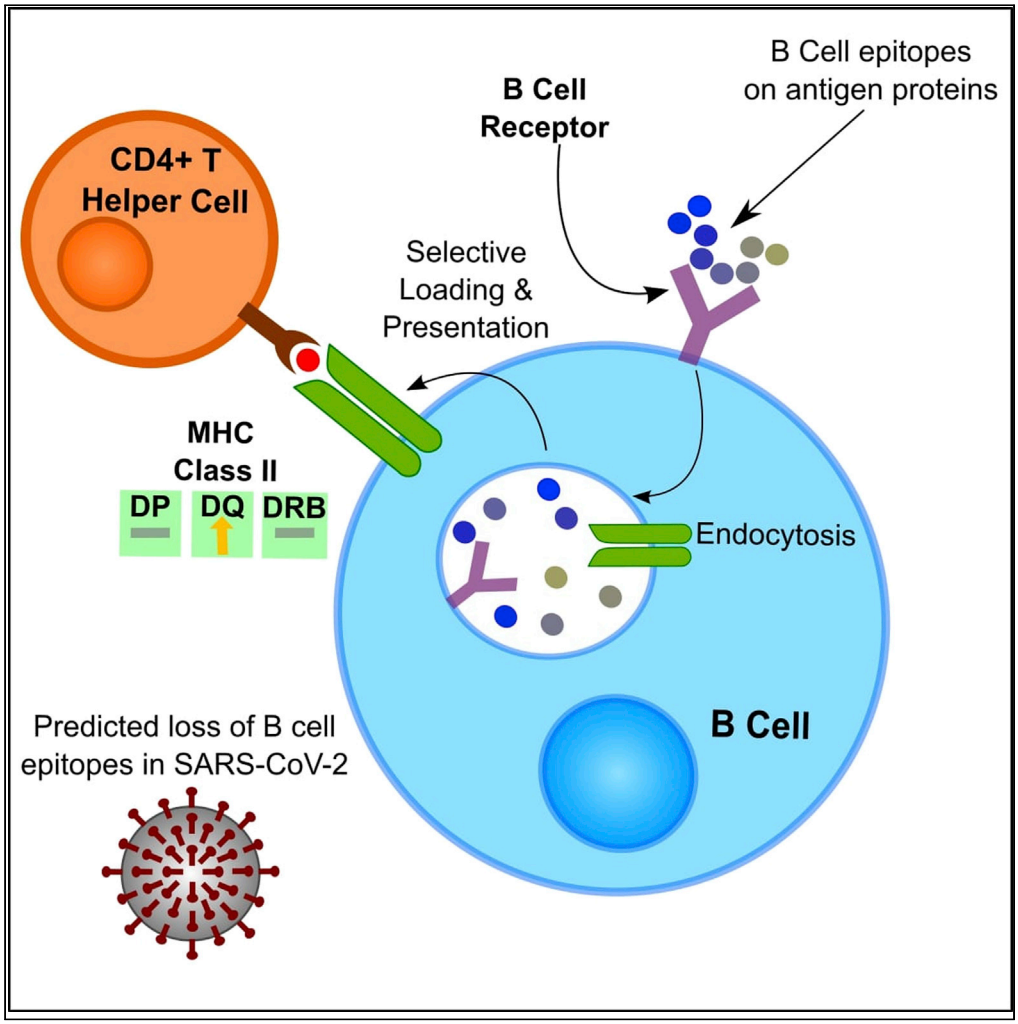
iScience

CellPress
OPEN ACCESS

Article

BepiTBR: T-B reciprocity enhances B cell epitope prediction



James Zhu, Anagha Gouru, Fangjiang Wu, Jay A. Berzofsky, Yang Xie, Tao Wang

yang.xie@utsouthwestern.edu (Y.X.)
tao.wang@utsouthwestern.edu (T.W.)

Highlights

BepiTBR leverages T-B reciprocity to enhance the prediction of B cell epitopes

We demonstrated the critical role of T-B reciprocity in B cell epitope generation

HLA class II DQ allele binders positively contribute to B cell epitope formation

BepiTBR identified a loss of B cell epitopes in SARS-CoV-2 proteins because of variants

Zhu et al., iScience *25*, 103764
February 18, 2022 © 2022
https://doi.org/10.1016/j.isci.2022.103764

iScience

CellPress
OPEN ACCESS

Article

BepiTBR: T-B reciprocity enhances B cell epitope prediction

James Zhu,[1] Anagha Gouru,[1] Fangjiang Wu,[1] Jay A. Berzofsky,[2] Yang Xie,[1,3,*] and Tao Wang[1,4,5,*]

SUMMARY

The ability to predict B cell epitopes is critical for biomedical research and many clinical applications. Investigators have observed the phenomenon of T-B reciprocity, in which candidate B cell epitopes with nearby CD4$^+$ T cell epitopes have higher chances of being immunogenic. To our knowledge, existing B cell epitope prediction algorithms have not considered this interesting observation. We developed a linear B cell epitope prediction model, BepiTBR, based on T-B reciprocity. We showed that explicitly including the enrichment of putative CD4$^+$ T cell epitopes (predicted HLA class II epitopes) in the model leads to significant enhancement in the prediction of linear B cell epitopes. Curiously, the positive impact on B cell epitope generation is specific to the enrichment of DQ allele binders. Overall, our work provides interesting mechanistic insights into the generation of B cell epitopes and points to a new avenue to improve B cell epitope prediction for the field.

INTRODUCTION

Antibodies are the key agents of humoral immunity produced by B cells, which target B cell epitopes and are involved in infectious diseases, cancers, and autoimmune diseases. B cell epitopes are the parts of antigen proteins recognized by antibodies. A deeper understanding of the features of B cell epitopes will propel the development of serological tests, vaccines, and therapies for various human diseases. Therefore, critical innovation is necessary to develop bioinformatics algorithms that can accurately predict B cell epitopes. B cell epitopes are either linear or conformational, meaning that the epitopes that induce antibodies could either be a linear stretch of amino acid residues of the antigen proteins, or could be formed by noncontiguous parts of the antigens that are close to each other in three-dimensional space (Benjamin et al., 1984; Berzofsky et al., 1982). The prediction of linear B cell epitopes has been the major focus of research. Popular linear epitope prediction algorithms include BepiPred 1.0 and 2.0 (Jespersen et al., 2017; Petersen et al., 2009), SVMTriP (Yao et al., 2012), LBEEP (Saravanan and Gautham, 2015), LBtope (Singh et al., 2013), *etc.* Popular conformational epitope prediction algorithms include Discotope1.1 (Haste Andersen et al., 2006), ElliPro (Ponomarenko et al., 2008), SEPPA 3.0 (Zhou et al., 2019), *etc.* Although discrepancies exist among different evaluation studies, the consensus appears to indicate that these methods perform only slightly better than a random classifier, and significantly fall behind HLA binding prediction software in quality (Galanis et al., 2019; Sanchez-Trincado et al., 2017). Specifically, the performance of B cell epitope prediction software achieved Area Under the ROC Curve (AUC) between 0.5 and 0.6, whereas HLA binding prediction software can now achieve AUC of >0.9. Usually, an AUC of 0.6–0.7 is considered acceptable and an AUC>0.7 is considered good (Mandrekar, 2010; Trifonova et al., 2014).

The recent SARS-CoV-2/COVID-19 pandemic has made it even more apparent the importance of studying B cell epitopes that can induce antibody responses. Robust antibody response can neutralize viral infectivity in a number of ways, such as interfering with viral binding to cellular receptors or blocking viral uptake into cells. Characterization of how the genetic variants that recently emerged in SARS-CoV-2 (van Dorp et al., 2020; Hassan et al., 2020; Plante et al., 2021; Toyoshima et al., 2020; Zhu et al., 2020) impact its B cell epitopes will reveal the selective pressures the human humoral immune system places on this virus, which could be critical for understanding the long-term effectiveness of the recently deployed SARS-CoV-2 vaccines. Importantly, Greaney et al., 2021) found that mutations at the Spike E484 position lead to >10-fold reduction in antibody neutralization, demonstrating the importance of monitoring emerging viral variants. However, given the large numbers of SARS-CoV-2 genomes that have been sequenced and shared publicly, it is impractical to test each viral epitope in every strain experimentally. A systematic and accurate bioinformatics profiling of all these SARS-CoV-2 strains *in silico* can help elucidate the

[1]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[2]Vaccine Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA

[3]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[4]Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[5]Lead contact

*Correspondence: yang.xie@utsouthwestern.edu (Y.X.), tao.wang@utsouthwestern.edu (T.W.)

https://doi.org/10.1016/j.isci.2022.103764

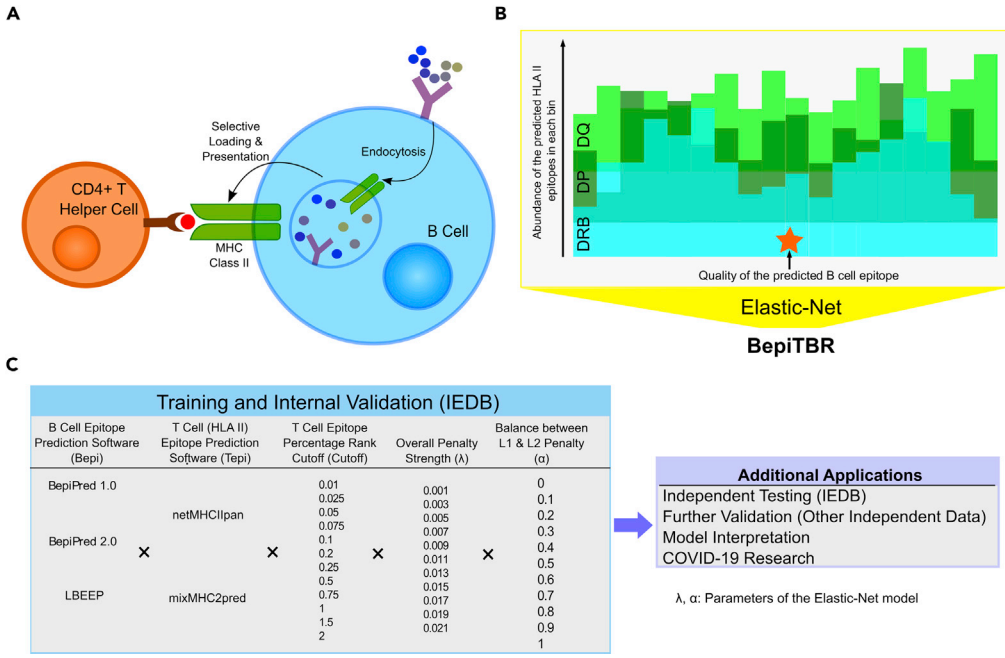


Figure 1. The rationale of the proposed model

(A) The process of B cell maturation involves help from $CD4^+$ T cells, which results in selective peptide loading of the MHC II complex.

(B) Cartoon of the format of the input data that are utilized by the BepiTBR model. The candidate B cell epitope is shown in blue. A window centering around each B cell epitope is examined in the antigen protein sequence and divided into bins. The MHC class II DP, DQ, and DRB allele binders are counted in each bin. The B cell epitope confidence score of the base model, the MHC class II binder counts, and their interaction terms form the input data.

(C) The process of model training and internal validation. The proposed BepiTBR model has incorporated three different base B cell epitope prediction models (Bepi.) and two different HLA class II epitope (T cell epitope) prediction models (Tepi.). To evaluate the model performance, we tested all possible combinations of base B cell epitope prediction models and HLA class II epitope prediction models, together with different parameters in the model: HLA class II epitope rank cutoff (cutoff), overall penalty strength (lambda), and balance between L1 and L2 penalty (alpha). The internal validation set was used to select the best parameter combination for each base model. The final models were further validated in other independent data, examined for model interpretation, and applied to COVID-19 research.

immune clearance mechanism, predict vaccine effectiveness, and facilitate antiviral antibody development, which is highly complementary with experimental approaches.

It has been observed that the activation of follicular B cells and the selection of high affinity B cell receptors are aided by $CD4^+$ T helper cells in an epitope-dependent manner, a phenomenon known as T-B reciprocity (Berzofsky, 1983; Ozaki and Berzofsky, 1987; Sabhnani et al., 2003; Zhang et al., 2014). As a result, the B cell epitopes with nearby $CD4^+$ T cell epitopes are more likely to be truly immunogenic and to induce mature B cell receptors (BCRs) and antibodies. For example, Brumeanu et al. observed that T or B viral synthetic epitopes from HA of the PR8 influenza virus were immunogenic not by themselves, but only when assembled as a contiguous dipeptide (Brumeanu et al., 1997). Mechanistically, Moss et al. proposed a direct "hand over" of antigen fragments from the BCRs to MHC II proteins (Moss et al., 2007). Alternatively, in the germinal centers, the protection from proteolysis of antigen epitopes by the bound antibody may lead to preferential MHC II-mediated presentation of the protected adjacent helper epitopes by the same B cells (Berzofsky, 1983; Ozaki and Berzofsky, 1987; Sabhnani et al., 2003; Zhang et al., 2014). Either case results in a selective loading (likely spatially constrained) of MHC II epitopes from BCR-internalized antigens (Figure 1A), which has been observed by Barroso et al. (Barroso et al., 2015). However, the detailed mechanisms of T-B reciprocity need to be further elucidated.

For prediction of B cell epitopes, no existing algorithm has leveraged this observation of T-B reciprocity. In this work, we showed that the incorporation of the intensities of nearby HLA class II epitopes (which are

potentially recognized by $CD4^+$ T-cells) significantly enhanced the prediction of B cell epitopes. We developed and validated a machine learning model, named BepiTBR, that incorporated this mechanism in the model specification. We showed that T-B reciprocity is a general biological principle that can be applied to enhance the prediction performance of different B cell epitope prediction software.

RESULTS

BepiTBR: leveraging T-B reciprocity to enhance the prediction of B cell epitopes

Owing to the important roles of linear B cell epitopes in immunizations and antibody production, we focused on linear B cell epitopes in this study. The core rationale of our model is that the enrichment of $CD4^+$ T epitopes/HLA class II epitopes in the neighborhood of candidate B cell epitopes could contain useful information to aid in the judgment of whether the potential B cell epitopes are truly immunogenic. We considered previous B cell epitope prediction models, BepiPred 1.0 and 2.0 (Jespersen et al., 2017) and LBEEP (Saravanan and Gautham, 2015), as our base prediction models for model improvements and comparisons. They were chosen because of the availability of standalone software packages. For prediction of HLA class II epitopes, we considered the classical netMHCIIpan software (Jensen et al., 2018) and the newly developed MixMHC2pred (Racle et al., 2019). All HLA class II alleles that are available from the HLA class II epitope prediction algorithms were considered.

The B cell assay data of human host from The Immune Epitope Database (IEDB) were used as our primary data source for the model development and validation/testing. The IEDB data were split into a training cohort (n = 10,764), a validation cohort (n = 3,588), and a test cohort (n = 1,251). Details of training and validation/test data creation are described in the STAR methodsection. For each IEDB record, the epitope sequence and the full antigen protein sequence were extracted. Full antigen protein sequences were obtained from NCBI'|'s Entrez protein database using the Entrez Programming Utilities. The HLA class II epitopes were predicted by either of the two software applications, for all available DRB alleles, all DP alleles, and all DQ alleles. Specifically, the DRB, DP, and DQ binders were predicted in a 15 a.a. residue-by-residue moving window from 180 a.a. upstream from the center of the B cell epitopes to 180 a.a. downstream. Then, the center positions of the predicted HLA class II epitopes were used to assign each epitope into one of the 18 nonoverlapping bins of 20 a.a. width. Because none of the available databases have recorded the HLA allele types of the host, we can only count the binders for all available HLA class II alleles, and treat those as a population average. However, as shown later, this approach is already providing a significant improvement to the prediction of B cell epitopes.

For the training of the enhanced B cell epitope prediction models, we include both the predicted confidence scores of one of the B cell epitopes by three existing algorithms, BepiPred 1.0, BepiPred 2.0, or LBEEP, and the binned counts of DRB, DP, and DQ binders in the neighborhood of the B cell epitopes, predicted by either netMHCIIpan or mixMHC2pred (Figure 1B). We also included interaction terms between the B cell epitope prediction score and each of the binned counts in the model. As there are many variables in the model, the Elastic-Net framework was used, which introduced both L1 and L2 penalties to the input variables, as shown in Figure 1B. The L1 penalty helps perform regularization and only retains the variables that are most important for prediction. The L2 penalty is beneficial to limit colinearity as the counts of DRB, DP, and DQ allele binders of each bin could have high correlation. The model was trained on the IEDB training cohort, and is referred to as "B cell epitope prediction enhanced by T-B reciprocity", BepiTBR, hereupon.

The performances of the BepiTBR models (hereby referred to as the "enhanced" models) were evaluated based on each of the three base B cell epitope prediction software on the full range of each tuning parameter. In particular, there were four tuning parameters (Figure 1C): (1) "Tepi. software", the choice of the HLA class II epitope prediction software (netMHCIIpan or mixMHC2pred), (2) "cutoff", ranging from 0.01 to 2, which was the percentile rank cutoff for determining HLA binders (smaller ranks refer to stronger binding), (3) "lambda", ranging from 0.001 to 0.021, which was the overall penalty strength, and (4) "alpha", ranging from 0 to 1, which was the balance between L1 and L2 penalties. For controls, we applied each of the three base B cell epitope prediction software applications on the same validation/test data ("base" model), and we also trained another set of Elastic-Net models with only the HLA class II epitope data ("HLA II-only" model). The HLA II-only models were trained with the same combinations of tuning parameters for head-to-head comparisons.

Improved prediction of B cell epitopes from considering T-B reciprocity

The Area Under the Curve of Receiver Operating Characteristic (AUROC) was used as the primary validation metric, which incorporated both the sensitivity and specificity of the prediction models. For ease of

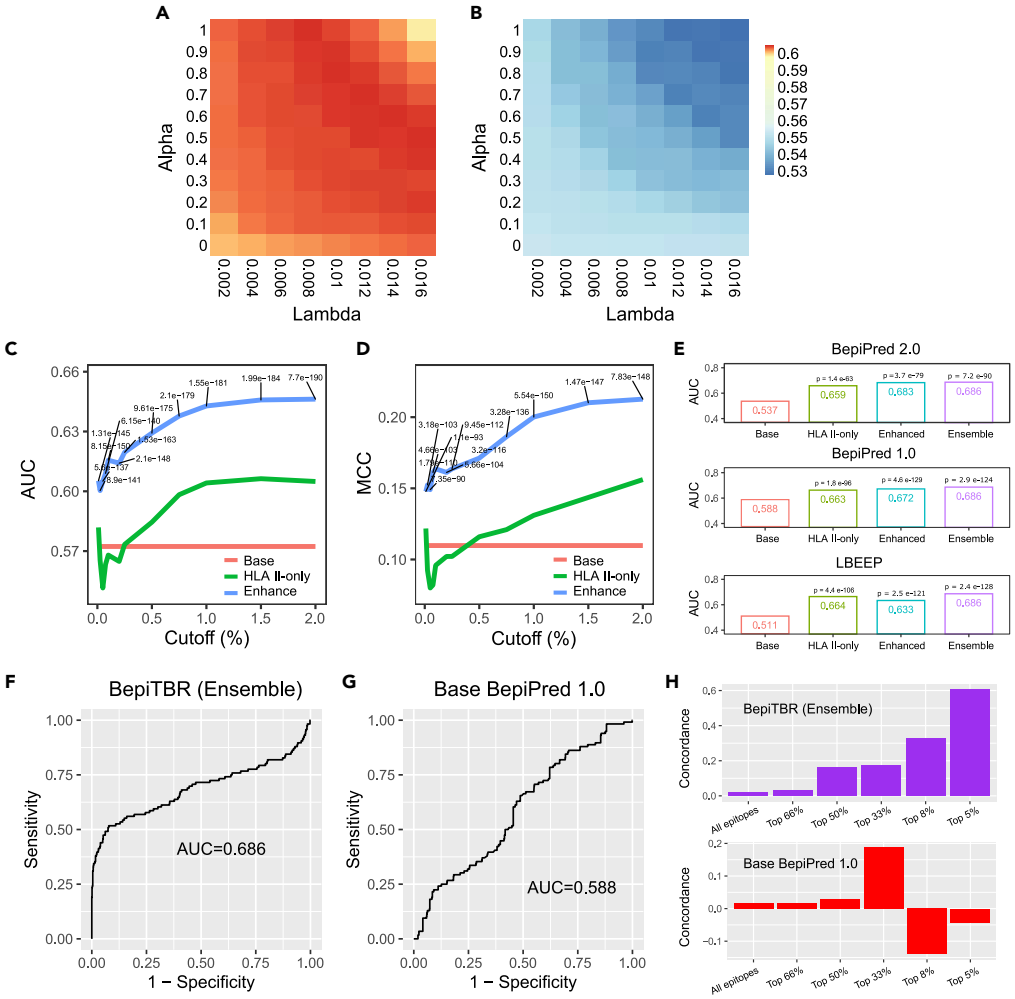


Figure 2. Prediction performance evaluation of the BepiTBR model

(A) A heatmap showing the AUCs in the validation dataset with the BepiPred1.0 B cell epitope prediction software and netMHCIIpan HLA class II epitope prediction software as the base models, at a rank percentile cutoff of 1. The parameter lambdas are shown on the X axis, and the parameter alphas are shown on the Y axis.

(B) A heatmap showing the AUCs in the validation dataset without using any base B cell epitope prediction model. In this model, only the HLA class II epitope prediction scores estimated by the netMHCIIpan software were incorporated in the model.

(C) A curve showing the best AUCs of all combinations of Tepi., lambdas, and alphas at each percentile cutoff vs. the cutoffs employed. The BepiTBR, the HLA II-only model (the model trained to predict B cell epitope only includes HLA class II epitope counts), and the base B cell epitope prediction model (LBEEP)' results are shown together.

(D) The same analyses as in (C), but performed with MCC as the benchmark metric.

(E) Barplots showing the AUCs of the BepiTBR models (enhanced model), the matching HLA II-only models (matched to the tuning parameters of the corresponding BepiTBR models), and the base B cell epitope prediction models, on the test cohort. The AUC of the ensemble BepiTBR model is also shown in each panel.

(F and G) The AUC of ROC plots for BepiTBR (ensemble) and for BepiPred 1.0.

(H) Pearson correlation between the similarity in B cell epitopes of any pair of Env proteins and the similarity in Libra-seq scores for the same pair of Env proteins across all sampled B cells. There are a total of five different Env proteins, and therefore, 10 possible pairs. Either all candidate B cell epitopes without any confidence score filtering or B cell epitopes with confidence scores larger than a cutoff are included.

See also Figures S1 and S2, and Table S1

comprehension and as an example, the performances of BepiTBR (BepiPred1.0-netMHCIIpan) for all combinations of the alpha and lambda parameters are shown in Figure 2A. The cutoff for HLA class II epitope percentile rank was 1. The performances of the model formed a gradient with the highest AUROC achieved

being 0.608. We also performed the same analysis for the base B cell epitope model (BepiPred1.0, AUC = 0.568) and the HLA II-only model (Figure 2B), and we showed that BepiTBR outperforms both control models across almost all ranges of parameters. The combination of BepiPred1.0 and netMHCIIpan is shown here as an example, whereas the results for other combinations are similar. The results were summarized by identifying the best combinations of the Tepi., alpha, and lambda parameters for the BepiTBR and the HLA II-only model (the base B epitope model is independent of these three parameters) for each choice of the rank percentile cutoffs. Figure 2C shows the performance of the "best" BepiTBR and HLA II-only model given each cutoff, as well as the base B cell epitope prediction model. BepiTBR outperforms both control models for all cutoffs chosen (LBEEP as the base model as an example). Figure S1A shows the same conclusion for BepiPred 1.0 and 2.0. Similar conclusions were also reached when the training and validation data were assigned using different random seeds (Figure S2). In addition, Matthews correlation coefficient (MCC), which has been used for benchmarking the accuracy of B cell epitope prediction tools in prior works (Galanis et al., 2019; Jespersen et al., 2017), was also used for scoring. The results are shown in Figures 2D, S1B, and S2, and confirmed that the three enhanced B cell epitope prediction models outperform the base models and the HLA II-only models.

Next, for each of the three BepiTBR models, we chose the best parameter configuration of HLA class II epitope prediction software, rank percentile cutoff, alpha, and lambda, according to the performances in the validation cohort. We then validated the chosen BepiTBR models on the independent test cohort (Figure 2E) and observed that all three BepiTBR models outperform the original B cell epitope prediction models by 0.1–0.15 in AUC. BepiTBR (BepiPred 1.0) and BepiTBR (BepiPred 2.0) also still outperform the corresponding HLA II-only models. BepiTBR (LBEEP) is not as successful as the corresponding HLA II-only model, but BepiTBR (LBEEP) also seems to be the least accurate of all three BepiTBR models. The Bayes Factor (Guinney et al., 2017; Seyednasrollah et al., 2017) was employed for deriving the statistical significance of model comparison, which again confirmed the significant improvement in the performance of BepiTBR models (Table S1).

To further improve the prediction performance, an ensemble model was created, BepiTBR ("ensemble") that averaged the prediction scores of the three base models: BepiTBR (BepiPred 1.0), BepiTBR (BepiPred 2.0) and BepiTBR (LBEEP). This ensemble model achieved an AUROC of 0.686 on the test cohort, and achieved the best performance compared to all base B cell epitope prediction models, all HLA II-only models, and all separate BepiTBR models (Figure 2E). The ROC plots for BepiTBR (ensemble) and the base BepiPred 1.0 model as a control are shown in Figure 2F. We noticed that the BepiTBR model possesses very strong sensitivity characteristics (shown by the bulge at the lower left corner), which is usually preferred given the same AUROC. However, this is not the case for the base BepiPred 1.0 model (Figure 2G), or any of base BepiPred 2.0 or base LBEEP models (data not shown).

To further validate the performance of the model in real data of antibody response, BepiTBR was tested on the Libra-seq data from Setliff et al., 2019, which allows high-throughput mapping of BCR sequences to antigen specificity. Setliff et al. performed the Libra-seq experiment on an HIV patient, n90. Five strains of HIV-1 Env sequences (KNH1144, BG505, ZM197, ZM106.9, and B41) were screened in this patient, and Libra-seq yielded a total of 1,465 B cells with BCR specificity known for each of the five Envs. The binding specificity, Libra-seq score, is measured as a continuous variable. When the Env sequences between two strains are more similar, the BCR repertoire will likely show a more similar response and vice versa. Under this rationale, the sequence similarity was measured by the number of shared candidate B cell epitopes between any two strains. The similarity in Libra-seq scores was defined by the correlation between the Libra-seq scores across all BCRs for the same pairs of Env antigens. For each given B cell epitope prediction model, we included all predicted B cell epitopes on the Env proteins (15-mers in overlapping moving windows), as well as only epitopes with confidence scores that meet different cutoffs, because we hypothesize that this definition of B cell epitope similarity is more relevant when considering B cell epitopes of higher immunogenicity. For all pairs of Env antigens, a Spearman correlation was calculated between similarity in Env B cell epitopes and similarity in Libra-seq scores. There is indeed a higher concordance (correlation) between the predicted Env immunogenicity and the measured Libra-seq scores when considering B cell epitopes of higher BepiTBR confidence scores (Figure 2H). This trend is less clear when using epitopes predicted by BepiPred 1.0 (Figure 2H), and even less so for BepiPred 2.0 and LBEEP (data not shown).

For the sake of generating unified prediction scores and consistent conclusions, the BepiTBR model will subsequently refer to this ensemble model if no distinctions are provided. Overall, the incorporation of

the HLA class II epitopes provides significant improvement for the prediction performances of B cell epitopes.

MHC DQ allele binders are positively associated with immunogenic B cell epitopes

The three main classes of MHC class II alleles are DR alleles, DQ alleles, and DP alleles, which were considered separately in our BepiTBR models. For HLA-DR, only DRB alleles were considered as DRA alleles are not used by any of the HLA class II epitope prediction software. To investigate which of these alleles are most critical for the generation of B cell epitopes, and the ways in which these alleles participate in this process, we extracted the estimated coefficients of the Elastic-Net models in BepiTBR (BepiPred 1.0), BepiTBR (BepiPred 2.0), and BepiTBR (LBEEP). We examined the coefficient of each 20 a.a. bin (from −180 a.a. to 180 a.a.) for each allele class and for each of the three BepiTBR models. Overall, the coefficients for the DQ alleles' bins were mostly positive, meaning a higher density of DQ allele binders would lead to higher chances of the B cell epitopes being predicted as true (Figure 3A). Bootstrap resampling was performed to generate the CIs of these coefficients, which confirmed that most of these coefficients for DQ binders are significantly larger than 0. Comparatively, the coefficients for DRB and DP alleles across all bins do not have a clear trend of being positively associated with B cell epitope generation (Figure S3). In fact, there seems to be a modest negative correlation between DRB allele binders and B cell epitope immunogenicity.

Next, we directly tested whether the counts of the DQ allele binders in each bin of the positive training cases (truly immunogenic B cell epitopes) were different from those of the negative cases (Figure 3B, positive values mean higher counts in positive cases than negative cases, and vice versa. Same in Figure S4). Many of the bins showed significantly higher DQ allele binder counts in the positive than in the negative cases. Moreover, considering the results from both Figures 3A and 3B together, it seems that the DQ binders around the candidate B cell epitopes, and to a lesser extent, ∼100 amino acid downstreams of the B cell epitopes are both critical for the B cell epitopes to be truly immunogenic. On the other hand, many of the bins showed significantly higher DRB allele binder counts in the negative cases than in the positive cases (Figure S4).

One potential confounding variable of the observed associations between HLA-DQ/DP/DRB ligands and B cell epitopes is the hydrophobicity of the amino acids. Amino acids on the protein surface are less likely to be hydrophobic; therefore, if the predicted DQ binders in our dataset happen to be enriched in less hydrophobic residues, then the observed signal may be an artifact of surface accessibility and not an indication of T-B reciprocity. To rule out this possibility, we tested whether HLA ligands for each class of alleles, predicted by mixMHC2pred and netMHCIIpan, as well as experimentally validated ligands documented by IEDB, were enriched in hydrophobic amino acids. If our observations were because of an artifact related to surface exposure, it would be expected that the true or predicted DQ ligands have less hydrophobic amino acids compared with DRB ligands. However, our analyses showed this is not the case (Figure S5).

Furthermore, if DRB/DQ/DP allele binders indeed participated in the generation of mature BCRs and epitope-specific B cells in the germinal center, we should expect that the expression levels of DRB/DQ/DP alleles in the B cells to also play a role. Five B cell scRNA-seq datasets generated from PBMCs of healthy donors with paired BCR-sequencing were examined (Figure 3C). In most datasets, the higher expression of DQ alleles was associated with higher levels of B cell activation, characterized by the switching of heavy chain classes to IgG/IgE/IgA, in sharp contrast to DP and DRB alleles (one example dataset shown in Figure S6). The expression of DRB, DQ, and DP alleles was usually correlated in B cells, which presented a confounding factor to this analysis. A multivariate regression against the B cell activation status (class switching) was performed based on the gene expression of DRB, DQ, and DP alleles to remove confounding correlations. Indeed, across all five datasets, the higher expression of DQ alleles was positively associated with the class switching of B cells, with statistical significance achieved in three datasets (Figure 3D). Moreover, across all datasets, the lower expression of DRB alleles was associated with activation of B cells.

A natural question follows that whether the investigation of expression of B cell activation markers in these B cell scRNA-seq data would yield the same conclusion. However, the B cells in these scRNA-seq datasets are all from peripheral blood, whereas BCR/B cell maturation (and also class switching) happens in the germinal centers. Once exiting the germinal centers, the expressional programs of the B cells may not

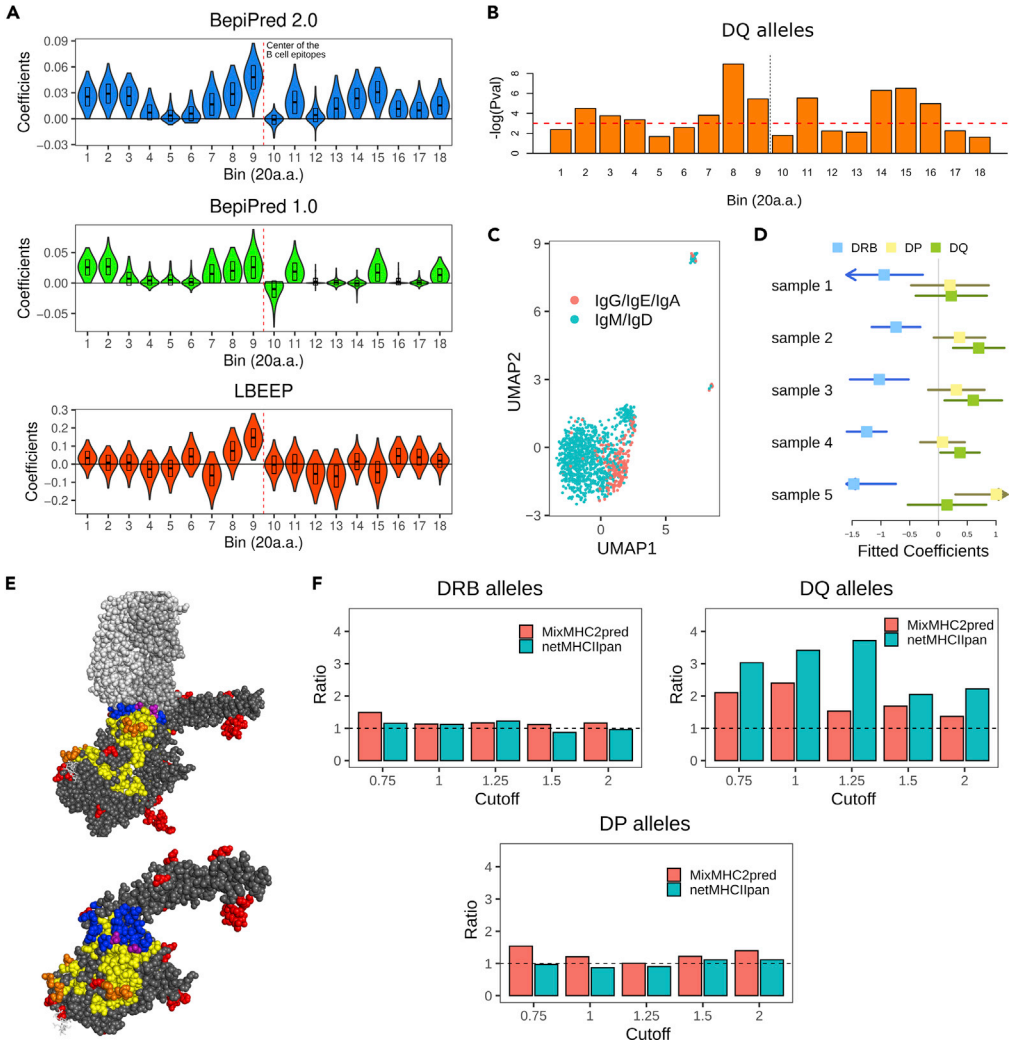


Figure 3. BepiTBR reveals insight into the generation of B cell epitopes

(A) The Elastic-Net model coefficients, for DQ alleles, of the BepiTBR (BepiPred 2.0), BepiTBR (BepiPred 1.0), and BepiTBR (LBEEP) models. 1–18 represents the 18 bins covering (−180 a.a., 180 a.a.) in 20 a.a. intervals. 200 bootstraps were performed to derive the distributions of coefficients.

(B) Negative log p values of the Mann-Whitney U tests investigating whether the true B cell epitopes have different counts of DQ allele binders in each of the 18 bins, compared with the negative cases. Positive direction: true B cell epitopes have higher counts of binders; Negative direction: true B cell epitopes have lower counts of binders.

(C) UMAP plot of single B cell scRNA-seq data showing the clustering of B cells by their class switching status. The vdj_v1_hs_pbmc dataset was shown as an example.

(D) Multivariate logistic regression of the expression of DP, DQ, and DRB alleles in each B cell against the status of class switching. Forest plots were used to display the fitted coefficients and their CIs. Samples one to five are sc5p_v2_hs_PBMC, vdj_nextgem_hs_pbmc3, vdj_v1_hs_pbmc2, vdj_v1_hs_pbmc, and vdj_v1_hs_pbmc3, respectively.

(E) The 3D structure of 3N85 from PDB, showing the antigen (bottom dark gray), antibody (top light gray), the curated conformational B cell epitopes (purple), the predicted conformational B cell epitopes by discotope (red), and the predicted HLA class II epitopes by MixMHC2pred (yellow). Blue color refers to the overlap between the curated B cell epitopes and HLA class II epitopes. Orange color refers to the overlap between HLA class II epitopes and the predicted B cell epitopes. Close-up image of the same 3 d structure, with the antibody removed, is also shown.

(F) Barplots showing the ratios of the number of curated conformational B cell epitope residues that are closer to predicted HLA class II epitopes on the same antigen proteins divided by the number of B cell epitope residues that are closer to epitopes not predicted to bind HLA class II proteins. The distances to HLA class II epitopes are averaged for each B cell epitope residue and the same is done for non-binding epitopes. All B cell epitopes of all structures available from PDB are aggregated.

See also Tables S2, S3, and S4; Figures S3, S4, S5, S6, and S7.

be representative of their progenitors in the germinal centers. Nevertheless, we tried to correlate the expression of *DQ*/*DP*/*DRB* with *CD19*'s expression. CD19 is a marker of B cell activation/maturation during the developmental stage (Wang et al., 2012) and is thus a good fit for correlation with MHC class II proteins. In these scRNA-seq data, all of *DP*, *DQ*, and *DRB* demonstrate a strongly positive correlation with *CD19* expression. But *DQ* does not have a more positive expression than *DP* or *DRB* (Table S2). Therefore, as we explained above, this analysis may be biased and may not best capture the real relationship between MHC class II proteins and B cell/BCR maturation (against B cell epitopes) during development. Therefore, it will be more informative to examine the correlation between *CD19* and *DQ*/*DP*/*DRB* in germinal center B cells. We analyzed the correlation between the expression of the *HLA-DQ*/*DP*/*DRB* genes and *CD19* in germinal center B cells from a 10X human lymph node Visium dataset (Figure S7, processed data from (Wang et al., 2021a)). We showed that the correlation between *HLA-DQ* and *CD19* is the most positive and much higher than those of *HLA-DP*/*DRB* (Table S3). Overall, our results strongly support the role of DQ alleles in B cell/BCR maturation, and thus the importance of DQ binders for B cell epitope formation.

Conformational B cell epitopes are also associated with enrichment of DQ epitopes

Though linear B cell epitopes are the main focus of this study, we also tested whether we could make similar observations with conformational B cell epitopes. Because of the relative rarity of conformational epitope data and the need to possibly consider the complicated three dimensional relationships, it was infeasible to evaluate T-B reciprocity in a machine learning model, as above for BepiTBR. Instead, we quantitatively evaluated whether validated conformational epitopes were closer to amino acids within epitopes predicted to bind or not bind HLA class II proteins. A total of 92 structures were extracted from IEDB that were protein complexes of interacting human antigen-antibody with valid full antigen sequences. The true antibody-binding residues have already been curated by IEDB (one example shown in Figure 3E). Again, we predicted the HLA class II epitopes using netMHCIIpan and mixMHC2pred, for all available DRB alleles, DP alleles, and DQ alleles. We examined whether each amino acid of the antigen protein is part of a predicted HLA class II epitope presented by any of the available DRB, DP, and DQ alleles for netMHCIIpan or mixMHC2pred, respectively.

We calculated the average spatial distances between amino acid residues that were part of the antibody-interacting B cell epitopes and amino acids that were or were not in predicted HLA class II epitopes (putatively $CD4^+$ T cell epitopes). The "B epi.-T epi." average distance and the "B epi.-non-T epi." average distance were compared, for all curated conformational B cell epitope residues of all structures. For example, for the mixMHC2pred software, at a rank percentile cutoff of 1, 70.6% of the B cell epitope residues were closer to DQ binding amino acids than non-DQ binding amino acids. For netMHCIIpan, 78.8% of B cell epitopes were closer to predicted DQ binders at a rank percentile cutoff of 1.25. This analysis was systematically conducted for both HLA class II epitope prediction software, across several cutoffs between 0.75 and 2 (comparable to the range in which BepiTBR performs best, Figures 2C and 2D), and for each allele class. Across the different settings for DQ alleles, more B cell epitope residues demonstrate the phenotype of having shorter "B epi.-T epi." distances than "B epi.-non-T epi." distances (Figure 3F). In contrast, there does not seem to be a positive nor negative enrichment for the DP and DRB alleles (Figure 3F). Therefore, for conformational epitopes, we made the same observations for DQ alleles as with linear B cell epitope data.

BepiTBR predicts B cell immunogenicity loss in SARS-CoV-2 strains

A systematic bioinformatics profiling of all these SARS-CoV-2 strains *in silico* can help elucidate the immune clearance mechanism, predict vaccine effectiveness, and facilitate antiviral antibody development, which is highly complementary with experimental approaches. We analyzed all of the collected viral sequences (N = 1,959,135) and identified 21,917 unique strains with high sequencing quality (detailed preprocessing steps described in the STAR methodsection). Variant calling for protein-changing mutations (Figures 4A and S8) was performed. Most of the top mutations detected have been reported before, such as P4715L (Orf1ab) (Toyoshima et al., 2020), D614G (S) (Plante et al., 2021), Q57H (Orf3a) (Hassan et al., 2020), *etc.*, confirming the validity of our variant calling. Further, it is interesting to note that many variants in the S protein are outside the S receptor-binding domain (RBD), which is responsible for engaging with the receptor on host cells (Zhou et al., 2020).

BepiTBR was used to profile the B cell immunogenicity map of the SARS-CoV-2 proteins in all unique viral sequences. Figure 4B shows the confidence scores of the BepiTBR predictions for epitopes from all

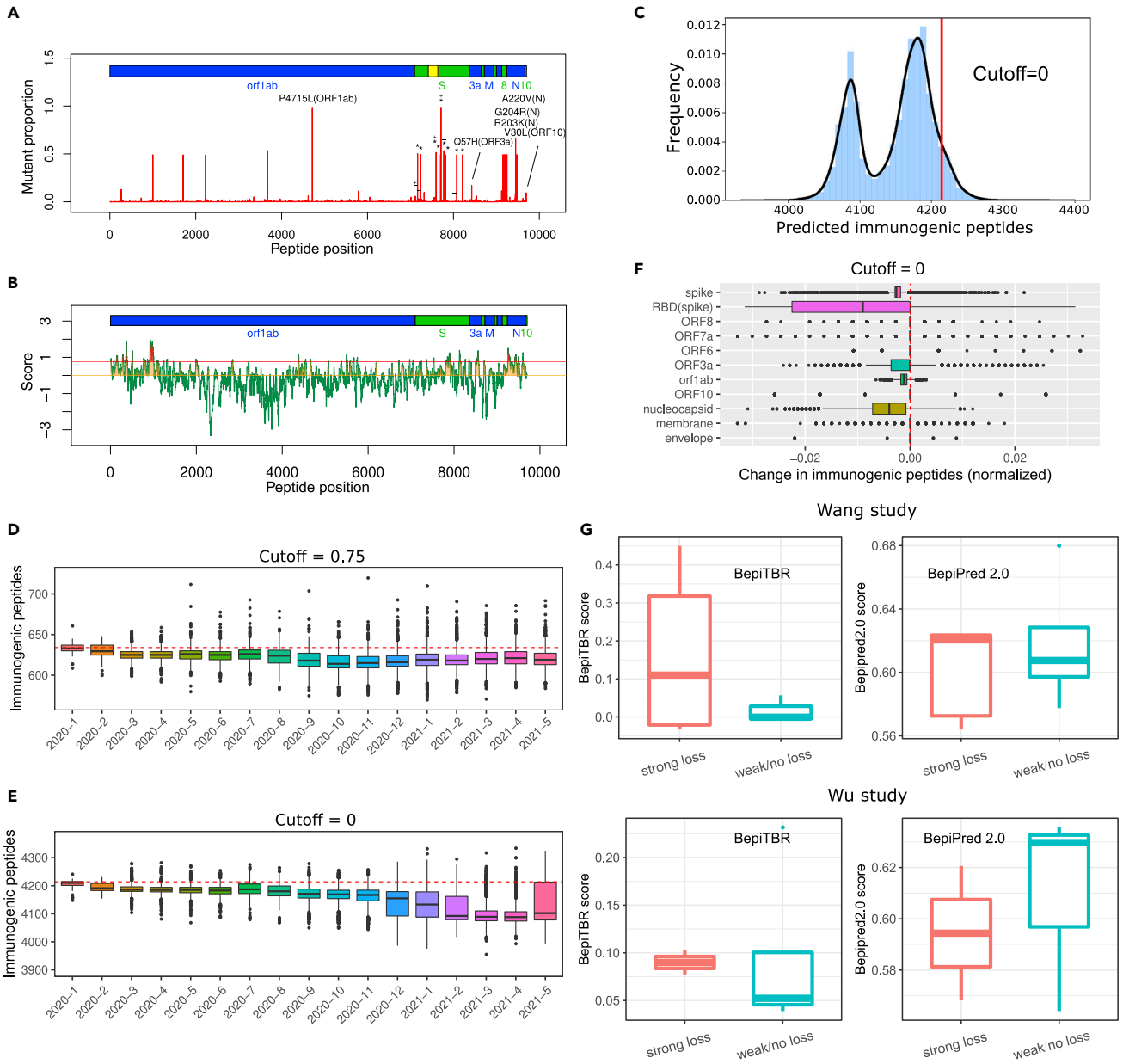


Figure 4. BepiTBR predicts B cell immunogenicity loss in SARS-CoV-2

(A) Variants detected in the SARS-CoV-2 strains, compared to the reference genome MN908947. Y axis shows the proportion of viral strains with a particular mutation. Mutations with relatively high abundances in the analyzed sequences are labeled. In the S protein region, "*" indicates mutations associated with (B)1.1.7 (Alpha/UK variant),"+" denotes mutations present in (B)1.351 (Beta/South African variant), and "-" denotes mutations present in (B)1.617.2 (Delta/Indian variant).

(B) The B cell epitope confidence scores by the BepiTBR model for each epitope of each SARS-CoV-2 viral protein. The cutoff of 0 is shown by the yellow line and the cutoff of 0.75 is shown by the red line.

(C) The number of predicted B cell epitopes of all viral proteins of all SARS-CoV-2 strains. The red line shows the number for the reference strain. Cutoff = 0.

(D) The number of predicted B cell epitopes of all viral proteins of all SARS-CoV-2 strains, broken down into the months by which they were first discovered. Cutoff = 0.75.

(E) The same analyses as in (D), but for cutoff = 0.

(F) The change in the number of predicted B cell epitopes of each viral protein of all SARS-CoV-2 strains, with respect to those of the reference genome, normalized by protein lengths. Counts for the Spike RBD are normalized by the length of the S RBD. Cutoff = 0.

Figure 4. *Continued*
(G) For the Wang et al. and Wu et al. studies, we investigated the B cell epitopes that were lost in the mutated S protein sequences compared to the reference S sequence. We calculated the average B cell immunogenicity confidence score, by BepiTBR (left) and by BepiPred 2.0 (right), for the lost B cell epitopes in each mutated S protein sequence.
See also Table S4; Figures S8, S9, S10, and S11.

SARS-CoV-2 viral proteins of the reference genome MN908947. To demonstrate the robustness of our analyses, we chose the cutoffs of 0 and 0.75 to pick the top predicted immunogenic epitopes. In particular, at the cutoff of 0, the predicted number of B cell epitopes is close to the predicted number of B cell epitopes by BepiPred 2.0 for SARS-CoV-2 (Zhu et al., 2020). B cell epitopes are predicted in most of the viral proteins by BepiTBR, including abundant epitopes in the nucleocapsid (N) protein. This corresponds well with the observation of Hachim et al (Hachim et al., 2020), whose Luciferase Immunoprecipitation System (LIPS) assay discovered antibody response against most SARS-CoV-2 proteins, with N being the most immunogenic. Further examination of the N protein showed that the predicted high quality B cell epitopes are concentrated in the N terminal of the N protein. This is in alignment with the observation of Phan et al., whose ELISA assays (Phan et al., 2021) showed that the N terminal of the N protein is much more immunogenic than the other regions of N.

We investigated the total number of predicted epitopes in all viral strains (Figures 4C and S9A). The SARS-CoV-2 strains tend to have a loss of immunogenicity compared to the reference genome. To more clearly demonstrate this, the collection times of all viral strains were examined to bin all viral strains into the months in which they were first discovered. Through binning, we observed that, from the onset of the pandemic outbreak, there has been an overall continuous decrease in B cell immunogenicity in SARS-CoV-2 (Figures 4D and 4E). Each of the SARS-CoV-2 proteins was examined to calculate the changes in the number of predicted B cell epitopes compared to the reference genome, normalized by protein lengths (Figures 4E and S9B). This analysis also included the S RBD. As expected, the S protein showed a high rate of B cell immunogenicity loss and the Orf1 protein also showed some loss of immunogenicity (Figures 4E and S9B).

We further investigated whether the HLA class II epitopes in the SARS-CoV-2 proteins also showed any changes over time, for DQ, DP, and DRB alleles (Figure S10). Consistent with the loss of predicted B cell epitopes, DQ allele binders showed a clear trend of decrease, especially in recent months. The trend for DP and DRB allele binders was not clear, whereas the number of DP/DRB allele binders displayed a slight increase before November of 2020 and a large subsequent decrease. Overall, the HLA class II epitopes in the SARS-CoV-2 proteins demonstrated a pattern consistent with our expectations.

Wu et al. examined whether the Moderna mRNA-1273 vaccine still induced neutralizing antibodies against SARS-CoV-2 Spike mutants (Wu et al., 2021). They discovered that D614G, D614G/N501Y, N501Y/P681H/dH69V70, and B.1.1.7 (see definition in STAR methodsection) induced little or no effect on neutralization in mRNA-1273 Phase 1 participants' sera, whereas D614G/N501Y/K417N/E484K and B.1.351 induced significant loss in neutralizing titers. BepiTBR was used to predict the B cell epitopes in the reference S protein sequence. Figure 4G shows the confidence scores of the predicted B cell epitopes in the reference S protein sequence that were lost in these variant sequences. Importantly, we only predicted and examined immunogenic B cell epitopes in the reference S sequence in this analysis. The antibodies are from the vaccine's sera, which are developed against the original S, and the new B cell epitopes that emerged in the variant S sequences are irrelevant. The lost epitopes of S sequences that are associated with severe decrease of serum neutralizing activities indeed have higher confidence scores than the other lost epitopes from variant sequences of minimum/no decrease in neutralization. Similarly, Wang et al., 2021b examined another group of recently reported variants, including N501Y, E484K/R683G, K417N/N501Y/E484K/R683G, K417N, N439K, A475V, Q493R, and N440K, and showed that they induced significant loss of neutralizing activity in either vaccine plasma (Moderna or Pfizer-BioNTech) or against neutralizing monoclonal antibodies. In contrast, the other variants, including R346S, Y453F, S477R, R683G, and D614G, did not induce significant loss. In alignment with the experimental validation, our model also correctly predicted which variants will induce severe loss of neutralization (Figure 4G). For the Wu et al. and Wang et al. studies, we performed the same analyses with BepiPred 2.0 as control, which showed that BepiPred 2.0 is not as good as BepiTBR in predicting the variants that induced loss of neutralization (Figure 4G). These analyses proved again that BepiTBR is a much improved B cell epitope prediction model, compared to other predecessor software.

BepiTBR predicts minimum loss of B cell immunogenicity in SARS-CoV and MERS-CoV

There are multiple forces that constrain the evolution of viruses, including humoral immune responses as well as other factors that affect the fitness advantages of the virus through mechanisms independent of antibody responses. To serve as a control for SARS-CoV-2, all viral genomic sequences available for SARS-CoV (N = 19) and MERS-CoV (N = 519) were also analyzed. The SARS-CoV virus has a much shorter circulation time in the human population than SARS-CoV-2. The MERS-CoV virus is characterized by spillovers from camels into humans independently over multiple years (Killerby et al., 2020). Given this, we anticipate the genetic variants in SARS-/MERS-CoV to be mostly independent of human humoral immune responses because of the limited time of sustained human circulation and human-to-human transmission. Thus, BepiTBR was used to predict the number of B cell epitopes in the reference genomes of SARS-/MERS-CoV, which are some of the earliest appearing strains for each virus, and compared each against all other strains of SARS-/MERS-CoV (Figure S11). As expected, BepiTBR predicted no obvious loss of B cell immunogenicity for either SARS-CoV or MERS-CoV.

DISCUSSION

In this work, we showed that T-B reciprocity is an overlooked process in terms of B cell epitope prediction, and built BepiTBR to incorporate T-B reciprocity in the B cell epitope prediction process. To our knowledge, none of the publicly available B cell epitope prediction software algorithms have leveraged this observation. Moreover, by building upon several off-the-shelf B cell epitope and HLA class II epitope prediction software, we showed that T-B reciprocity provides general enhancement of performances, rather than being specific to a particular combination of B cell epitope and HLA class II epitope prediction algorithms. The final BepiTBR model achieved an AUC of 0.686 on an independent validation cohort, representing an improvement of AUROC of ~0.1–~0.17 compared to the three base models. A "reasonable prediction performance" should take into consideration the specific research domain, difficulty level of the question, and what prior works have achieved. Prior studies (Galanis et al., 2019; Sanchez-Trincado et al., 2017) have achieved AUC only between 0.5 and 0.6, reflecting the high degree of difficulty in prediction of B cell epitopes and the relative lack of success of prior methods. Our performance of 0.686 should be considered satisfactory in this context. The analyses of the linear and conformational B cell epitopes both yield this similar conclusion. Our work points to a new direction for future works to further improve the prediction of B cell epitopes bioinformatically. More importantly, this work reiterated and reinforced the importance of T-B reciprocity, which has been underappreciated as a valid biological phenomenon that governs the generation of B cell epitopes.

One might argue that any epitope, when exposed to B cells, has the potential to induce antibody responses. However, as shown by the results documented by IEDB, some epitopes were tested more than 10 times in different experimental settings and were negative in all tests. Therefore, not all epitopes are capable of inducing immunogenic responses or are at least highly inefficient in doing so. Moreover, our work showed that T-B reciprocity results in preferences for certain candidate epitopes to be more or less potent in inducing antibody responses, governed by restrictions on spacing to HLA class II epitopes. Thus, a high performance predictive model can offer great value by unveiling the potentially large variations of B cell epitope immunogenicity within an antigen.

We made several interesting observations when constructing the BepiTBR model. First, for all three base models, the best HLA class II epitope prediction software was mixMHC2pred, which seems to confirm mixMHC2pred as an improved MHC class II binder prediction software compared to netMHCIIpan (Racle et al., 2019), at least for the purpose of BepiTBR. Next, as shown in the test cohort, both in terms of the contributions to the performances of the BepiTBR (BepiPred 1.0, BepiPred 2.0, and LBEEP) models and the performances of the base models, BepiPred 1.0 and 2.0 appear to be comparable with each other, whereas both appear better than LBEEP. This is consistent with the rankings of these models' performances from other studies (Galanis et al., 2019; Jespersen et al., 2017). Third, comparing the HLA II-only models (trained on HLA class II epitope data for the task of B cell epitope prediction) and the base models, we found that the HLA II-only models generally perform much better than the three base models, which are dedicated prediction models for B cell epitopes. The three base models were all developed with overall narrow consideration of the B cell epitopes themselves, whereas the HLA II-only models (and also BepiTBR) explicitly considered the information of a much wider neighborhood. Therefore, it seems that the neighborhood of the final B cell epitopes is important for the induction of B cell responses and contains essential information for improving prediction of B cell epitopes.

We unexpectedly found that the existence of DQ allele binders particularly enhances the probabilities for the candidate B cell epitopes to be truly immunogenic, in contrast to DRB and DP allele binders. Furthermore, we

found that the DQ allele binders immediately adjacent to the candidate B cell epitopes are most critical for this process. On the other hand, the results for DP and DRB alleles were dramatically different. In fact, we observed a negative correlation between DRB alleles and B cell epitope immunogenicity in several of our analyses. These observations confirm that the improvement in prediction accuracy in BepiTBR is not an artifact related to caveats such as antigen protein length. Otherwise, the same pattern should have been observed for DRB, DP, and DQ allele binders. These observations also shed light on the biological process of T-B reciprocity, especially the curious opposing effect between DQ and DRB. One possible explanation is that DRB allele binders might be more likely to be T_{reg} epitopes rather than T_{helper} epitopes, whereas only T helper cells participate in T-B reciprocity. In fact, the T_{reg} epitopes first described by Cousens et al. (Cousens et al., 2013, 2014) indeed have high affinity binding to DRB alleles. These interesting observations and hypotheses are worth future mechanistic investigation, but are beyond the scope of this current study.

Applying BepiTBR in SARS-CoV-2 sequences, we found a possible decrease in B cell immunogenicity over time. With more than a year of human circulation, the antigenic drift of SARS-CoV-2 has already become obvious. Such B cell antigenic drift has previously been observed in influenza, and Li et al. showed that individuals typically receive re-infection with a drifted strain every 5 to 10 years (Li et al., 2019). This raises great concerns for reinfection of SARS-CoV-2, which has been reported increasingly frequently (Cohen and Burbelo, 2020). This also raises concerns for long-term effectiveness of the recently developed vaccines, especially given the rise of the Delta and Omicron variant strains. On the other hand, there is also the possibility that some antigenic drift could make the virus less pathogenic and less "visible" to the human immune system, which could increase the survival fitness of this virus. This coincides with the reported increase in proportions of asymptomatic COVID-19 cases (Ren et al., 2021). All of these observations further advocate for development of better B cell epitope prediction models, such as BepiTBR, which can be applied to quickly screen the large amounts of collected viral sequences and to monitor the overall trend of B cell antigenic drift in different regions and populations.

In conclusion, our study provided independent confirmation of the concept of T-B reciprocity and has paved the way for future works on the development of B cell epitope prediction algorithms. This is critical for developing better preventive and therapeutic measures not only against infectious pathogens, but also for other diseases like cancers, with more recent studies reporting curious roles B cells play in the tumorigenesis process.

Limitations of the study

One limitation of the current study is that we did not have the opportunity to consider the individual-specific HLA alleles in the model. As each individual has a unique set of HLA class II alleles, the HLA class II epitopes on the antigen proteins will also be different. However, currently available data from IEDB or other sources do not have this level of detailed information. Therefore, we predicted HLA class II epitopes with all available alleles from the epitope prediction algorithms and treated the predicted epitopes as a population average. Strikingly, we found that this approach had already significantly improved the prediction of B cell epitopes, which serves as a foundation for future work that seeks to generate B cell epitope immunogenicity data with matched HLA typing for the host to further improve BepiTBR.

We note that the prediction of B cell immunogenicity, the focus of this study, is tasked with the prediction of whether an epitope will induce any antibody response (without knowing the many possible antibodies that would arise), whereas a related but different prediction problem is to predict the pairing between an epitope and a given antibody. Both questions are valid and have different areas of application. The former could be used for determining candidate epitopes for vaccine development or monitoring antigenic epitope drift, whereas the latter may be useful in scenarios such as engineering of bispecific antibodies (Fan et al., 2015).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - B cell epitope data collection

- Linear and conformational B cell epitope predictions
- HLA class II epitope prediction
- BepiTBR model
- Acquisition of the coronavirus genome sequences
- Mutation calling in SARS-CoV-2 protein sequences

- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.103764.

ACKNOWLEDGMENTS

This study was supported by the National Institutes of Health (NIH) [CCSG 5P30CA142543/TW and YX, 1R35GM136375/YX, NIH 1R01CA258584/TW, NIH 2P50CA070907-21A1/TW, and YX], Cancer Prevention Research Institute of Texas [CPRIT RP190208/TW, RP180805/YX].

AUTHOR CONTRIBUTIONS

J.Z. performed all bioinformatics analyses. A.G. contributed to the graphics. T.W., and Y.X. supervised the study. J.B. provided input on the biological interpretation of the bioinformatics results. J.Z., A.G., T.W., J.B., and Y.X. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 17, 2021
Revised: November 5, 2021
Accepted: January 10, 2022
Published: February 18, 2022

REFERENCES

Barroso, M., Tucker, H., Drake, L., Nichol, K., and Drake, J.R. (2015). Antigen-B cell receptor complexes associate with intracellular major histocompatibility complex (MHC) class II molecules. J. Biol. Chem. *290*, 27101–27112.

Benjamin, D.C., Berzofsky, J.A., East, I.J., Gurd, F.R., Hannum, C., Leach, S.J., Margoliash, E., Michael, J.G., Miller, A., and Prager, E.M. (1984). The antigenic structure of proteins: a reappraisal. Annu. Rev. Immunol. *2*, 67–101.

Berzofsky, J.A. (1983). T-B reciprocity. An Ia-restricted epitope-specific circuit regulating T cell-B cell interaction and antibody specificity. Surv. Immunol. Res. *2*, 223–229.

Berzofsky, J.A., Buckenmeyer, G.K., Hicks, G., Gurd, F.R., Feldmann, R.J., and Minna, J. (1982). Topographic antigenic determinants recognized by monoclonal antibodies to sperm whale myoglobin. J. Biol. Chem. *257*, 3189–3198.

Brumeanu, T.D., Casares, S., Bot, A., Bot, S., and Bona, C.A. (1997). Immunogenicity of a contiguous T-B synthetic epitope of the A/PR/8/34 influenza virus. J. Virol. *71*, 5473–5480.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Cohen, J.I., and Burbelo, P.D. (2020). Reinfection with SARS-CoV-2: implications for vaccines. Clin. Infect. Dis. *73*, e4223–e4228.

Cousens, L.P., Su, Y., McClaine, E., Li, X., Terry, F., Smith, R., Lee, J., Martin, W., Scott, D.W., and De Groot, A.S. (2013). Application of IgG-derived natural treg epitopes (IgG Tregitopes) to antigen-specific tolerance induction in a murine model of type 1 diabetes. J. Diabetes Res. *2013*, 621693.

Cousens, L., Najafian, N., Martin, W.D., and De Groot, A.S. (2014). Tregitope: immunomodulation powerhouse. Hum. Immunol. *75*, 1139–1146.

van Dorp, L., Richard, D., Tan, C.C.S., Shaw, L.P., Acman, M., and Balloux, F. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat. Commun. *11*, 5986.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Fan, G., Wang, Z., Hao, M., and Li, J. (2015). Bispecific antibodies and their applications. J. Hematol. Oncol. *8*, 130.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22.

Galanis, K.A., Nastou, K.C., Papandreou, N.C., Petichakis, G.N., and Iconomidou, V.A. (2019). Linear B-cell epitope prediction: a performance review of currently available methods. BioRxiv. https://doi.org/10.1101/833418.

Greaney, A.J., Loes, A.N., Crawford, K.H., Starr, T.N., Malone, K.D., Chu, H.Y., and Bloom, J.D. (2021). Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe *29*, 463–476.e6.

Guinney, J., Wang, T., Laajala, T.D., Winner, K.K., Bare, J.C., Neto, E.C., Khan, S.A., Peddinti, G., Airola, A., Pahikkala, T., et al. (2017). Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. Lancet Oncol. *18*, 132–142.

Hachim, A., Kavian, N., Cohen, C.A., Chin, A.W.H., Chu, D.K.W., Mok, C.K.P., Tsang, O.T.Y., Yeung, Y.C., Perera, R.A.P.M., Poon, L.L.M., et al. (2020). ORF8 and ORF3b antibodies are accurate serological markers of early and late SARS-CoV-2 infection. Nat. Immunol. *21*, 1293–1301.

Hassan, S.S., Choudhury, P.P., Roy, B., and Jana, S.S. (2020). Missense mutations in SARS-CoV2 genomes from Indian patients. Genomics *112*, 4622–4627.

Haste Andersen, P., Nielsen, M., and Lund, O. (2006). Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. *15*, 2558–2567.

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

Jespersen, M.C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. *45*, W24–W29.

Khare, S., Gurry, C., Freitas, L., Schultz, M.B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R.T., Yeo, W., et al. (2021). Gisaid's role in pandemic response. China CDC Weekly *3*, 1049–1051.

Killerby, M.E., Biggs, H.M., Midgley, C.M., Gerber, S.I., and Watson, J.T. (2020). Middle East respiratory syndrome coronavirus transmission. Emerg. Infect. Dis. *26*, 191–198.

Larsen, J.E., Lund, O., and Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. Immunome Res *2*, 2.

Li, Z.-R.T., Zarnitsyna, V.I., Lowen, A.C., Weissman, D., Koelle, K., Kohlmeier, J.E., and Antia, R. (2019). Why are CD8 T cell epitopes of human influenza A virus conserved? J. Virol. *93*, e01534–e01618.

Mandrekar, J.N. (2010). Receiver operating characteristic curve in diagnostic test assessment. J. Thorac. Oncol. *5*, 1315–1316.

Moss, C.X., Tree, T.I., and Watts, C. (2007). Reconstruction of a pathway of antigen processing and class II MHC peptide capture. EMBO J. *26*, 2137–2147.

Ozaki, S., and Berzofsky, J.A. (1987). Antibody conjugates mimic specific B cell presentation of antigen: relationship between T and B cell specificity. J. Immunol. *138*, 4133–4142.

Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct. Biol. *9*, 51.

Phan, I.Q., Subramanian, S., Kim, D., Murphy, M., Pettie, D., Carter, L., Anishchenko, I., Barrett, L.K., Craig, J., Tillery, L., et al. (2021). In silico detection of SARS-CoV-2 specific B-cell epitopes and validation in ELISA for serological diagnosis of COVID-19. Sci. Rep. *11*, 4290.

Plante, J.A., Liu, Y., Liu, J., Xia, H., Johnson, B.A., Lokugamage, K.G., Zhang, X., Muruato, A.E., Zou, J., Fontes-Garfias, C.R., et al. (2021). Spike mutation D614G alters SARS-CoV-2 fitness. Nature *592*, 116–121.

Ponomarenko, J., Bui, H.-H., Li, W., Fusseder, N., Bourne, P.E., Sette, A., and Peters, B. (2008). ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics *9*, 514.

Racle, J., Michaux, J., Rockinger, G.A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., et al. (2019). Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat. Biotechnol. *37*, 1283–1286.

Ren, R., Zhang, Y., Li, Q., McGoogan, J.M., Feng, Z., Gao, G.F., and Wu, Z. (2021). Asymptomatic SARS-CoV-2 infections among persons entering China from April 16 to october 12, 2020. JAMA *325*, 489–492.

Sabhnani, L., Manocha, M., Sridevi, K., Shashikiran, D., Rayanade, R., and Rao, D.N. (2003). Developing subunit immunogens using B and T cell epitopes and their constructs derived from the F1 antigen of Yersinia pestis using novel delivery vehicles. FEMS Immunol. Med. Microbiol. *38*, 215–229.

Sanchez-Trincado, J.L., Gomez-Perosanz, M., and Reche, P.A. (2017). Fundamentals and methods for T- and B-cell epitope prediction. J. Immunol. Res. *2017*, 2680160.

Saravanan, V., and Gautham, N. (2015). Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor. OMICS *19*, 648–658.

Setliff, I., Shiakolas, A.R., Pilewski, K.A., Murji, A.A., Mapengo, R.E., Janowska, K., Richardson, S., Oosthuysen, C., Raju, N., Ronsard, L., et al. (2019). High-throughput mapping of B cell receptor sequences to antigen specificity. Cell *179*, 1636–1646.e15.

Seyednasrollah, F., Koestler, D.C., Wang, T., Piccolo, S.R., Vega, R., Greiner, R., Fuchs, C., Gofer, E., Kumar, L., Wolfinger, R.D., et al. (2017). A DREAM challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. JCO Clin. Cancer Inform. *1*, 1–15.

Shean, R.C., Makhsous, N., Stoddard, G.D., Lin, M.J., and Greninger, A.L. (2019). VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank. BMC Bioinformatics *20*, 48.

Singh, H., Ansari, H.R., and Raghava, G.P.S. (2013). Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One *8*, e62216.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.

Song, S., Ma, L., Zou, D., Tian, D., Li, C., Zhu, J., Chen, M., Wang, A., Ma, Y., Li, M., et al. (2020). The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. Genomics Proteomics Bioinformatics *18*, 749–759.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., and Kiyotani, K. (2020). SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J. Hum. Genet. *65*, 1075–1082.

Trifonova, O.P., Lokhov, P.G., and Archakov, A.I. (2014). [Metabolic profiling of human blood]. Biomed. Khim. *60*, 281–294.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The immune epitope database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343.

Wang, K., Wei, G., and Liu, D. (2012). CD19: a biomarker for B cell development, lymphoma diagnosis and therapy. Exp. Hematol. Oncol. *1*, 36.

Wang, Y., Song, B., Wang, S., Chen, M., Xie, Y., Xiao, G., Wang, L., and Wang, T. (2021a). De-noising spatial expression profiling data based on in situ position and image information. BioRxiv. https://doi.org/10.1101/2021.11.03.467103.

Wang, Z., Schmidt, F., Weisblum, Y., Muecksch, F., Barnes, C.O., Finkin, S., Schaefer-Babajew, D., Cipolla, M., Gaebler, C., Lieberman, J.A., et al. (2021b). mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. BioRxiv. https://doi.org/10.1101/2021.01.15.426911.

Wu, K., Werner, A.P., Moliva, J.I., Koch, M., Choi, A., Stewart-Jones, G.B.E., Bennett, H., Boyoglu-Barnum, S., Shi, W., Graham, B.S., et al. (2021). mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. BioRxiv. https://doi.org/10.1101/2021.01.25.427948.

Yao, B., Zhang, L., Liang, S., and Zhang, C. (2012). SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. PLoS One *7*, e45152.

Zhang, J., Alam, S.M., Bouton-Verville, H., Chen, Y., Newman, A., Stewart, S., Jaeger, F.H., Montefiori, D.C., Dennison, S.M., Haynes, B.F., et al. (2014). Modulation of nonneutralizing HIV-1 gp41 responses by an MHC-restricted TH epitope overlapping those of membrane proximal external region broadly neutralizing antibodies. J. Immunol. *192*, 1693–1706.

Zhou, C., Chen, Z., Zhang, L., Yan, D., Mao, T., Tang, K., Qiu, T., and Cao, Z. (2019). SEPPA 3.0-enhanced spatial epitope prediction enabling glycoprotein antigens. Nucleic Acids Res. *47*, W388–W394.

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

Zhu, J., Kim, J., Xiao, X., Wang, Y., Luo, D., Jiang, S., Chen, R., Xu, L., Zhang, H., Moise, L., et al. (2020). The immune vulnerability landscape of the 2019 novel coronavirus, SARS-CoV-2. BioRxiv. https://doi.org/10.1101/2020.02.08.939553.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
SARS-CoV-2 reference sequence	NCBI Nucleotide	https://www.ncbi.nlm.nih.gov/nuccore/MN908947
SARS-CoV-2 sequences (NGDC)	Song et al. 2020	Various (https://bigd.big.ac.cn/ncov)
SARS-CoV-2 sequences (GISAID)	Khare et al. 2021	Various (https://www.gisaid.org/)
SARS-CoV reference sequence	NCBI Nucleotide	NC_004718
SARS-CoV sequences	NCBI Nucleotide	Various (https://www.ncbi.nlm.nih.gov/nuccore/?term=txid694009%5BOrganism%3Anoexp%5D+and+complete+genome)
MERS-CoV reference sequence	NCBI Nucleotide	NC_019843
MERS-CoV sequences	NCBI Nucleotide	Various (https://www.ncbi.nlm.nih.gov/nuccore/?term=txid1335626%5BOrganism%3Anoexp%5D+and+complete+genome)
B cell epitope sequences	Vita et al. 2019	Various (https://www.iedb.org/downloader.php?file_name=doc/bcell_full_v3_single_file.zip)
Curated training, validation, and test data for BepiTBR	Vita et al. 2019	Github: https://github.com/zzhu33/BepiTBR/blob/main/data.zip
10X B cell scRNA-seq datasets	10X Inc.	https://www.10xgenomics.com/resources/datasets
10X human lymph node Visium dataset	10X Inc.	Processed data from https://www.biorxiv.org/content/10.1101/2021.11.03.467103v2
Software and algorithms		
Python 3.7.3	Python Software Foundation	https://www.python.org/downloads/release/python-373/
Raku/Perl6 (Rakudo implementation)	Rakudo.org	https://rakudo.org/downloads
Biopython 1.78	Cock et al. 2009	https://github.com/biopython/biopython
R 3.6.3	The R Foundation	https://cran.r-project.org/src/base/R-3/R-3.6.3.tar.gz
R glmnet package	Friedman et al., 2010	https://cran.r-project.org/package=glmnet
R ROCR package	Sing et al., 2005	https://cran.r-project.org/package=ROCR
R umap package	Tomasz Konopka (tokonopka@gmail.com)	https://cran.r-project.org/package=umap
R forestplot package	Max Gordon (max@gforge.se)	https://cran.r-project.org/package=forestplot
R vioplot package	Daniel Adler and S. Thomas Kelly (tomkellygenetics@gmail.com)	https://cran.r-project.org/package=vioplot
BepiPred 1.0	Larsen et al. 2006	https://services.healthtech.dtu.dk/
BepiPred 2.0	Jespersen et al., 2017	https://services.healthtech.dtu.dk/service.php?BepiPred-2.0
NetMHCIIpan 3.2	Jensen et al., 2018	https://services.healthtech.dtu.dk/services/NetMHCIIpan-4.0/9-Downloads.php#
MixMHC2pred	Racle et al., 2019	https://github.com/GfellerLab/MixMHC2pred
Discotope 1.1	Haste Andersen et al. 2006	https://services.healthtech.dtu.dk/services/DiscoTope-2.0/9-Downloads.php#
ElliPro	Ponomarenko et al. 2008	https://tools.iedb.org/ellipro/download/
PyMOL 2.0	Schrödinger, LLC.	https://pymol.org/installers/
DataBase of Actionable Immunology	Tao Wang (Tao.Wang@UTSouthwestern.edu)	https://dbai.biohpc.swmed.edu/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BepiTBR	This paper	https://dbai.biohpc.swmed.edu/bepitbr and https://github.com/zzhu33/BepiTBR
VAPiD	Shean et al. 2019	https://github.com/rcs333/VAPiD
Other		
BioHPC supercomputing facility	Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center	https://portal.biohpc.swmed.edu

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. Tao Wang (Tao.Wang@UTSouthwestern.edu).

Materials availability

This study did not generate new unique reagents

Data and code availability

The training, validation, and test data for BepiTBR are made available at Github: https://github.com/zzhu33/BepiTBR/blob/main/data.zip. We also shared the called mutations and predicted B and HLA class II epitopes for all viral genomes of this study publicly at this link, as a resource for the research community. The 10X B cell scRNA-seq datasets are available from the 10X Genomics website (10X: https://www.10xgenomics.com/resources/datasets), with accession codes: sc5p_v2_hs_PBMC, vdj_nextgem_hs_pbmc3, vdj_v1_hs_pbmc, vdj_v1_hs_pbmc2, and vdj_v1_hs_pbmc3. The structures of the antibody-antigen complexes were exported from PDB with accession codes in PDB: Table S4.

The BepiTBR software codes are available at https://github.com/zzhu33/BepiTBR. A free cloud-based BepiTBR computation service will be provided at the DataBase of Actionable Immunology: https://dbai.biohpc.swmed.edu/.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

B cell epitope data collection

For our training and validation/testing, we mainly considered linear B cell epitope data of the human host from The Immune Epitope Database (IEDB) (Vita et al., 2019). We filtered the B cell epitope data by type (linear epitopes only) and length (peptides greater than 6 a.a.). We did not impose an upper length limit for the B cell epitopes. Only 5% of these B cell epitopes are longer than 20 a.a. in length. Entries corresponding to the same B cell epitopes were combined into a single record. We also only used epitopes that had valid corresponding antigen protein sequences, which is necessary for HLA class II epitope prediction. We note that BepiPred 1.0 and 2.0, as well as LBEEP were also trained on the IEDB data. The goal of our work is to show that HLA class II epitope (putative $CD4^+$ T cell epitope) prediction can enhance the base B epitope prediction model. Namely, we are interested in the relative difference of prediction accuracy, induced by incorporation of HLA class II epitope information. Importantly, the HLA class II epitope information has never been considered in these prior models. Therefore, although some of these models also incorporated data from IEDB in training, the use of IEDB data in our work does not constitute any leak of information or overfitting. However, to remove any concern of overfitting, we divided all IEDB records by the year 2018 (all three B epitope prediction software were published before 2018). The IEDB records before 2018 were split randomly into a training cohort and a validation cohort. The IEDB records after 2018 formed the completely independent test cohort.

Linear and conformational B cell epitope predictions

BepiPred 1.0 and 2.0 were applied to predict B cell epitopes with default settings. These two software algorithms give a confidence score for each amino acid residue of a given peptide sequence. We then took the maximum of the scores for each amino acid within the given peptide sequence as the peptide level confidence score. We found this approach to be very important for improving the performance of the BepiPred 1.0 and 2.0 software for a given peptide sequence. LBEEP directly gives a prediction score for the whole given peptide. It was applied with the following parameters: "-m pep -M C -t 0.001". Conformational epitopes were predicted by ElliPro and discotope using the default options (maximum distance of 6 Å and minimum score of 0.5 for ElliPro, and contact distance of 10.0 Å and threshold of −7.7 for discotope).

HLA class II epitope prediction

For MixMHC2pred, we used the default parameters for prediction of HLA class II epitopes. We considered all HLA class II alleles available for MixMHC2pred. For netMHCIIpan software, we used the "-inptype 0" parameter, and considered the same alleles to match mixMHC2pred for consistency. For both prediction tools, we scanned all 15-mer candidate epitopes in an overlapping moving window from −180a.a. upstream of the middle of the B cell epitope of interest to 180a.a. downstream of the middle of the B cell epitope. DRB, DQ and DP allele binders were counted in non-overlapping bins of 20a.a. within this range. Counts were assigned to bins based on the center positions of the moving windows. When the centers of the bins extend beyond the start or the end of the antigen protein, counts of 0s were assigned to those affected bins.

BepiTBR model

The model considered the confidence score of the B cell epitope from each base B cell epitope prediction software, and the abundance of the HLA class II epitopes ($CD4^+$ T cell epitopes) predicted by either of the two epitope prediction algorithms in each bin. Interaction terms were added between the B cell epitope confidence score and each bin of HLA class II epitope count. Penalty terms were added to only the HLA class II epitope counts and the interaction terms. Overall, the model can be described as:

$$\boldsymbol{\beta} = argmin_{\beta}\left(-\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(p(x_i)\right) + (1-y_i)\log\left(1-p(x_i)\right)\right] + \lambda\alpha\left\|\left(\boldsymbol{\beta}^{tepi}, \boldsymbol{\beta}^{int}\right)\right\|_1 + \frac{\lambda(1-\alpha)}{2}\left\|\left(\boldsymbol{\beta}^{tepi}, \boldsymbol{\beta}^{int}\right)\right\|_2^2\right)$$

where refers to the label of B cell epitope immunogenicity, x refers to the features of B cell epitope confidence scores, HLA class II epitope counts, and their interaction terms (see below), and refers to each of the B cell epitopes in the training set. The total penalty strength is denoted as λ,and the proportion of penalty assigned to L1 (LASSO) is denoted as α.

The link function taken for the regression is:

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \boldsymbol{x}_i\boldsymbol{\beta} + \beta_0.$$

And

$$\boldsymbol{x}_i = \left(x_i^{bepi}, \boldsymbol{x}_i^{tepi}, \boldsymbol{x}_i^{int}\right)$$

denotes the observed data in a row vector format, and

$$\boldsymbol{\beta} = \left(\beta^{bepi}, \boldsymbol{\beta}^{tepi}, \boldsymbol{\beta}^{int}\right)^T$$

denotes the coefficients to be learned, in a column vector format. We implemented the BepiTBR model using the R glmnet package (Friedman et al., 2010). No standardization of the covariates was employed.

In the training phase, as there are many more negative cases from IEDB than positive cases, we randomly subsampled the negative training data points so its sample size is twice that of the positive training data points. We bootstrapped the training dataset 200 times, and fitted the same model (with the same parameters) on the boostrapped training data. The coefficients of these 200 models form a distribution to derive

statistics like CIs. In the prediction phase, the linear predictor (continuous variable) for each epitope is output as the predicted score of that B cell epitope.

Acquisition of the coronavirus genome sequences

The SARS-CoV-2 complete genome sequences and meta data were downloaded from the NGDC (https://bigd.big.ac.cn/ncov) (Song et al., 2020) and GISAID (https://www.gisaid.org/) (Khare et al., 2021) databases. We downloaded a total of 1,959,135 viral sequences before the data lock of June 15th, 2021. Sequences that are the same between the two databases (0.03% of all collected strains) were combined into the same records. Sequences that have any ambiguous amino acids in any coding sequence (49.8%) were removed. We then translated the genomic sequences to proteins, kept only sequences with greater than three appearances, and then removed viral strains with duplicated protein sequences. In order to keep the computation time manageable, for sequences collected after April 2021, we randomly sampled ~1000 unique sequences for each month. We identified a total of 21,917 unique and high quality sequences after all filterings. The reference genome was acquired from NCBI. This sequence is one of the first few isolates of SARS-CoV-2 collected and sequenced in late December of 2019: https://www.ncbi.nlm.nih.gov/nuccore/MN908947.

The complete genome SARS-CoV and MERS-CoV sequences are also downloaded from NCBI: =%%%+++https://www.ncbi.nlm.nih.gov/nuccore/?term=txid694009%5BOrganism%3Anoexp%5D+and+complete+genome and =%%%+++https://www.ncbi.nlm.nih.gov/nuccore/?term=txid1335626%5BOrganism%3Anoexp%5D+and+complete+genome. The reference genomes (SARS-CoV: NC_004,718.3 and MERS-CoV: NC_019,843.3) were determined by NCBI, which were one of the earliest few sampled isolates for each coronavirus, respectively (sometimes only the months of sample collection were shown). Analysis procedures for SARS-CoV and MERS-CoV follow that of SARS-CoV-2.

Mutation calling in SARS-CoV-2 protein sequences

The genome sequences were annotated and translated using a modified version of the VAPiD (Shean et al., 2019) pipeline. Each isolate's protein sequence was then aligned to the reference protein sequence using MUSCLE (Edgar, 2004). In order to process the large number of sequences, alignments were performed in batches of 1,000 sequences for each coding sequence. The reference sequence was included in each batch, and its alignment output was used as the index to combine the results of different alignment batches. Mutations were called based on the alignment output. Silent mutations were ignored as their effects are outside the scope of this study.

Following Wu et al.(Wu et al., 2021), B.1.1.7 refers to the mutation combination of D614G/N501Y/P681H/A570D/T716I/S982A/D1118H/dH69V70Y144, and B.1.351 refers to the combination of D614G/L18F/D80A/R246I/N501Y/K417N/E484K/A701V/dL242A243L244.

QUANTIFICATION AND STATISTICAL ANALYSIS

All computations were performed in the R, Python and Raku programming languages. MCC was implemented by a customized R script, where the cutoff value for determining positive predicted epitopes was determined by the ratio of positive to negative epitopes in the training data. AUC of ROC was implemented by the R ROCR package (Sing et al., 2005). We used the Mann-Whitney U test to test the alternative hypothesis of whether true B cell epitopes have higher or lower densities of MHC class II binders nearby (implemented by the R wilcox.test function). The UMAP analyses of the single cells were performed by the R umap package. Protein 3-dimensional structures were visualized using the PyMOL Molecular Graphics System (Version 2.0 Schrödinger, LLC.). The forest plot was generated by the R forestplot package. Bayesian Factors were calculated by bootstrapping the observed data and counting the number of times one model performed better than the other model, divided by the number of times when this was reversed (Guinney et al., 2017). For all boxplots appearing in this study, box boundaries represent interquartile ranges, whiskers extend to the most extreme data point within 1.5 times the IQR, and the line in the middle of the box represents the median. Violin plots were generated by the R vioplot package.

ADDITIONAL RESOURCES

Free cloud-based BepiTBR computation service: https://dbai.biohpc.swmed.edu/bepitbr/.