, Wu TT. Kabat database and its applications: 30 years after the first variability plot. Nucleic Acids Res 2000;28(1):214–8.
[43] Lefranc M.-P., Lefranc G. The immunoglobulin factsbook. San Diego: Academic Press; 2001. xiv, 457 p. p.
[44] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinforma 2009;10:421.
[45] Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974;185(4154):862–4.
[46] Deleage G, Roux B. An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1987;1(4):289–94.
[47] Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 1968;21(2):170–201.
[48] Janin J. Surface and inside volumes in globular proteins. Nature 1979;277(5696):491–2.
[49] Manavalan P, Ponnuswamy PK. Hydrophobic character of amino acid residues in globular proteins. Nature 1978;275(5681):673–4.
[50] Guy HR. Amino acid side-chain partition energies and distribution of residues in soluble proteins. Biophys J 1985;47(1):61–70.
[51] Lifson S, Sander C. Antiparallel and parallel beta-strands differ in amino acid residue preferences. Nature 1979;282(5734):109–11.
[52] Mohana Rao JK, Argos P. A conformational preference parameter to predict helices in integral membrane proteins. Biochim Biophys Acta 1986;869(2):197–214.
[53] Levitt M. Conformational preferences of amino acids in globular proteins. Biochemistry 1978;17(20):4277–85.
[54] Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzym Relat Areas Mol Biol 1978;47:45–148.
[55] Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18(3):534–52.
[56] Bhaskaran R, Ponnuswamy Pk. Positional flexibilities of amino acid residues in globular proteins. Int J Pept Protein Res 1988;32(4):241–55.
[57] Zhao G, London E. An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. Protein Sci 2006;15(8):1987–2001.
[58] Cooper GM, Hausman RE. The Cell: A Molecular Approach. ASM Press,; 2007.
[59] Chothia C. The nature of the accessible and buried surfaces in proteins. J Mol Biol 1976;105(1):1–12.
[60] Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229(4716):834–8.
[61] Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol 1985;55(3):836–9.