Critical review of conformational B-cell epitope prediction methods

Gabriel Cia [1,2], Fabrizio Pucci [1,2,†], Marianne Rooman [1,2,†]

[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles,
F. Roosevelt Avenue, 1050, Brussels, Belgium

[2] Interuniversity Institute of Bioinformatics in Brussels, Belgium

[†]Contributed equally to this work

Email: Gabriel.Cia@ulb.be, Fabrizio.Pucci@ulb.be, Marianne.Rooman@ulb.be

Abstract

Accurate *in-silico* prediction of conformational B-cell epitopes would lead to major improvements in disease diagnostics, drug design and vaccine development. A variety of computational methods, mainly based on machine learning approaches, have been developed in the last decades to tackle this challenging problem. Here, we rigorously benchmarked nine state-of-the-art conformational B-cell epitope prediction webservers, including generic and antibody-specific methods, on a dataset of over 250 antibody-antigen structures. The results of our assessment and statistical analyses show that all the methods achieve very low performances, and some do not perform better than randomly generated patches of surface residues. In addition, we also found that commonly used consensus strategies that combine the results from multiple webservers are at best only marginally better than random. Finally, we applied all the predictors to the SARS-CoV-2 spike protein as an independent case study, and showed that they perform poorly in general, which largely recapitulates our benchmarking conclusions. We hope that these results will lead to greater caution when using these tools until the biases and issues that limit current methods have been addressed, promote the use of state-of-the-art evaluation methodologies in future publications, and suggest new strategies to improve the performance of conformational B-cell epitope prediction methods.

Keywords: Conformational B-cell epitope prediction, Antibody-specific epitope prediction, Benchmarking, Immunoinformatics.

1 Introduction

The ever increasing amounts of biological data that are being generated and deposited in publicly accessible databases [1, 2] have boosted the development of machine learning (ML) models that are being used to help in advancing a variety of problems in the fields of genomics, proteomics and molecular evolution [3, 4]. The availability of 3-dimensional (3D) structural information from either experiments [5] or accurate prediction tools [6, 7] has further led to substantial improvements of *in-silico* prediction and modeling tools. One of the fields that has seen the development of a large number of structure-based ML models is B-cell epitope prediction. B-cell epitopes are typically protein surface regions which are bound by antibodies, and knowledge of the residues that form an epitope is key for unraveling disease mechanisms [8, 9] or for applications such as vaccine design, immunotherapy and immunoassay development [10].

Several experimental methods are available to determine B-cell epitopes [10], but they are expensive, time-consuming and some require a high level of lab expertise. This is why the development of *in-silico* tools has attracted a lot of attention. Initial methods focused on linear B-cell epitopes and relied on features derived from antigen sequences, but early on their predictive power was shown to be no better than random [11], a conclusion that was further confirmed in a recent study [12]. As more X-ray structures of antibody-antigen complexes were deposited in the Protein Data Bank (PDB) [5], a number of structure-based methods were developed to predict conformational or discontinuous B-cell epitopes, which contain residues that are not necessarily contiguous along the protein sequence. Many of these methods have reported significantly better than random predictive power [13] (see [14] for a historic presentation of B-cell epitope prediction methods).

Nevertheless, a number of voices [15, 16, 17, 18, 19, 20] have raised concerns regarding the feasibility of generic epitope predictions, *i.e.* predicting all the epitopes on a given antigen for all possible antibodies. Indeed, some evidence suggests that antibodies may be raised against virtually any part of the surface of any given protein [21, 22, 23], except in the case of chemical modifications such as glycosylation which are known to often block antibody binding [24, 25, 26]. The case of extensively studied proteins such as lysozyme, HIV-gp120 and, more recently, the receptor binding domain (RBD) of the SARS-CoV-2 spike protein show that it is possible to find epitopes on almost the entire surface of an antigen. If this were to be the general case, generic epitope prediction approaches would be futile.

As a result, a new trend has emerged in the field that challenges the generic epitope prediction paradigm and instead attempts to develop antibody-specific epitope predictors [15]. The main advantage of this approach is that it deals with a more constrained and tractable task as opposed to generic epitope prediction. Its downside is that it requires prior knowledge of the antibodies that need to be screened, which greatly limits the number of use cases compared to generic epitope prediction methods. Moreover, current antibody-specific epitope predictors are not fast enough to screen even a small fraction of the space of all possible antibodies, which is currently estimated at 10^{12} for naive antibodies and up to 10^{16} - 10^{18} for all possible antibodies [27].

To advance the field of B-cell epitope prediction, we have benchmarked and analyzed some of the most popular generic and antibody-specific B-cell epitope prediction methods by testing whether they are able to accurately identify experimentally validated epitopes.

This evaluation was performed on a dataset of over 250 non-redundant antibody-antigen structures using a rigorous benchmarking methodology.

2   Materials and methods

2.1   Surface residues

Residues were considered as part of the surface if they have a relative solvent accessible surface area (RSA) of at least 10%. The RSA of a residue X in a given protein structure, expressed in %, is defined as the sum of the accessible surface areas (ASA) of its heavy atoms divided by its maximal ASA reached when included in a Gly-X-Gly tripeptide in extended conformation. The ASA and RSA values were computed using an in-house program [28].

2.2   Epitope residues

Antigen surface residues residues that undergo a change in RSA of at least 5% upon binding with an antibody ($\Delta RSA = RSA_{unbound} - RSA_{bound} \geq 5\%$) were considered as epitope residues.

2.3   Structure datasets

The structures of complete antibodies (heavy and light chain) in complex with protein antigens were downloaded from the AntiBody DataBase [29] in PDB format. This represented 3,000 complexes at the date of 10/2020. Structures with a resolution greater than 3.0 Å, an R-factor greater than 0.30, or in which less than 80% of the residues have atomic coordinates were overlooked. Complexes in which the antigen has less than 50 residues were also dropped. This resulted in a quality filtered set of 1,151 antibody-antigen structures. The epitopes on the antigens were determined as described in the previous subsection. This set is referred to as $\mathcal{E}_{Ag}$.

In order to avoid redundancy and correctly evaluate the benchmarked generic B-cell epitope predictors, the antigens from the $\mathcal{E}_{Ag}$ set were clustered according to their sequence identity using CD-hit [30] with a 70% sequence identity threshold. This yielded 268 distinct antigen clusters. The representative antigen structure of each cluster was chosen to be the one identified by CD-hit. The epitope residues of all antigens in a given cluster were mapped onto the representative antigen structure by aligning their sequences using Biopython's local alignment algorithm [31] with the same default parameter settings as EMBOSS [32] (substitution matrix = BLOSUM62, open gap penalty = -10, extension gap penalty = -0.5); epitope residues were only mapped if the aligned residues were identical. The dataset of representative antigen structures with all epitopes mapped onto them is referred to as $\mathcal{E}_{Ag}^{rep}$.

The list of structures of the two datasets $\mathcal{E}_{Ag}$ and $\mathcal{E}_{Ag}^{rep}$ along with their PDB files are available at `https://github.com/3BioCompBio/BCellEpitope`.

2.4 Evaluation metrics

To estimate the prediction performance of the benchmarked predictors, we used a number of well established performance metrics [33], including the balanced accuracy (BAC), the Matthews correlation coefficient (MCC), the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC), defined as:

- $\text{BAC} = \dfrac{1}{2} \left(\dfrac{\text{TP}}{\text{TP} + \text{FN}} + \dfrac{\text{TN}}{\text{TN} + \text{FP}} \right)$

- $\text{MCC} = \dfrac{\text{TP TN} - \text{FP FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$

- ROC-AUC, *i.e.* the area under the curve (AUC) of the recall or sensitivity (TP/(TP+FN)) versus the false positive rate or specificity (FP/(FP+TN)).

- PR-AUC, *i.e.* the AUC of the positive predictive value (PPV) or precision (TP/(TP+FP)) versus the recall or sensitivity (TP/(TP+FN)).

where TP are correctly predicted epitope residues, FP non-epitope residues incorrectly predicted as epitope residues, TN correctly predicted non-epitope residues, and FN epitope residues incorrectly predicted as non-epitope residues. The mean random value is equal to 0.5 for BAC and ROC-AUC, and 0 for MCC; for PR-AUC it is dataset-dependent.

2.5 Random epitope prediction procedure

In order to assess the statistical significance of the different methods against random predictions, we defined two procedures, one that predicts random surface residues, and a second that predicts random patches of surface residues, *i.e* groups of residues that are nearby in the 3D structure. Each procedure is repeated 2,000 times in order to generate bootstrap distributions that are then used to calculate p-values.

The first procedure randomly predicts N_r random surface residues as epitopes on each antigen structure. We tested two strategies for setting the value of N_r: (1) $N_r = 18$, which corresponds to the average epitope size in our $\mathcal{E}_{Ag}$ dataset; (2) N_r chosen dynamically to match the number of residues predicted by each method for each structure. The latter strategy allowed us to assess how our random procedure compares to each method in the case of equivalent prediction thresholds, and led to method-specific bootstrap distributions for each metric (MCC, BAC, ROC-AUC, PR-AUC).

The above procedure is overly simplistic as epitope residues are not randomly scattered across the protein surface, but rather form patches of nearby surface residues. We therefore developed a second procedure that predicts N_p random patches of N_r surface residues each. The patches were constructed by randomly selecting a surface residue and adding its N_r - 1 closest surface residues. Here we also used two strategies to set the values of N_p and N_r: (1) $N_p = 1$ and $N_r = 18$; (2) dynamical number of epitope residues N matching the number of predicted residues on a per-method and per-structure basis, distributed in N_p

patches of N_r residues as: $N_p = \lfloor N/18 \rfloor$ and $N_r = 18$ for all patches but one for which $N_r = N - (N_p - 1) * 18$.

Method	ROC-AUC	BAC	MCC	PR-AUC	$N_{antigens}$	$F_{predicted}$
Sequence-based generic methods						
BepiPred2	0.53	0.52	0.02	0.24	101/268	52 %
CBTOPE	0.46	0.47	-0.06	0.19	229/268	38 %
Structure-based generic methods						
SEPPA3	0.53	0.51	0.00	0.23	105/268	47 %
DiscoTope2	**0.58**	**0.53**	**0.06**	0.26	220/268	19 %
ElliPro	0.56	**0.53**	0.04	0.23	259/268	63 %
EPSVR	0.53	0.52	0.03	0.26	236/268	52 %
BEpro	**0.58**	**0.53**	**0.06**	0.27	235/268	12 %
epitope3D	0.41	0.41	-0.17	0.09	221/268	3 %
Structure-based antibody-specific methods						
EpiPred	0.50	0.50	-0.01	**0.35**	746/1151	49 %

Table 1: Average performance of each of the benchmarked conformational B-cell epitope predictors. The highest score(s) of each metric are in bold. $N_{antigens}$ corresponds to the number of structures on which each method has been evaluated out of the total number of eligible antigens. $F_{predicted}$ is the mean fraction (in %) of surface residues predicted as epitopes. Note that the reason why DiscoTope2 and BEpro have a low $F_{predicted}$ comes from the fact that these methods use very high prediction thresholds.

3 Results and discussion

3.1 Epitope dataset analysis

We used two datasets: $\mathcal{E}_{Ag}$ that contains 1,151 good-quality antigen structures, each carrying a single epitope, and the non-redundant and non-homologous $\mathcal{E}_{Ag}^{rep}$ dataset, which contains 268 representative antigen structures onto which we mapped all the epitopes identified on homologous structures in $\mathcal{E}_{Ag}$ (see Methods for details). Mapping multiple known epitopes onto a single antigen structure prevents as much as possible erroneous false positive annotations that would arise if different epitopes from the same antigen were evaluated independently from each other [34].

The number of mapped epitopes per antigen structure in $\mathcal{E}_{Ag}^{rep}$ follows a decreasing exponential-type distribution in which 85% of the structures have less than 5 mapped structures (see Supplementary Figure S1). For some extensively studied antigens such as lysozyme and HIV-gp-120, this number increases to more than 35. In the case of lysozyme, the epitopes cover almost the entire surface: 70 out of 85 surface residues belong to at least one of

the many lysozyme epitopes found in our $\mathcal{E}_{Ag}$ dataset. The generality of this observation is currently an open question; we will come back to it in the discussion section.

3.2   Benchmarking methodology

We assessed conformational B-cell epitope prediction methods with a functioning webserver. The list of methods that were selected includes two generic sequence-based methods: Bepipred-2.0 [35] and CBTOPE [36]; six generic structure-based methods: SEPPA3 [37], DiscoTope2 [34], ElliPro [38], EPSVR [39], BEpro [40] and epitope3D [41]; and one antibody-specific structure-based predictor: EpiPred [42].

For evaluating each of the generic epitope prediction tools, we used the subset of the $\mathcal{E}_{Ag}^{rep}$ dataset that is not contained in the training dataset of the method considered. More precisely, we removed from $\mathcal{E}_{Ag}^{rep}$ any antigen that has a sequence identity of more than 99% with any antigen in the training dataset of the method that is being assessed. Note that using a lower sequence identity threshold of 70% has virtually no effect on the score values reported in Table 1, as seen in Table S1. For assessing the antibody-specific epitope prediction tool EpiPred, we used the $\mathcal{E}_{Ag}$ set from which we removed all the PDB structures that are included in the method's training set. Although this procedure means that all the methods were assessed on different test sets, it avoids biases due to evaluating training data.

Furthermore, we only considered the predictions made for surface residues (as defined in Methods) in the assessment, as core residues can not be part of B-cell epitopes. Note that the prediction scores would be much better if both surface and core residues were considered. This does not make sense for structure-based predictors and basically boosts sequence-based predictor performance given that the identification of surface residues is an easier problem than epitope prediction.

We used the threshold-independent metric MCC and BAC for predictors evaluation as they account for all the categories of the confusion matrix. In addition, we also used ROC-AUC and PR-AUC as these performance metrics are independent of any classification threshold value and thus give complementary information.

3.3   Benchmarking results

We report the average BAC, MCC, ROC-AUC and PR-AUC scores of the benchmarked methods in Table 1 and their statistical significance against different random procedures in Table S2a. Additional metrics for evaluating the methods, including sensitivity, specificity, precision and F1 score, are available in Table S3. What clearly comes out is that all the methods have very low performances, as indicated by ROC-AUC and BAC values < 0.6 and MCC values < 0.1. BEpro and DiscoTope2 are the highest scoring methods, both achieving identical metrics (ROC-AUC=0.58, BAC=0.53, MCC=0.06). In contrast, the scores of epitope3D are even worse than random (ROC-AUC=0.41, BAC=0.41, MCC=-0.17), because almost all its predicted epitope residues are situated in the protein core. The first conclusion we can draw from these results is that even the highest scoring methods have very little predictive power.

Surprisingly, the antibody-specific epitope predictor, EpiPred, does not show better overall performance than the best generic epitope predictors in terms of ROC-AUC, BAC and
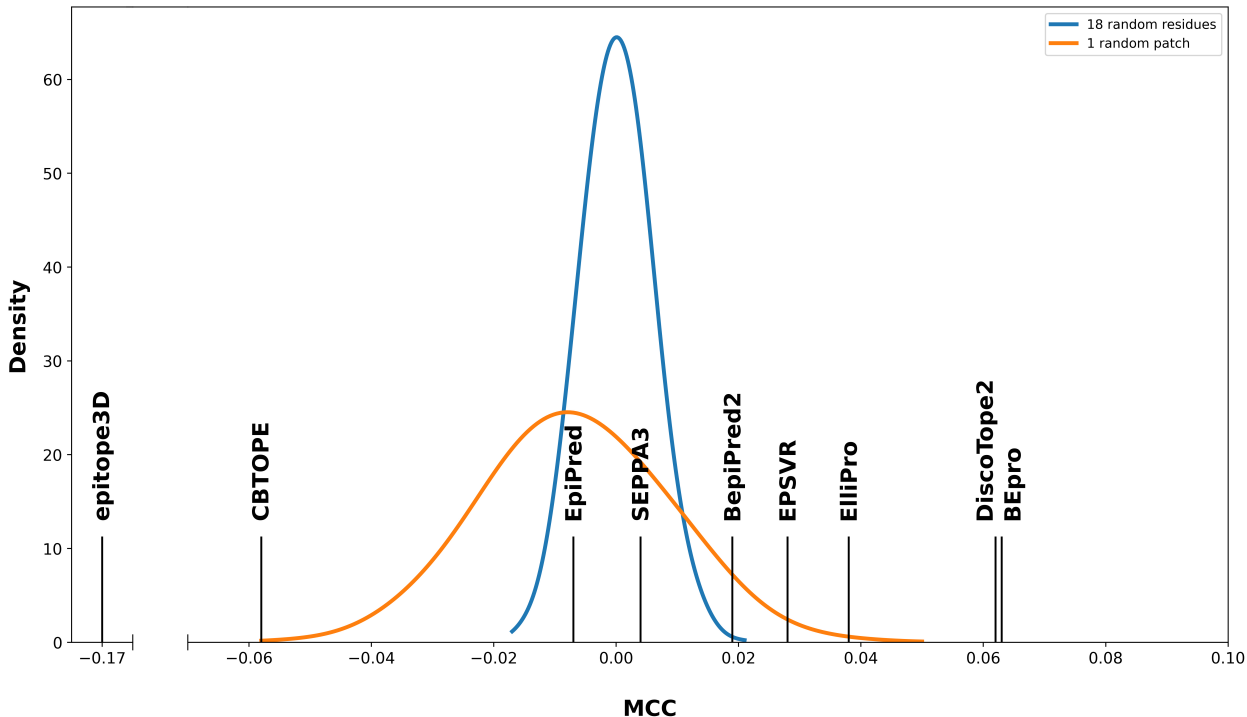
Figure 1: Matthews correlation coefficient (MCC) of the benchmarked conformational B-cell epitope prediction methods, along with the bootstrap distributions obtained from a random procedure that generates 18 random surface residues (blue) or 1 random patch of 18 nearby surface residues (orange) on each structure of the $\mathcal{E}_{Ag}^{rep}$ dataset, repeated 2,000 times.

MCC. However, it does have the highest PR-AUC score, which indicates that the knowledge of the antibody improves the precision of the predicted epitopes.

Note that all these results are independent of the chosen definition of epitopes. Indeed, comparing Table 1 with Table S4, we observe that the scores are almost identical whether using an RSA- or distance-based definition of epitope residues, the latter being another common definition used by many of the benchmarked methods. Moreover, we analyzed how the definition of surface residues influences the methods' scores. For that purpose, we computed the MCC scores as a function of the RSA thresholds used for defining surface residues, as shown in Figure S2. We see that all the prediction methods have systematically worse scores as the threshold increases, which indicates that the more we restrict the evaluation to residues that are truly at the surface, the worse the methods perform. Conversely, considering buried residues as belonging to the surface makes the predictions easier, because it artificially increases the difference between epitopes and non-epitopes by basically enriching the latter with hydrophobic residues much more than the former that are defined by an additional threshold on ΔRSA (see Methods).

To determine whether these results are statistically significantly better than random, we benchmarked the methods against a procedure which randomly predicts N_r surface residues as epitopes. We tested two strategies for the value of N_r: the first considers N_r equal to the average size of an epitope, and the second one uses a dynamic N_r that matches the number of epitope residues predicted by each method for each antigen (see Methods for

details). As shown in Table S2a, when benchmarked against these strategies, only Bepipred2, DiscoTope2, ElliPro, EPSVR and BEpro are significantly better than our random procedure across all four metrics. SEPPA3 and EpiPred are significantly better only for some metrics, while CBTOPE and epitope3D are no better than random across all metrics.

We also benchmarked the methods against a random procedure that predicts patches of surface residues as epitopes instead of random residues scattered across the antigen surface (see Methods). This time only DiscoTope2, EPSVR and BEpro are significantly better than our random patches procedure across all metrics (Table S2a). BepiPred2, SEPPA3, ElliPro and EpiPred are significantly better for some of the metrics, while CBTOPE and epitope3D are no better than random patches across all metrics. Note that the reason behind the differences between the residue- and patch-based bootstrap distributions comes from the fact that the patch-based random procedure has a higher standard deviation than its residue-based counterpart (see Figure 1). Indeed, predicting patches of residues leads to a higher possibility of predicting either many correct or incorrect residues than when predicting randomly scattered residues, given true epitopes are themselves patches of nearby residues.

One of the reasons that could explain why the majority of the methods did not perform better than random is the presence of false negative annotations in the dataset, corresponding to epitopes not yet identified. In order to improve the confidence in the epitope/non-epitope annotation of surface residues, we repeated the above benchmarking on the subset of $\mathcal{E}_{Ag}^{rep}$ consisting of antigens bound by at least 5 epitopes, noted $\mathcal{E}_{Ag}^{rep(5)}$. EpiPred was excluded from this analysis given it is not affected by the issue of erroneous false negatives. The results are reported in Figure S3 and Table S2b, which shows similar results than the previous evaluation on the full benchmark set, with BepiPred2, DiscoTope2 and BEpro performing better than random across all metrics.

In conclusion, all the methods showed very poor performances in absolute terms, and only two methods, namely DiscoTope2 and BEpro, achieved better than random performances across all metrics and benchmarks.

3.4 Consensus predictions

Often, conformational B-cell epitopes are predicted using a combination of several methods and a consensus scheme whereby a residue is considered as an epitope if at least M methods predict it as such (see [43, 44, 45, 46] for recent examples). We therefore tested whether combining the predictions of all the generic epitope prediction methods gave better results than each one individually.

We first removed the structures that were in any of the selected methods' training datasets, resulting in a dataset of 65 structures on which the consensus predictions were evaluated. We predicted a residue as an epitope if at least M of the selected methods agreed. This consensus strategy gave the highest results for $M = 4$, resulting in a ROC-AUC = 0.56, BAC = 0.56, MCC = 0.10 and PR-AUC = 0.34, which is slightly better than any individual predictor. Nonetheless, one should keep in mind that this result was obtained through optimization of the M value in direct validation and is thus probably a bit overestimated.

Method	$r_{\mathcal{I}-score}$	$N_{residues}$
BepiPred2	**0.20***	3491
CBTOPE	0.08*	5799
SEPPA3	0.15*	3107
DiscoTope2	0.15*	6287
ElliPro	0.04	6910
EPSVR	0.09*	6490
BEpro	0.14*	6287
epitope3D	-0.03	4795

Table 2: Spearman correlation coefficient $r_{\mathcal{I}-score}$ between the estimated immunodominance $\mathcal{I}$ and the per-residue epitope score outputted by each method on the $\mathcal{E}_{Ag}^{rep(5)}$ set. $N_{residues}$ corresponds to the number of residues on which the correlation was calculated. The highest correlation value is in bold. Due to the large sample size, we considered correlations as statistically significant if their p-value is ≤ 0.001, which are labeled with an asterisk.

In summary, even though consensus prediction schemes might be slightly better than single-method approaches or randomly predicted residues, they do not overcome the fundamental issue of the under-performance of B-cell epitope prediction methods.

3.5   Epitope immunodominance

The question of whether antibodies can be raised against any part of any antigen's surface has currently no definitive answer, but the poor performance of all the methods evaluated in the previous sections suggest a positive answer to this question. Even if the entire surface of any antigen can be bound by antibodies, some regions are undoubtedly targeted much more often than others by the immune system and more easily trigger the immune response. This phenomenon, known as epitope immunodominance [47, 48], is important, for example when designing epitope-based vaccines that attempt to generate an immune response towards subdominant but functionally conserved sites where escaping mutations are less likely to occur [49, 50, 51]. Knowledge of the immunodominant and subdominant epitopes of an antigen can therefore be of great value.

One can reasonably expect that the scores attributed to each surface residue by conformational B-cell epitope predictors are correlated, at least to some extent, with residue immunodominance. To test this hypothesis, we estimated the immunodominance $\mathcal{I}$ of a given residue as the number of times it appears in an epitope; for this purpose, we restricted ourselves to the $\mathcal{E}_{Ag}^{rep(5)}$ dataset which only contains antigen structures that are bound by at least 5 different antibodies. As for the benchmarking, antigen structures that were part of the training dataset of a given method were removed. $\mathcal{I}$ was min-max scaled on a per-antigen basis to adjust for differences in number of epitopes per antigen structure.

We computed the Spearman correlation between $\mathcal{I}$ and the predicted scores of each method, with the exception of EpiPred given immunodominance is irrelevant for antibody-specific methods.

As shown in Table 2, six out of the eight evaluated methods have a statistically significant Spearman correlation and are therefore better than random. However, the correlation values are very low, the highest being 0.20 for BepiPred2; note that this is a sequence-based predictor. These results are in accordance with the low prediction scores observed in the previous subsection and indicate that the scores of the methods are not accurate enough to be used to deduce which epitopes are immunodominant.

Note that immunodominance is a highly complex phenomenon and that the $\mathcal{I}$ value that we used to estimate it, though intuitive, is clearly an approximation. Indeed, our dataset is biased towards most (e.g., clinically) promising antibodies, and moreover contains highly engineered antibodies which do not necessarily reflect the preferences of the immune system. In addition, $\mathcal{I}$ does not account for natural biases of the immune system such as antigenic imprinting [52] or original antigenic suppression [53], where the immune system preferentially uses or avoids the immunological memory based on previous infections.

3.6   SARS-CoV-2 case study

As a case study, we evaluated the B-cell epitope predictors on the spike protein of the SARS-CoV-2 virus. This protein enables cell invasion through binding to the host's ACE2 receptor [54], and its receptor binding domain has been shown to be the preferential target of the host's immune response [55]. The SARS-CoV-2 spike protein is included neither in our $\mathcal{E}_{Ag}$ benchmark dataset nor in the training sets of the methods, and therefore constitutes an independent test case. Note that another series of epitope prediction tools applied to SARS-CoV-2 has been reviewed in [56].

In order to evaluate the ability of the benchmarked methods to predict the known epitopes in the spike protein, we gathered 83 spike protein-antibody complexes resolved by X-ray crystallography or cryo-electron microscopy and deposited in the PDB [5] (see [57, 58] for the list of PDB ids). We extracted 83 epitopes from these complexes; 75 of them are localized on the receptor binding domain (RBD) of the spike protein while the remaining 8 target its N-terminal domain (NTD). We evaluated the methods by mapping the 83 epitopes on the PDB structure 6VYB, which is a complete trimeric spike protein structure with one chain in open conformation [59]. Note that we do not have the results for all the predictors evaluated in the previous subsection as some of them failed to run on the large protein trimer or their webserver was down.

The prediction scores of the methods are given in Table 3 and the localization of the predicted and real epitopes in the 3D structure of the spike protein trimer are shown in Figure 2. Some of the predictors have relatively good scores, especially when compared with the performances on the large benchmark dataset analyzed in the previous subsection. EPSVR obtained the best results, reaching a ROC-AUC of 0.75 and a score-immunodominance correlation of 0.45. This can clearly be seen in Figure 2 where EPSVR predicts very well a large portion of the RBD and NTD epitope residues. The worse performing method is again epitope3D which, as previously observed, is biased towards non-epitope core residues. The remaining methods did not perform too well, as they either overpredicted (ElliPro), underpredicted (DiscoTope2) or made predictions all over the surface (BepiPred2 and CBTOPE).

Method	MCC	ROC-AUC	$r_{\mathcal{I}-score}$
BepiPred2	0.19	0.64*	0.30*
CBTOPE	0.02	0.51	0.17*
DiscoTope2	**0.27***	0.60	0.24*
ElliPro	0.17	0.66*	0.35*
EPSVR	0.22	**0.75***	**0.45***
epitope3D	-0.20	0.38	0.039

Table 3: MCC, ROC-AUC and Spearman correlation coefficient $r_{\mathcal{I}-score}$ between immunodominance $\mathcal{I}$ and prediction scores of each method on the SARS-CoV-2 spike protein trimer. Statistically significant results (p-value ≤ 0.001) are labeled with an asterisk. The highest value for each metric is in bold.

All epitope residues in the spike protein trimer

BepiPred2   CBTOPE   DiscoTope2
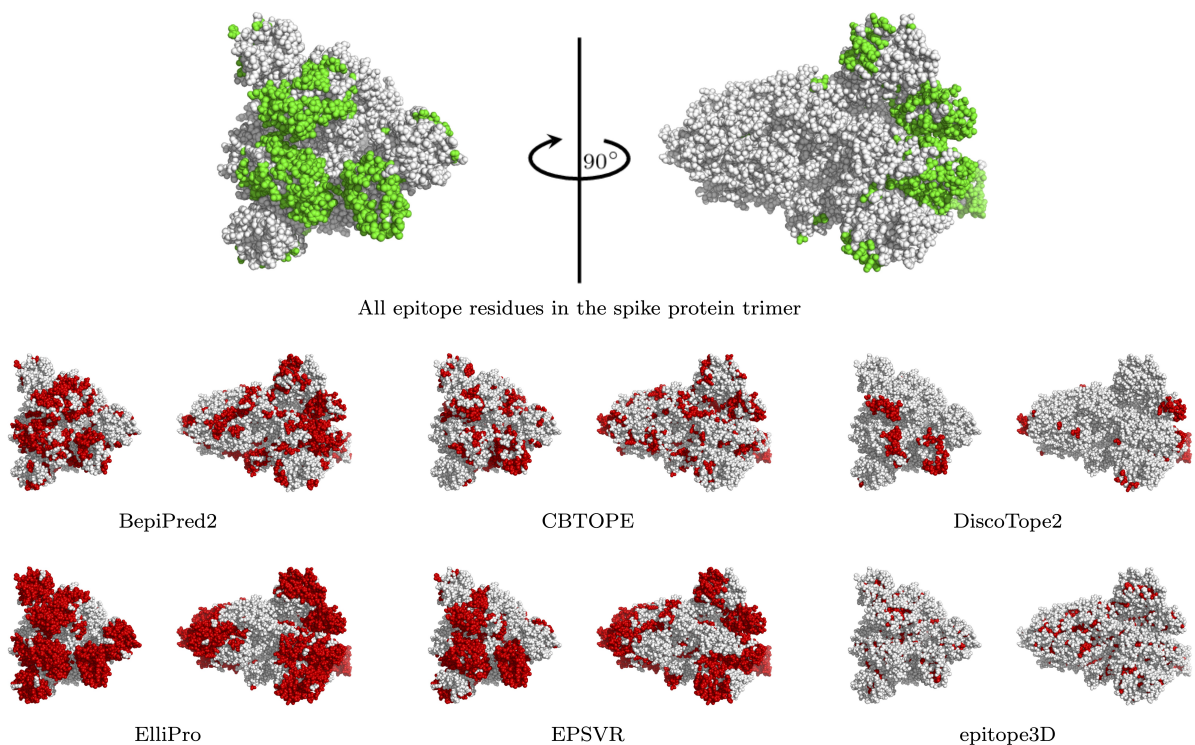
ElliPro   EPSVR   epitope3D

Figure 2: Visual representation of all the true epitope residues of the SARS-CoV-2 spike protein trimer (green) and the residues predicted by each of the evaluated conformational B-cell epitope prediction methods (red). All the figures were generated with PyMOL [60]

3.7 Per-antigen performance analysis

The fact that the overall SARS-CoV-2 results are better than those on the large benchmark is interesting and prompted us to dig further into our results. We analyzed the methods' performances on a per-antigen basis by plotting the distribution of MCC and Spearman correlation coefficients of each method; they are shown in Figures S4 and S5. Both figures show that all the methods have very variable performances according to the antigen, in other

11

words, they have a high standard deviation. Indeed, some entries are well predicted with high scores, while others are completely wrongly predicted.

It could be interesting to further analyze such results to understand whether the well predicted antigens are due to randomness, to biases towards the method's training dataset, or to truly learned features that distinguish epitope and non-epitope residues. Understanding this could boost the development of improved predictors and help to better understand their reliability.

Note that, despite the fact that predictions on the SARS-CoV-2 spike protein show better results on the average, B-cell epitope predictors did not contribute much to antibody development during the pandemic. Indeed, experimental characterization of antibodies starting from plasma of infected COVID-19 patients followed by 3D cryo-electron microscopy has been the method of choice to design therapeutic monoclonal antibodies [61].

4 Conclusion

Many *in-silico* tools to predict conformational B-cell epitopes using sequence- and/or structure-derived features have been published in the last 20 years. They have all compared themselves to each other and almost systematically claim to outperform all the other methods. However, no independent benchmarking has been published in recent years [62]. In this paper we have carefully assessed nine of the most popular and recent prediction methods available through a webserver, including eight generic predictors and an antibody-specific predictor, on a large and well-curated set of antigen-antibody structures.

Our benchmarking results show that the overall performance of the methods is very poor, and that many of them do not perform significantly better than randomly predicted patches of residues. Indeed, only 2 out of the 9 evaluated methods perform significantly better than random both on the benchmark dataset and on the subset of antigens for which we have the highest confidence about epitope/non-epitope residue labels. Note that some of the tested methods were trained over 10 years ago and their performance might have been better if they had been retrained on an up-to-date training dataset.

In addition, we evaluated the performance of consensus strategies that combine multiple predictors, which is a frequently used approach in B-cell epitope prediction. Our results show that even this strategy is not much better than random predictions.

Regarding the evaluation methodology used in this benchmark, we combined both threshold-dependent (BAC and MCC) and threshold-independent (ROC-AUC and PR-AUC) metrics in order to capture the overall performance of the methods. They all point towards the same conclusion regarding the predictive performance of the methods, which is much lower than what we expected.

Our benchmarking also provides hints about open questions and improvements that could be made to current prediction methods. First of all, our analysis highlights the pervasive issue of the incomplete knowledge of all the true epitopes of any given protein, especially in lesser studied ones. This has a major impact on evaluation procedures given there is always the uncertainty that a non-epitope residue might be part of a yet unknown epitope. For that reason, we performed our benchmark analysis on both the full dataset of representative structures $\mathcal{E}_{Ag}^{rep}$ and on the subset of antigens with at least 5 mapped epitopes, in order

to assess if there is a difference in the methods' performance when evaluated on antigen structures for which we have a higher confidence in the labels of their surface residues, but our results showed no significant difference between these two benchmarks.

Related to this, the question arises of whether an antibody can be obtained against any surface region or if only certain parts with favourable structural and physico-chemical features are valid candidates. In the former case, the strong immunodominance observed in nature could originate from biases specific to the naive immune system of each species. Although our benchmarking analysis cannot answer this question with certainty, the systematically low performance of all tested methods suggests that it is indeed the case. However, some regions that are more easily recognized by antibodies, known as immunodominant epitopes, have been shown to elicit a stronger immune response [63] and it should thus be possible to identify their characteristic features. A subsidiary question is whether natural and engineered antibodies are different in that respect.

Another interesting observation from our results is that each method predicts some proteins quite well, with high prediction scores, and others very poorly. The analysis of these outliers could be helpful to understand the advantages and limitations of each method and help design better performing methods. It must be noted that the predictors do not agree in general: a protein that is well predicted by one method is usually not well predicted by the others.

Regarding the features used by the different methods, we found that the large majority of predictors overlook some important features whose consideration would certainly boost their performances:

- Glycosylated antigen regions usually cannot be recognized by antibodies due their shielding effect [24, 25, 26]. The annotation or prediction of glycosylated regions should thus be included in the predictors to boost their performances.

- Antigens can undergo conformational changes where some regions get masked and become unavailable for antibody binding. The spike protein trimer of SARS-CoV-2 is an example of this: it occurs in open and closed conformations characterized by differences in solvent accessibility and epitopes [57].

- Oligomerization properties of antigens are also important for determining their immunogenicity. Indeed, oligomers hide regions from the solvent which are then no longer accessible to antibodies and, at the same time, lead to inter-chain surface regions that can be targeted by antibodies. Prediction methods often do not take these properties into account.

- It would be beneficial for webservers to give users the ability to provide additional information about the target protein, such as residues that cannot be bound and should therefore be ignored by the predictor.

- Another interesting improvement would be the possibility for the user to provide antibody sequences for which he wishes to identify potential epitopes, as this has been suggested to improve prediction performances [17].

- Finally, the application of recent advances in Natural Language Processing (NLP) could enable the development of novel methods [64, 65] that help advance the field.

We hope this work will be of use for future research in B-cell epitope prediction and help solve some of the critical issues. It is important to set up additional independent benchmarks as well as blind prediction experiments as they would contribute to a better understanding of the biases and limitations of epitope prediction methods and advance the field.

5 Competing interests

There is NO Competing Interest.

6 Acknowledgments

7 Author contributions statement

Conceptualization, F.P. and M.R.; G.C. conducted the experiment(s). All author analysed the results, wrote and reviewed the manuscript, and agreed to the published version of the manuscript.

References

[1] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[2] Chuming Chen, Hongzhan Huang, and Cathy H Wu. Protein bioinformatics databases and resources. *Protein Bioinformatics*, pages 3–39, 2017.

[3] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, José A Lozano, Rubén Armananzas, Guzmán Santafé, Aritz Pérez, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

[4] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

[5] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[6] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1):W296–W303, 2018.

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[8] Anne Davidson and Betty Diamond. Autoimmune diseases. *New England Journal of Medicine*, 345(5):340–350, 2001.

[9] Jan A Burger and Nicholas Chiorazzi. B cell receptor signaling in chronic lymphocytic leukemia. *Trends in immunology*, 34(12):592–601, 2013.

[10] Julia V Ponomarenko and Marc HV Van Regenmortel. B cell epitope prediction. *Structural bioinformatics*, 2:849–879, 2009.

[11] Martin J Blythe and Darren R Flower. Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Science*, 14(1):246–248, 2005.

[12] Kosmas A Galanis, Katerina C Nastou, Nikos C Papandreou, Georgios N Petichakis, Diomidis G Pigis, and Vassiliki A Iconomidou. Linear B-cell epitope prediction for in silico vaccine design: A performance review of methods available via command-line interface. *International journal of molecular sciences*, 22(6):3210, 2021.

[13] Jason A Greenbaum, Pernille Haste Andersen, Martin Blythe, Huynh-Hoa Bui, Raul E Cachau, James Crowe, Matthew Davies, AS Kolaskar, Ole Lund, Sherrie Morrison, et al. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *Journal of Molecular Recognition: An Interdisciplinary Journal*, 20(2):75–82, 2007.

[14] Jose L Sanchez-Trincado, Marta Gomez-Perosanz, and Pedro A Reche. Fundamentals and methods for T-and B-cell epitope prediction. *Journal of immunology research*, 2017, 2017.

[15] Inbal Sela-Culang, Yanay Ofran, and Bjoern Peters. Antibody specific epitope prediction—emergence of a new paradigm. *Current opinion in virology*, 11:98–102, 2015.

[16] Marc HV Van Regenmortel. Specificity, polyspecificity and heterospecificity of antibody-antigen recognition. *HIV/AIDS: Immunochemistry, Reductionism and Vaccine Design*, pages 39–56, 2019.

[17] Martin Closter Jespersen, Swapnil Mahajan, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. Antibody specific B-cell epitope predictions: leveraging information from antibody-antigen protein complexes. *Frontiers in immunology*, 10:298, 2019.

[18] Inbal Sela-Culang, Shaul Ashkenazi, Bjoern Peters, and Yanay Ofran. PEASE: predicting B-cell epitopes utilizing antibody sequence. *Bioinformatics*, 31(8):1313–1315, 2015.

[19] Inbal Sela-Culang, Vered Kunik, and Yanay Ofran. The structural basis of antibody-antigen recognition. *Frontiers in immunology*, 4:302, 2013.

[20] Liang Zhao, Limsoon Wong, and Jinyan Li. Antibody-specified B-cell epitope prediction in line with the principle of context-awareness. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1483–1494, 2011.

[21] David C Benjamin, Jay A Berzofsky, Iain J East, Frank RN Gurd, Charles Hannum, Sydney J Leach, Emanuel Margoliash, J Gabriel Michael, Alexander Miller, Ellen M Prager, et al. The antigenic structure of proteins: a reappraisal. *Annual review of immunology*, 2(1):67–101, 1984.

[22] Jens Vindahl Kringelum, Morten Nielsen, Søren Berg Padkjær, and Ole Lund. Structural analysis of B-cell epitopes in antibody: protein complexes. *Molecular immunology*, 53(1-2):24–34, 2013.

[23] Vered Kunik and Yanay Ofran. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Engineering, Design & Selection*, 26(10):599–609, 2013.

[24] Deborah Chang and Joseph Zaia. Why glycosylation matters in building a better flu vaccine. *Molecular & Cellular Proteomics*, 18(12):2348–2358, 2019.

[25] Lorenzo Casalino, Zied Gaieb, Jory A Goldsmith, Christy K Hjorth, Abigail C Dommer, Aoife M Harbison, Carl A Fogarty, Emilia P Barros, Bryn C Taylor, Jason S McLellan, et al. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS central science*, 6(10):1722–1734, 2020.

[26] René Wintjens, Amanda Makha Bifani, and Pablo Bifani. Impact of glycan cloud on the B-cell epitope prediction of SARS-CoV-2 Spike protein. *npj Vaccines*, 5(1):1–8, 2020.

[27] Bryan Briney, Anne Inderbitzin, Collin Joyce, and Dennis R Burton. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, 566(7744):393–397, 2019.

[28] Georgios A Dalkas, Fabian Teheux, Jean Marc Kwasigroch, and Marianne Rooman. Cation–π, amino–π, π–π, and H-bond interactions stabilize antigen–antibody interfaces. *Proteins: Structure, Function, and Bioinformatics*, 82(9):1734–1746, 2014.

[29] Saba Ferdous and Andrew CR Martin. AbDb: antibody structure database—a database of PDB-derived antibody structures. *Database*, 2018, 2018.

[30] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[31] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[32] Fábio Madeira, Matt Pearce, Adrian RN Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo Lopez. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic acids research*, 50(W1):W276–W279, 2022.

[33] Yasen Jiao and Pufeng Du. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, 4(4):320–330, 2016.

[34] Jens Vindahl Kringelum, Claus Lundegaard, Ole Lund, and Morten Nielsen. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS computational biology*, 8(12):e1002829, 2012.

[35] Martin Closter Jespersen, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic acids research*, 45(W1):W24–W29, 2017.

[36] Hifzur Rahman Ansari and Gajendra PS Raghava. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome research*, 6(1):1–9, 2010.

[37] Chen Zhou, Zikun Chen, Lu Zhang, Deyu Yan, Tiantian Mao, Kailin Tang, Tianyi Qiu, and Zhiwei Cao. SEPPA 3.0—enhanced spatial epitope prediction enabling glycoprotein antigens. *Nucleic acids research*, 47(W1):W388–W394, 2019.

[38] Julia Ponomarenko, Huynh-Hoa Bui, Wei Li, Nicholas Fusseder, Philip E Bourne, Alessandro Sette, and Bjoern Peters. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC bioinformatics*, 9(1):1–8, 2008.

[39] Shide Liang, Dandan Zheng, Daron M Standley, Bo Yao, Martin Zacharias, and Chi Zhang. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. *BMC bioinformatics*, 11(1):1–6, 2010.

[40] Michael J Sweredoski and Pierre Baldi. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics*, 24(12):1459–1460, 2008.

[41] Bruna Moreira da Silva, YooChan Myung, David B Ascher, and Douglas EV Pires. epitope3D: a machine learning method for conformational B-cell epitope prediction. *Briefings in Bioinformatics*, 23(1):bbab423, 2022.

[42] Konrad Krawczyk, Xiaofeng Liu, Terry Baker, Jiye Shi, and Charlotte M Deane. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, 30(16):2288–2294, 2014.

17

[43] Yengkhom Damayanti Devi, Arpita Devi, Hemanga Gogoi, Bondita Dehingia, Robin Doley, Alak Kumar Buragohain, Ch Shyamsunder Singh, Partha Pratim Borah, C Durga Rao, Pratima Ray, et al. Exploring rotavirus proteome to identify potential B-and T-cell epitope using computational immunoinformatics. *Heliyon*, 6(12):e05760, 2020.

[44] Fisayo A Olotu and Mahmoud ES Soliman. Immunoinformatics prediction of potential B-cell and T-cell epitopes as effective vaccine candidates for eliciting immunogenic responses against Epstein–Barr virus. *biomedical journal*, 44(3):317–337, 2021.

[45] Jerome Rumdon Lon, Yunmeng Bai, Bingxu Zhong, Fuqiang Cai, and Hongli Du. Prediction and evolution of B cell epitopes of surface protein in SARS-CoV-2. *Virology journal*, 17(1):1–9, 2020.

[46] Sangeeta Khare, Marli Azevedo, Pravin Parajuli, and Kuppan Gokulan. Conformational changes of the receptor binding domain of sars-cov-2 spike protein and prediction of a b-cell antigenic epitope using structural data. *Frontiers in Artificial Intelligence*, 4:630955, 2021.

[47] Manpreet Kaur, Hema Chug, Harpreet Singh, Subhash Chandra, Manish Mishra, Meenakshi Sharma, and Rakesh Bhatnagar. Identification and characterization of immunodominant B-cell epitope of the C-terminus of protective antigen of Bacillus anthracis. *Molecular immunology*, 46(10):2107–2115, 2009.

[48] Hiro-O Ito, Toshihiro Nakashima, Takanori So, Masato Hirata, and Masakazu Inoue. Immunodominance of conformation-dependent B-cell epitopes of protein antigens. *Biochemical and biophysical research communications*, 308(4):770–776, 2003.

[49] Catharine I Paules, Hilary D Marston, Robert W Eisinger, David Baltimore, and Anthony S Fauci. The pathway to a universal influenza vaccine. *Immunity*, 47(4):599–603, 2017.

[50] Davide Angeletti and Jonathan W Yewdell. Is it possible to develop a "universal" influenza virus vaccine? Outflanking antibody immunodominance on the road to universal influenza vaccination. *Cold Spring Harbor perspectives in biology*, 10(7):a028852, 2018.

[51] Seth J Zost, Nicholas C Wu, Scott E Hensley, and Ian A Wilson. Immunodominance and antigenic variation of influenza virus hemagglutinin: implications for design of universal vaccine immunogens. *The Journal of infectious diseases*, 219(Supplement_1):S38–S45, 2019.

[52] Anup Vatti, Diana M Monsalve, Yovana Pacheco, Christopher Chang, Juan-Manuel Anaya, and M Eric Gershwin. Original antigenic sin: a comprehensive review. *Journal of autoimmunity*, 83:12–21, 2017.

[53] Davide Angeletti, James S Gibbs, Matthew Angel, Ivan Kosik, Heather D Hickman, Gregory M Frank, Suman R Das, Adam K Wheatley, Madhu Prabhakaran, David J Leggat, et al. Defining B cell immunodominance to viruses. *Nature immunology*, 18(4):456–463, 2017.

[54] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *cell*, 181(2):271–280, 2020.

[55] Nathan Post, Danielle Eddy, Catherine Huntley, May CI Van Schalkwyk, Madhumita Shrotri, David Leeman, Samuel Rigby, Sarah V Williams, William H Bermingham, Paul Kellam, et al. Antibody response to SARS-CoV-2 infection in humans: A systematic review. *PloS one*, 15(12):e0244126, 2020.

[56] Syed Nisar Hussain Bukhari, Amit Jain, Ehtishamul Haq, Abolfazl Mehbodniya, and Julian Webber. 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*, 11(2):146, 2022.

[57] Gabriel Cia, Fabrizio Pucci, and Marianne Rooman. Analysis of the Neutralizing Activity of Antibodies Targeting Open or Closed SARS-CoV-2 Spike Protein Conformations. *International journal of molecular sciences*, 23(4):2078, 2022.

[58] Gabriel Cia, Jean Marc Kwasigroch, Marianne Rooman, and Fabrizio Pucci. SpikePro: a webserver to predict the fitness of SARS-CoV-2 variants. *Bioinformatics*, 2022.

[59] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181(2):281–292, 2020.

[60] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, 40(1):82–92, 2002.

[61] Peter C Taylor, Andrew C Adams, Matthew M Hufford, Inmaculada De La Torre, Kevin Winthrop, and Robert L Gottlieb. Neutralizing monoclonal antibodies for treatment of COVID-19. *Nature Reviews Immunology*, 21(6):382–393, 2021.

[62] Bo Yao, Dandan Zheng, Shide Liang, and Chi Zhang. Conformational B-cell epitope prediction on antigen protein structures: a review of current algorithms and comparison with common binding site prediction methods. *PloS one*, 8(4):e62249, 2013.

[63] Steven A Frank. *Immunology and evolution of infectious disease*. Princeton University Press, 2020.

[64] Joakim Clifford, Magnus Haraldson Hoeie, Morten Nielsen, Sebastian Deleuran, Bjoern Peters, and Paolo Marcatili. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *bioRxiv*, 2022.

[65] Minjun Park, Seung-woo Seo, Eunyoung Park, and Jinhan Kim. EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. *bioRxiv*, 2022.