DiscoTope-3.0 - Improved B-cell epitope prediction using AlphaFold2 modeling and inverse folding latent representations

Magnus Haraldson Høie[1], Frederik Steensgaard Gade[1], Julie Maria Johansen[1], Charlotte Würtzen[1], Ole Winther[2,3,4], Morten Nielsen[1] and Paolo Marcatili[1†]

[1*]Department of Health Technology, Section for Bioinformatics.
[2*]Section for Cognitive Systems, DTU Compute, Technical University of Denmark.
[3*]Center for Genomic Medicine, Copenhagen University Hospital.
[4*]Department of Biology, Bioinformatics Centre, University of Copenhagen, Denmark.

*Corresponding author: Morten Nielsen, morni@dtu.dk;
†P.M. and M.N. contributed equally to this work.

Abstract

Accurate computational identification of B-cell epitopes is crucial for the development of vaccines, therapies, and diagnostic tools. Structure-based prediction methods generally outperform sequence-based models, but are limited by the availability of experimentally solved structures. Here, we present DiscoTope-3.0, a B-cell epitope prediction tool that exploits inverse folding representations from solved or AlphaFold-predicted structures. On independent datasets, the method demonstrates improved performance on both linear and non-linear epitopes with respect to current state-of-the-art algorithms. Most notably, our tool maintains high predictive performance across solved and predicted structures, alleviating the need for experiments and extending the general applicability of the tool by more than 4 orders of magnitude. DiscoTope-3.0 is available as a web server and downloadable package, processing up to 50 structures per submission. The web server interfaces with RCSB and AlphaFoldDB, enabling large-scale prediction on all currently cataloged proteins. DiscoTope-3.0 is available at: https://services.healthtech.dtu.dk/service.php?DiscoTope-3.0.

Keywords: Structure-based, B-cell epitope prediction, AlphaFold, inverse-folding, ESM-IF1, antigen

1 Introduction

A key mechanism in humoral immunity is the precise binding of B-cell receptors and antibodies to their molecular targets, named antigens. The antigen regions that are involved in the binding are known as B-cell epitopes. B-cell epitopes are found on the surface of antigens, and in the case of proteins they can be classified as linear if the epitope residues are sequentially arranged along the antigen sequence, or discontinuous if they are only proximal in the antigen tertiary structure, but not in the primary structure. Identification of B-cell epitopes has large biotechnological applications, including rational development of vaccines and immunotherapeutics. However, experimental mapping of epitopes remains expensive and resource intensive. Computational tools for B-cell epitope prediction offer a viable alternative to experiments.

However, prediction of B-cell epitopes remains a challenging problem (Galanis et al., 2019; Sun et al., 2019). Historically, in-silico prediction methods have been either antigen sequence- or structure-based. Sequence-based methods such as BepiPred-2.0 (Clifford et al., 2022) are attractive given the high availability of protein sequences. BepiPred-2.0 utilizes a random forest trained on structural features predicted from the antigen sequence, but has limited accuracy and struggles to predict non-linear epitopes (Klausen et al., 2019). In a recent work, BepiPred-3.0 further improves the method, demonstrating large gains by exploiting sequence representations from the protein language model ESM-2 (Lin et al., 2022).

On the other hand, structure-based methods benefit from having direct access to the antigen tertiary structure, and in particular, its surface topology. DiscoTope-2.0 (Kringelum et al., 2012) was published in 2012, and it estimates epitope propensity from the local geometry of each residue, taking into consideration both its solvent accessibility and the direction of its side chain. Newer methods such as SEMA (Shashkova et al., 2022), epitope3D (da Silva et al., 2021), BeTop (Zhao et al., 2012), EPSVR, EPMeta (Liang et al., 2010), and ElliPro (Ponomarenko et al., 2008) have shown marginal improvements in the prediction accuracy. Recently, ScanNet demonstrated a new state of the art with the use of a geometric deep-learning neural network (Tubiana et al., 2022b), explicitly considering geometric details at both the resolution of individual atoms and side-chains. However, while structure-based prediction tools in general outperform sequence-based methods, they are limited by the availability of antigen structures.

Data scarcity affects the accuracy of prediction tools in different ways. Firstly, they constraint the amount of data on which such tools can be trained. As of January 2023, less than 5500 antibody structures are available in the antibody-specific structural database SabDab (Dunbar et al., 2013), which are in complex with an antigen. After filtering this dataset for redundancy, one is left with less than 1500 structures for training, which limits the complexity of the models that can be reliably trained without incurring in overfitting (Clifford et al., 2022).

Secondly, the available data is a biased sampling of the possible antibody-antigen complexes. We see that most of the antigens are found only once in the dataset, while others, because of medical or biological interest, have been resolved in complex with as many as 43 (Dunbar et al., 2013) different antibodies. This means that one cannot confidently annotate negative residues; they might be part of antibody-antigen complexes yet to be solved.

Lastly, undersampling will also result in imprecise assessment of the tools' accuracy; predictions that appear as false positives might just be part of complexes yet to be solved. The last two points are typical of a class of problems known as Positive-Unlabeled (PU) training. In this scenario, we are only confident of positive annotations, while all remaining samples should be treated as unlabeled. Several approaches have been proposed for increasing the accuracy of methods and their estimated metrics in such cases (da Silva et al., 2021; Ren et al., 2015; Li et al., 2021). A simple yet effective strategy is to train ensemble predictors based on bootstrapping of samples in the Unlabelled class (Mordelet and Vert, 2014), also known as PU bagging, which is the approach that we propose in this work.

With recent advances in protein structure prediction, AlphaFold2 (Jumper et al., 2021) has enabled accurate prediction of protein structures directly from sequences. Currently, over 200 million pre-computed structures are available in AlphaFold DB (Varadi et al., 2021), covering every currently cataloged protein in UniProt (Consortium, 2022). The three-dimensional coordinate of the proteins, together with the local quality reported as pLDDT scores, are readily accessible from the database.

To fully exploit such dramatic advancement in the availability of accurate models, we need to create informative yet robust numeric representations of both predicted and solved structures. The ESM-IF1 inverse folding model is an equivariant graph neural network pre-trained to recover native protein sequences from protein backbones structures (Ca, C and N atoms). This model has been shown to outperform sequence-based representations on tasks such as predicting binding affinity and change in stability (Hsu et al., 2022). Crucially, ESM-IF1 is explicitly trained on both solved and predicted structures, enabling large-scale application on AlphaFold predicted structures.

In this work, we train DiscoTope-3.0, a structure-based B-cell epitope prediction tool exploiting inverse folding embeddings generated from either AlphaFold predicted or solved structures. DiscoTope-3.0 is explicitly trained on both predicted and solved antigen structures using an ensemble approach, enabling large-scale prediction of epitopes even when solved structures are unavailable. We compare the impact in performance when using predicted structures versus solved structures, in both cases showing unprecedented accuracy. DiscoTope-3.0 is implemented as a web server and downloadable package interfacing with both RCSB and AlphaFoldDB.

2 Results

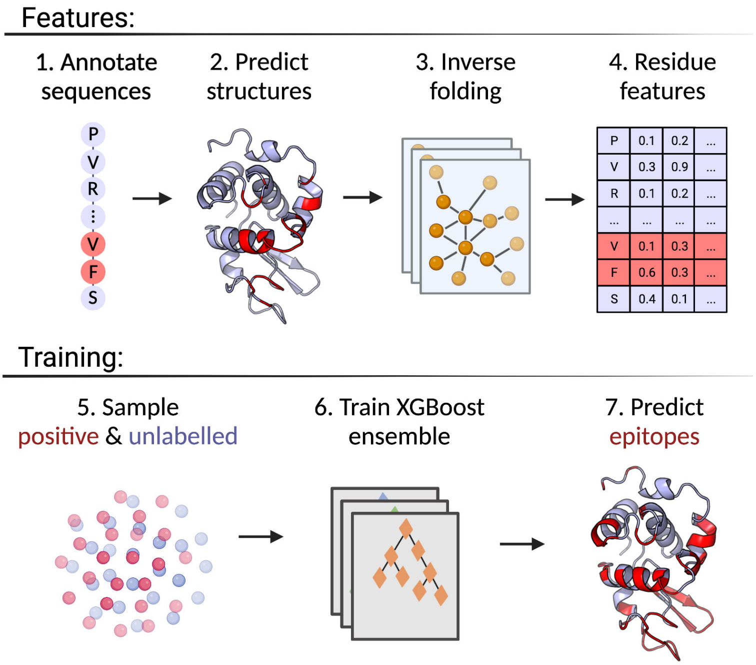


Fig. 1 Overview of the DiscoTope-3.0 method. Created with BioRender.com

The final training strategy for DiscoTope-3.0 is shown in Figure 1. First, epitopes from solved antibody-antigen complexes are mapped onto the antigen sequences (1). Using sequences as input, antigen structures are predicted using AlphaFold2 (2). Next, per-residue structural representations are extracted using the ESM-IF1 protein inverse folding model (3 and 4). During training, random subsets of epitopes and unlabelled residues are sampled across the dataset (5), before finally training an ensemble of XGBoost models on the individual data subsets (6). The final DiscoTope-3.0 score is given as the ensemble average (7).

Here, we present a quick overview of the training procedure. More details are available in the Methods section. DiscoTope-3.0 training and validation is based on the BepiPred-3.0 dataset of 582 antibody-antigens complexes, covering a total of 1466 antigen chains IEDB (Vita et al., 2018). Epitopes are defined as the set of residues within 4 Å of any antibody heavy atom (see Methods). The training is based on 2 different datasets: Training and Validation, while evaluation is performed on the Validation and External Test set. Briefly, we first removed any sequences with more than 20 % similarity to the BepiPred-3.0 test set, resulting in 1406 chains. After clustering the chains at

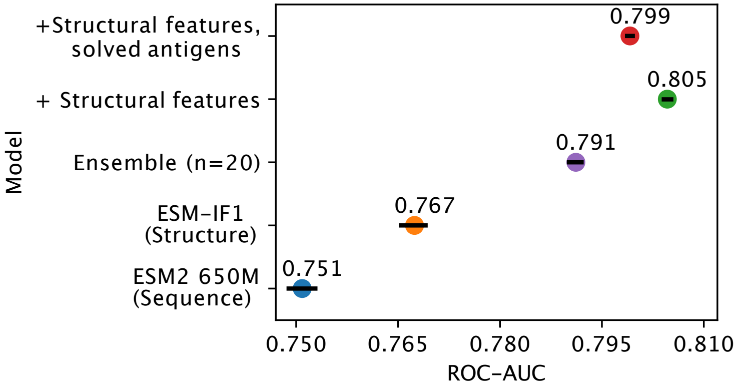


Fig. 2 Effects of inverse folding and bagging. Ablation results on the validation set of AlphaFold predicted structures, with less than 50 % sequence similarity to the training set. The plot reports the AUC-ROC for a single XGboost model trained on representations based on ESM-2 650M parameter (blue) and on ESM-IF1 (orange), for an ensemble of 20 XGboost models based on bootstraps of ESM-IF1 representations (purple), for models where additional structural features are included (see methods) tested on both AlphaFold models (green) and on the corresponding solved structures (red) (see Methods). Error bars indicate 95 % confidence interval.

50 % similarity, we selected 1125 chains for training and 281 for validation. The external test set consists of 24 antigens collected from SAbDab (Dunbar et al., 2013) and PDB (Berman et al., 2000) on October 20, 2022. These antigens share at most 20 % similarity to both our own, BepiPred and ScanNet's training datasets (see Methods).

In addition to using experimentally solved antigens for training, all individual antigen chains were predicted using AlphaFold2. Both the solved and predicted chains were then embedded with ESM-IF1. Further we extract for each residue its relative surface accessibility (RSA), AlphaFold local quality score (pLDDT) as well as the antigen length and a one-hot encoding for the antigen sequence (see Methods). These structural features (or subsets) were used to train an ensemble of XGBoost models and the ensemble average is used as the final prediction score.

Structure-based embeddings have been shown to be a powerful representation in different downstream tasks. To see if this is also the case for B-cell epitope prediction, we evaluated the results obtained using different feature encoding schemes on our validation set of AlphaFold structures (for details on this data set refer to Methods). First, we assess whether training a single XGBoost model using structure representations from predicted structures outperforms a similar model based on the sequence representations from ESM-2 (Figure 2). Here, we observe a marginal but consistent epitope prediction performance using the structure (AUC-ROC 0.767 ± 0.003) vs sequence representations (AUC-ROC 0.751 ± 0.003) ($p < 0.0001$).

As explained in the introduction, the B-cell epitope prediction problem can be categorized in the broad class of PU training. Incorrectly labeled negative examples can negatively affect the training, by introducing frustration in the learning process (Dietterich, 2000). We can observe that, by using an ensemble learning strategy with a dataset bagging approach based on previous works (Huang et al., 2009; Elkan and Noto, 2008; Dietterich, 2000) (see Methods), we can further improve performance (AUC-ROC 0.791) and generalization.

2.1 Effect of using predicted versus solved structures

One of the risks in training models on either solved structures or AlphaFold predictions is that the methods might over-specialize to one source and perform significantly worse on the other, or even be affected by data leakage. For example, a model may learn to identify conformational changes in the side chains of epitope residues in solved structures, remaining after the binding antibody is removed.

By training on both predicted and solved structures, we obtain a final model which performs well on both structure types, with an AUC-ROC 0.799 for predicted structures (Figure 2), and 0.807 when predicting solved structures (Supplementary Figure S1). We note that training separate models, namely using only solved or only predicted structures, does indeed improve performance slightly when tested on the same class (AUC-ROC 0.813 and 0.805 respectively), but comes at added complexity. To simplify comparison with other tools, we therefore chose the DiscoTope-3.0 version trained on both structure types for further analysis.

2.2 Benchmark comparison to state-of-the-art methods

To further test the effect of using predicted versus solved structures, we used the external test set of 24 antigens. These antigens share at most 20 % sequence similarity to both our own, BepiPredm and ScanNet's training datasets (see Methods). The three tools use the same definition for epitope residues, thus ensuring a fair comparison.

The precision and recall scores of the tools were calculated on this test set, both on experimentally solved structures and their AlphaFold predicted counterparts for DiscoTope-3.0 and ScanNet. As a reference, we also benchmark BepiPred-3.0, which is purely sequence-based and independent of the different structural variations, and a naïve predictor using relative surface accessibility as its score.

The results of this evaluation are displayed in Figure 3. Here, DiscoTope-3.0 outperforms ScanNet, both on predicted (AUC-PR 0.232 ± 0.020 vs 0.127 ±0.011) and solved structures (0.223 ± 0.018 vs 0.157 ± 0.012). Our tool performance is largely unaffected by the type of structures used for prediction. To further test the robustness of the tools to minor differences in the antigen structures, we performed an energy minimization on the solved structures using the software FoldX (Schymkowitz et al., 2005). This minimization only impacts the side chain, thus leaving the backbone of the native structure unaltered. The ESM-IF model does not use the side chain atoms in its predictions, and

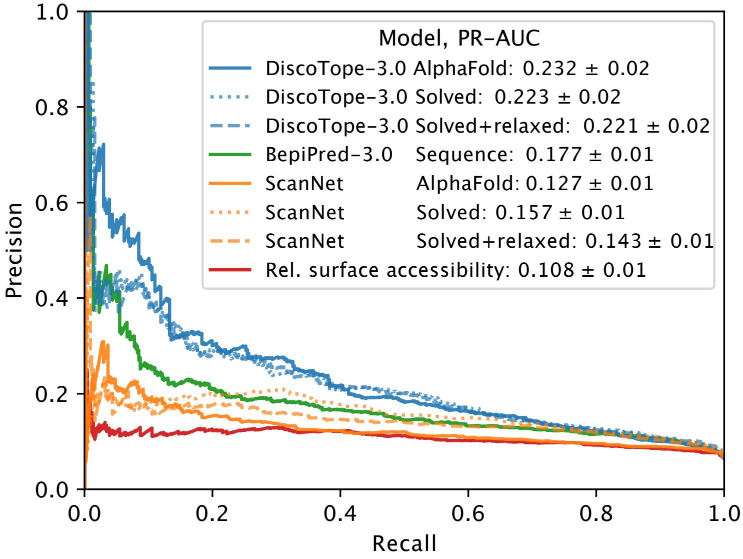


Fig. 3 **Improved performance on solved and predicted structures**. AUC-PR curve plots for DiscoTope-3.0 (blue), BepiPred-3.0 (green) and ScanNet (orange) on the external test set of 24 antigen chains, at most 20 % similar to the training set of all models. Structures provided as AlphaFold predicted, experimentally solved, or sequence in the case of BepiPred-3.0. Standard deviation calculated from bootstrapping 1000 times (see Methods).

consequently our tool should not be affected by the relaxation process. Indeed, we see that the performance of DiscoTope-3.0 is almost identical when shifting between solved (0.223), relaxed (0.221) and modeled (0.232) structures. On the other hand, ScanNet, despite being structure-based, was outperformed by the sequence-based BepiPred-3.0 in the case of predicted structures (0.177 ± 0.015). It demonstrates a large drop in performance when shifting from solved to predicted structures (0.157 vs 0.127), and a smaller yet significant drop on relaxed structures (0.143). Similar trends were observed on the external test set as evaluated with AUC-ROC (Supplementary Figure SS2).

A typical real-case scenario would be for users to submit individual antigens, and then to analyze the top scoring epitope residues, regardless of the specific score. The epitope rank score is a metric highly relevant for this case. It analyzes what is the rank of an epitope residue predicted score when compared to all the scores for the same antigen. Here, an epitope rank score of 70 % would mean that epitopes on average score in the upper 70th percentile. Using this metric, DiscoTope-3.0 consistently outperforms both ScanNet and BepiPred-3.0 in the case of predicted structures, while performing similarly to ScanNet on solved structures (Figure SS3)). We observe that after side-chain relaxation in solved structures, ScanNet's epitope rank scores are reduced by ∼3.1 percentile points, while swapping solved for predicted structures leads to

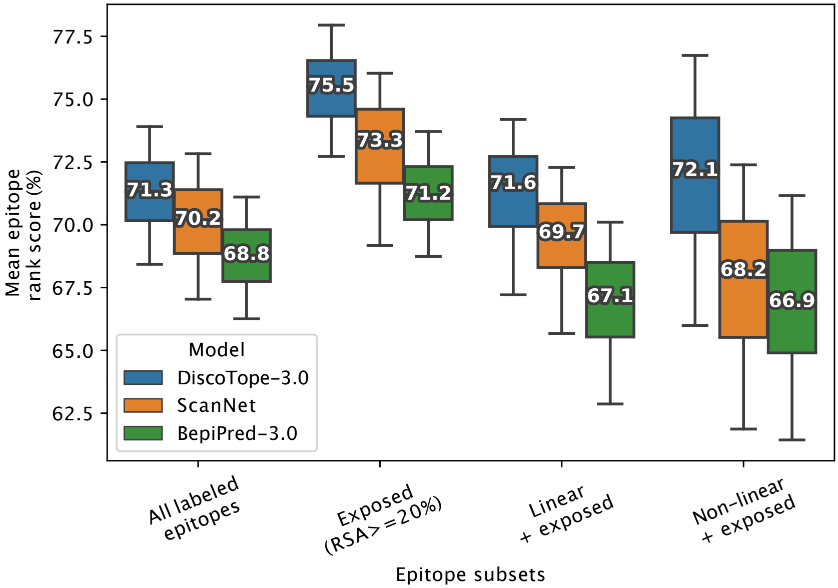


Fig. 4 Improved performance on linear and non-linear epitopes. External test set mean epitope rank scores across the following epitope subsets: All labeled epitopes, Exposed (RSA > 20 % relative surface accessibility, Exposed Linear epitopes and exposed Non-linear epitopes (see text and Methods). Mean values calculated after bootstrapping 1000 times, with error bars indicating 95 % confidence interval.

a loss of ∼7.5 percentile points (see Methods). In contrast to this, DiscoTope-3.0 only lost ∼0.1 and ∼0.6 percentile points respectively, again indicating robustness to the modeling process (Supplementary Figure S3, S4).

These observations can be at least in part explained by how structural features are processed by the two models. DiscoTope-3.0's structure representations are generated from the protein backbone using ESM-IF1, making it robust to the quality of side-chain modeling. ScanNet, on the other hand, explicitly processes side-chain atomic coordinates. This may suggest a loss of signal if side-chains are perturbed away from their bound conformation, or structures are modeled in an unbound state.

2.3 Improved prediction on exposed and non-linear epitopes

We also investigated if the structural information available to DiscoTope-3.0 and ScanNet affects the prediction of different kinds of epitopes. To this aim, epitopes were split into different sub-categories (Exposed, Buried, Linear and Non-linear). Exposed and Buried epitope residues are defined depending on whether their relative surface accessibility was above or below 20 %, respectively. Linear epitopes are defined as any group of 3 or more epitope residues found sequentially along the antigen sequence, allowing for a possible gap of up to 1 unlabeled residue in between. Finally non-linear epitopes were defined as epitopes not satisfying the conditions of the linear group.

The result of this performance evaluation in the external test set reveals a better performance of DiscoTope-3.0 across all epitope subsets (Figure 4). DiscoTope-3.0 performance is remarkably good for non-linear epitopes, where ScanNet performs similarly to BepiPred-3.0. In the case of buried epitopes (relative surface accessibility ¡20 %), all models score poorly in the 30-37th percentile. This low performance is likely an artifact of the epitope labeling process (shared between all tools), erroneously labeling buried residues in the structural proximity of an exposed epitope residue as part of an epitope patch, even though they are not directly involved in molecular interactions with the antibody.

2.4 Effect of predicted structural quality

Next, we investigate how the quality of the AlphaFold predicted structures affects the scores of exposed epitopes. Overall, lower structural quality leads to a decrease in predictive performance (Figure 5). High quality structures (pLDDT 95-100) have a mean epitope rank score of 84.2 %. As the structural quality gets worse, the score also decreases to 81.2 and 75.5 % for antigens with pLDDTs between 85-95 and 60-85, respectively (Figure 5A). Fitting a linear regression model for pLDDT versus mean epitope rank score, we observe a 5 percentile point drop in epitope scores for every 10 point decrease in pLDDT (Figure 5B), but it's worth noting that over 94 % of antigens exceed a quality of pLDDT 80.

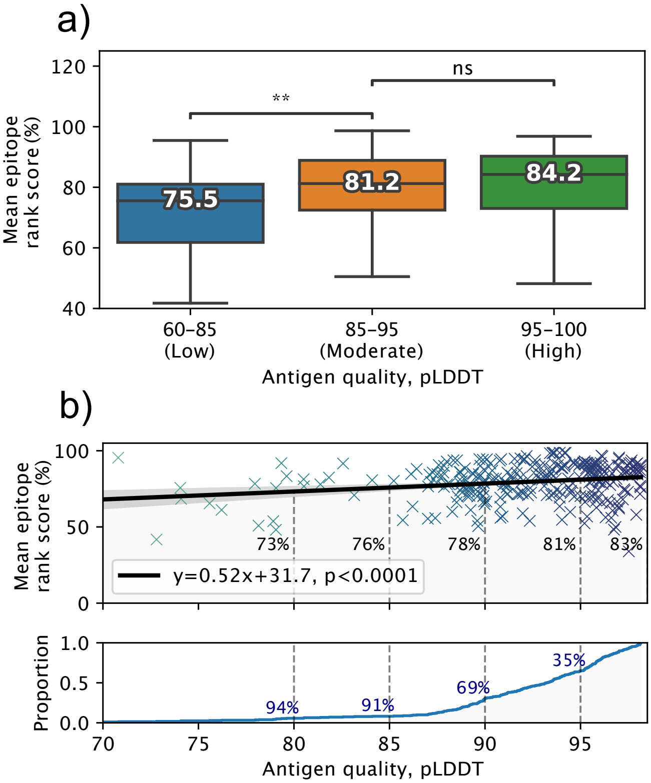


Fig. 5 Effects of predicted structural quality. Validation set performance on AlphaFold predicted antigens dependent on predicted structural quality, excluding buried epitopes. a) Epitope rank score distribution for antigens split into increasing quality bins of mean antigen pLDDT 60-85, 85-95 and 95-100. Median value for each distribution is shown, with paired one-tailed t-test comparison (** = p < 0.005). b) Mean antigen pLDDT versus mean epitope rank score, with a fitted linear model shown in black. Below, cumulative distribution of mean antigen pLDDT, with a 91 % proportion exceeding a pLDDT of 85, and 35 % exceeding 95 respectively.

2.5 Lysozyme case study with collapsed epitopes

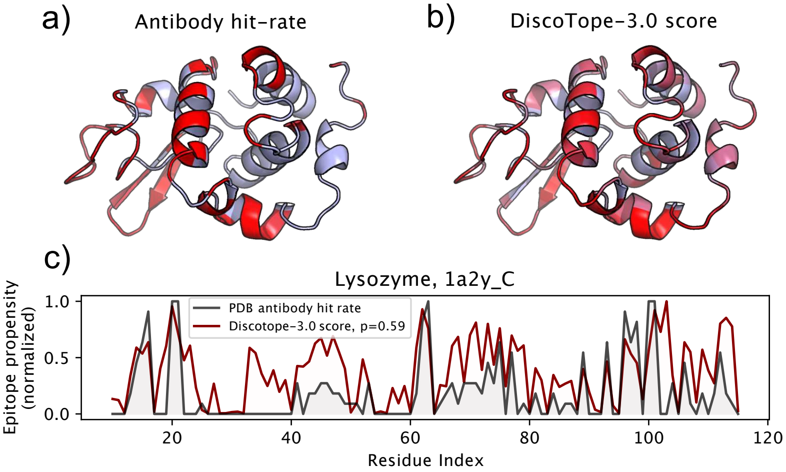


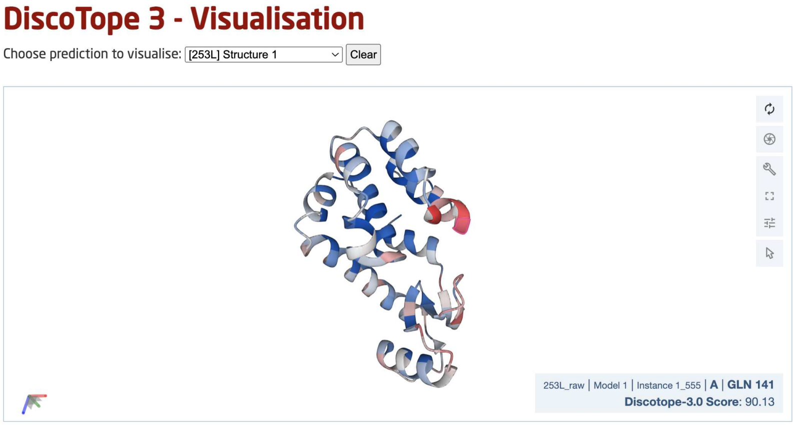
Fig. 6 DiscoTope-3.0 score significantly correlates with antibody hit rate. Lysozyme epitope propensity as predicted by DiscoTope-3.0, excluding all lysozyme chains from training. Increasing epitope propensity shown in red. a) PDB antibody hit rate mapped AlphaFold predicted structure (chain 1a2y_C) b) DiscoTope-3.0 score c) Epitope propensity visualized across the PDB amino-acid sequence. Normalized DiscoTope-3.0 score and PDB epitope counts shown, as measured from aligning the 12 epitope mapped lysozyme sequences (Spearman R = 0.59).

As a noteworthy test case, we evaluate the performance of DiscoTope-3.0 on lysozyme, a well-studied antigen extensively mapped against different antibodies. First, we identified all lysozyme chains at 90 % similarity to the chain C of the PDB structure 1A2Y, finding 12 chains that are in complex with antibodies. Next, DiscoTope-3.0 was re-trained excluding these chains. Finally, we compared DiscoTope's epitope propensity with the residue PDB antibody hit rate, calculated as the relative ratio of observed epitopes for each residue across all the aligned sequences. This analysis is shown in Figure 6.

The results showed that increasing DiscoTope-3.0 scores for a residue were associated with a higher probability of observing a bound epitope, as indicated by a Spearman correlation coefficient of 0.59 and a significant linear trend (Supplementary Figure SS5). Additionally, there was a significant improvement in the performance of the model, as measured by the PR-AUC and ROC-AUC, when most of the antigen surface had been mapped for epitope binding. This emphasizes the underestimation of performance caused by under labeling of epitope residues for most antigen structures.

We note that the residues at positions ~30-40 (Figure 6) score highly in DiscoTope but have no observed epitopes. Upon further investigation into the IEDB database, we found this region to be part of a discontinuous epitope patch (including K31, R32, G34, D36, G37, G40 ...) bound by a camelid antibody deposited under the PDB id 4I0C.

2.6 DiscoTope-3.0 web server



DiscoTope 3 - Visualisation

Choose prediction to visualise: [253L] Structure 1

253L_raw | Model 1 | Instance 1_555 | A | GLN 141
Discotope-3.0 Score: 90.13

Fig. 7 **DiscoTope-3.0 web server interface**. The web server provides an interactive 3D view for each predicted protein structure. DiscoTope-3.0 score on an example PDB, with increasing epitope propensity from blue to red. DiscoTope-3.0 is accessible at: https://services.healthtech.dtu.dk/service.php?DiscoTope-3.0

The DiscoTope-3.0 web server allows for rapidly predicting epitopes on either AlphaFold2 predicted or solved structures. The web server accepts batches of up to 50 structures at a time. Users may upload structures directly as PDB files, or automatically fetch existing structures submitted as lists of RCSB or AlphaFoldDB IDs. Output predictions are easily visualized through an interactive 3D view directly on the web server using Molstar (Sehnal et al., 2021). Predictions may be downloaded in both a CSV and PDB format.

3 Discussion

In this work, we present the DiscoTope- 3.0 method for prediction of B-cell epitopes. The method exploits structural information obtained from AlphaFold predicted and/or experimentally solved antigen structures, utilizing informative representations extracted by the ESM-IF1 inverse folding model. Extensive benchmarking of the tool demonstrated state of the art performance on both solved and predicted structures. Importantly this performance, in contrast to earlier proposed structure-based models, was found to be maintained when shifting to predicted and relaxed structures. This observation is of critical importance since it removes the need for experimentally solved structures imposed in current structure-based models and allows for predicted structures to be applied for accurate B-cell epitope predictions. This extends the applicability of the tool from the few thousand antigens for which a solved structure is available, to potentially any pathogen protein that can be structurally modeled using AlphaFold-2.0.

Our tool displays a remarkable robustness when applied to structural models, and it is, to the best of our knowledge, the first tool that presents highly accurate results (>0.75 ROC-AUC) on protein structural models. Unlike other structure-based methods such as ScanNet, we find that structure representations based on the protein backbone are robust towards minor modifications such as relaxation of the input structure side chains, and only marginally affected by the quality of the structural models.

Finally, DiscoTope-3.0 interfaces with AlphaFoldDB and RCSB, enabling rapid batch processing across all currently cataloged proteins in UniProt and deposited solved structures. In the case of predicted structures, confidence in predictions can be assessed by the antigen predicted quality. The web server is made freely available for academic use, accepting up to 50 input structures at a time.

Our tool has been trained and evaluated on individual antigen chains. One could envision that, for multimeric antigen structures, it would be possible to further increase the tool performance by training and testing on the antigen complex. At this time, AlphaFold2 modeling accuracy for complexes is not yet on par with its accuracy on individual chains, and predicted complexes are not yet available in the AlphaFoldDB. As the science and technology behind the structural modeling progresses, it will be likely possible to further improve B-cell epitope predictions.

On the other hand, the PU learning strategy based on ensemble and bagging we use displays a remarkable boost in performance. We can imagine that, given the large dimension of the potential antibody space, the large gap between potential and observed epitopes will not be easily filled. An alternative strategy, that could circumvent this problem and provide valuable information to users, would be to perform antibody-specific epitope predictions. This approach has been tested by us and others in the past (Jespersen et al., 2019; Krawczyk et al., 2014), but the results are yet to provide a significant improvement in accuracy.

In summary, DiscoTope-3.0 is the first structure-based B-cell epitope prediction model that accepts and maintains state-of-the-art predictive power on predicted antigen structures. We believe this advance will serve as an important aid for the community in the quest for novel rational methods for the design of novel immunotherapeutics.

4 Data and code availability

DiscoTope-3.0 web server, downloadable package and training datasets are freely available for academic use.

- Web server DTU: https://services.healthtech.dtu.dk/service.php?DiscoTope-3.0
- Web server Biolib: https://biolib.com/DTU/DiscoTope-3/
- Code availability: https://github.com/Magnushhoie/discotope3_web

5 Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Author contributions

MH, PM, MN contributed to conception and design of the study. MH, FG, JJ, CW implemented the methodology and software. FG, MH implemented the web server. MH, JJ, CW performed the statistical analysis and visualization of results. MH, MN, PM wrote sections of the manuscript. PM, MN, OW provided supervision. All authors contributed to manuscript revision, read, and approved the submitted version.

7 Funding

This work was in part funded by National Institute of Allergy and Infectious Diseases (NIAID), under award number 75N93019C00001. M.H.H. acknowledges the Sino-Danish Center [2021]. Funding for open access charge: Internal Funding from the University. Conflict of interest statement. None declared.

8 Methods

8.1 Training and evaluation of DiscoTope-3.0

The antigen training dataset as presented in BepiPred-3.0 was used as the starting point for our work. The dataset consists of 582 AbAg crystal structures from the PDB, filtered for a minimum resolution of 3.0 Å and R-factor 0.3. Epitopes are defined as any antigen residue containing at least 1 heavy atom within 4 Å of an antibody heavy atom. From this dataset, using the tool MMseqs2, we first remove any sequences with more than 20 % sequence identity to the BepiPred-3.0 test set, resulting in 1406 chains. Next, the antigen sequences are clustered at 50 % sequence identity. Each cluster has then been selected to be part of the validation (281 chains) or the training set (1125 chains).

In the ablation study, single XGBoost models (Chen and Guestrin, 2016) with default parameters were trained using representations from either the predicted structure or antigen sequence respectively. When testing feature combinations, ensemble size and effect of training on solved and predicted structures, error bars were estimated from re-training 20 times.

Hyperparameters for the XGBoost models were manually optimized using the validation dataset. We used the final parameters n_estimators=200, max_depth=4, learning_rate=0.3 and subsample=0.50. The gpu_hist tree method was used for faster training on a GPU.

8.2 Dataset bagging and ensemble training

When sampling residues for each model in the ensemble, we randomly select 70 % of available observed epitopes (positives) across the training dataset, then sample unlabelled residues (negatives) with a ratio of 5:2. When using both predicted and solved structures, these were sampled at a 1:1 ratio.

Ensembles were constructed by iteratively training independently trained XGBoost models on the randomly sampled datasets. When training an ensemble, we set a different random seed each time.

8.3 Feature calculation and data filtering

Each isolated chain was processed as a single PDB file with ESM-IF1, extracting for each residue its latent representation from the ESM-IF1 encoder output. pLDDT values were either extracted from the PDB files in the case of AlphaFold structures, or set to 100 for solved structures. In the case of training on both solved and predicted structures, we include a binary input feature set to 1 if the input is an AlphaFold2 model, and 0 for solved structures.

Residue solvent accessible surface area was calculated using the Shrake-Rupley algorithm using Biotite (Kunzmann and Hamacher, 2018), with default settings, and converted to relative surface accessibility using the Sander and Rost 1994 (Rost and Sander, 1994) scale as available in Biopython (Cock et al., 2009).

When training DiscoTope-3.0, we removed any antigen with less than 5 or more than 75 epitope residues, as well as PDBs with a mean pLDDT score below 85. Residues with a pLDDT below 75. No data filtering was performed during evaluation on the validation and external test datasets.

8.4 External test set evaluation

The external test set, used for comparing our tool to ScanNet and BepiPred-3.0, consists of antigens deposited in SAbDab and the PDB after April 2021. Any antigen with more than 20 % sequence identity to the training datasets used in this work, in ScanNet, or in BepiPred-3.0 were removed. We annotated epitopes using the same approach as in BepiPred-3.0, which is common to all the tools.

We submitted either solved or AlphaFold2 predicted structures to the ScanNet web server (Tubiana et al., 2022a), using the antibody-antigen binding mode and otherwise default parameters. BepiPred-3.0 predictions were generated from its online web server using the antigen sequence and default parameters.

When evaluating DiscoTope-3.0 on the external test set, we retrained the final model with an ensemble size of 100, on the full training and validation sets, including both solved and predicted chains at a 1:1 ratio.

8.5 AlphaFold2 modeling and structural relaxation

Sequences for each antigen chain containing at least 1 epitope were extracted and modeled with the ColabFold implementation of AlphaFold2 at default settings. We picked the top ranking PDB after 5 independent iterations of 3 recycles, as ranked by AlphaFold2's internal quality measure.

For relaxation of the solved structures we used the foldx_20221231 version of FoldX, with the RepairPDB command for relaxing residues with bad torsion angles, van der Waals clashes or high total energy.

8.6 Data analysis

To calculate the mean epitope rank score, the predicted residue scores for an antigen were first ranked in ascending order. Next, we calculated the average of the rank scores for all epitope residues.

Exposed epitopes were defined as all epitopes with a relative surface accessibility exceeding 20 %, while the remaining epitopes were defined as buried.

When reported, significance testing was performed with a one-sided paired t-test using scipy.stats.ttest_rel (Virtanen et al., 2020). The linear model on the mean antigen pLDDT vs mean epitope rank scores was fitted using scipy.stats.linregress with default parameters.

For confidence estimation with bootstrapping, the dataset was sampled fully with replacement 1000 times, with the bootstrapped datasets used to calculate means, epitope rank scores and standard deviation values.

References

Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. 2000, 01. The Protein Data Bank. *Nucleic Acids Research 28*(1): 235–242. https://doi.org/10.1093/nar/28.1.235 .

Chen, T. and C. Guestrin 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 785–794. ACM.

Clifford, J.N., M.H. Høie, S. Deleuran, B. Peters, M. Nielsen, and P. Marcatili. 2022. Bepipred-3.0: Improved b-cell epitope prediction using protein language models. *Protein Science 31*(12): e4497. https://doi.org/doi:10.1002/pro.449 .

Cock, P.J., T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and et al. 2009. Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics 25*(11): 1422–1423. https://doi.org/10.1093/bioinformatics/btp163 .

Consortium, T.U. 2022, 11. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research 51*(D1): D523–D531. https://doi.org/10.1093/nar/gkac1052 .

da Silva, B.M., Y. Myung, D.B. Ascher, and D.E.V. Pires. 2021. epitope3d: a machine learning method for conformational b-cell epitope prediction. *Briefings in Bioinformatics 23*(1). https://doi.org/https://doi.org/10.1093/bib/bbab423 .

Dietterich, T.G. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning 40*(2): 139–157. https://doi.org/10.1023/a:1007607513941 .

Dunbar, J., K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C.M. Deane. 2013. Sabdab: The structural antibody database. *Nucleic Acids Research 42*(D1). https://doi.org/10.1093/nar/gkt1043 .

Elkan, C. and K. Noto. 2008. Learning classifiers from only positive and unlabeled data. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. https://doi.org/10.1145/1401890.1401920 .

Galanis, K.A., K.C. Nastou, N.C. Papandreou, G.N. Petichakis, D.G. Pigis, and V.A. Iconomidou. 2019. Linear b-cell epitope prediction for in silico vaccine design: A performance review of methods available via command-line interface. https://doi.org/10.1101/833418 .

Hsu, C., R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives. 2022. Learning inverse folding from millions of predicted structures. *bioRxiv*. https://doi.org/10.1101/2022.04.10.487779 .

Huang, F., G. Xie, and R. Xiao. 2009. Research on ensemble learning. *2009 International Conference on Artificial Intelligence and Computational Intelligence* 3: 249–252. https://doi.org/10.1109/aici.2009.235 .

Jespersen, M.C., S. Mahajan, B. Peters, M. Nielsen, and P. Marcatili. 2019. Antibody specific b-cell epitope predictions: Leveraging information from antibody-antigen protein complexes. *Frontiers in Immunology* 10. https://doi.org/10.3389/fimmu.2019.00298 .

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. 2021. Highly accurate protein structure prediction with alphafold. *Nature 596*(7873). https://doi.org/10.1038/s41586-021-03819-2 .

Klausen, M.S., M.C. Jespersen, H. Nielsen, K.K. Jensen, V.I. Jurtz, C.K. Sønderby, M.O.A. Sommer, O. Winther, M. Nielsen, B. Petersen, and P. Marcatili. 2019. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics 87*(6). https://doi.org/10.1002/prot.25674 .

Krawczyk, K., X. Liu, T. Baker, J. Shi, and C.M. Deane. 2014. Improving b-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics 30*(16): 2288–2294. https://doi.org/10.1093/bioinformatics/btu190 .

Kringelum, J.V., O.L. Claus Lundegaard, and M. Nielsen. 2012. Reliable b cell epitope predictions: Impacts of method development and improved benchmarking. *PLoS Computational Biology 8*(12). https://doi.org/10.1371/journal.pcbi.1002829 .

Kunzmann, P. and K. Hamacher. 2018. Biotite: A unifying open source computational biology framework in python. *BMC Bioinformatics 19*(1). https://doi.org/10.1186/s12859-018-2367-z .

Li, F., S. Dong, A. Leier, M. Han, X. Guo, J. Xu, X. Wang, S. Pan, C. Jia, Y. Zhang, G.I. Webb, L.J.M. Coin, C. Li, and J. Song. 2021, 11. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Briefings in Bioinformatics 23*(1). https://doi.org/10.1093/bib/bbab461. https://arxiv.org/abs/https://academic.oup.com/bib/article-pdf/23/1/bbab461/42231477/bbab461.pdf .

Liang, S., D. Zheng, D.M. Standley, B. Yao, M. Zacharias, and C. Zhang. 2010. Epsvr and epmeta: prediction of antigenic epitopes using support vector regression and multiple server results. *BMC Bioinformatics 11*(1). https://doi.org/10.1186/1471-2105-11-381 .

Lin, Z., H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. 2022. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*. https://doi.org/10.1101/2022.07.20.500902 .

Mordelet, F. and J.P. Vert. 2014. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters* 37: 201–209. https://doi.org/https://doi.org/10.1016/j.patrec.2013.06.010 .

Ponomarenko, J., H.H. Bui, W. Li, N. Fusseder, P.E. Bourne, A. Sette, and B. Peters. 2008. Ellipro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics 9*(1). https://doi.org/10.1186/1471-2105-9-514 .

Ren, J., Q. Liu, J. Ellis, and J. Li. 2015. Positive-unlabeled learning for the prediction of conformational b-cell epitopes. *BMC Bioinformatics 16*(S18). https://doi.org/10.1186/1471-2105-16-s18-s12 .

Rost, B. and C. Sander. 1994. Conservation and prediction of solvent accessibility in protein families. *Proteins: Structure, Function, and Genetics 20*(3). https://doi.org/10.1002/prot.340200303 .

Schymkowitz, J., J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. 2005, 07. The FoldX web server: an online force field. *Nucleic Acids Research 33*(suppl_2): W382–W388. https://doi.org/10.1093/nar/gki387 .

Sehnal, D., S. Bittrich, M. Deshpande, R. Svobodová, K. Berka, V. Bazgier, S. Velankar, S.K. Burley, J. Koča, and A.S. Rose. 2021, 05. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research 49*(W1): W431–W437. https://doi.org/10.1093/nar/gkab314 .

Shashkova, T.I., D. Umerenkov, M. Salnikov, P.V. Strashnov, A.V. Konstantinova, I. Lebed, D.N. Shcherbinin, M.N. Asatryan, O.L. Kardymon, and N.V. Ivanisenko. 2022. Sema: Antigen b-cell conformational epitope prediction using deep transfer learning. *Frontiers in Immunology* 13. https://doi.org/10.3389/fimmu.2022.960985 .

Sun, P., S. Guo, J. Sun, L. Tan, C. Lu, and Z. Ma. 2019. Advances in in-silico b-cell epitope prediction. *Current Topics in Medicinal Chemistry 19*(2): 105–115. https://doi.org/10.2174/1568026619666181130111827 .

Tubiana, J., D. Schneidman-Duhovny, and H.J. Wolfson. 2022a. Scannet: A web server for structure-based prediction of protein binding sites with geometric deep learning. *Journal of Molecular Biology 434*(19): 167758. https://doi.org/https://doi.org/10.1016/j.jmb.2022.167758 .

Tubiana, J., D. Schneidman-Duhovny, and H.J. Wolfson. 2022b. Scannet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods 19*(6). https://doi.org/10.1038/s41592-022-01490-7 .

Varadi, M., S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, and S. Velankar. 2021, 11. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research 50*(D1): D439–D444. https://doi.org/10.1093/nar/gkab1061 .

Virtanen, P., R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, and et al. 2020. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods 17*(3): 261–272. https://doi.org/10.1038/s41592-019-0686-2 .

Vita, R., S. Mahajan, J.A. Overton, S.K. Dhanda, S. Martini, J.R. Cantrell, D.K. Wheeler, A. Sette, and B. Peters. 2018. The immune epitope database (iedb): 2018 update. *Nucleic Acids Research 47*(D1). https://doi.org/10.1093/nar/gky1006 .

Zhao, L., L. Wong, L. Lu, S.C. Hoi, and J. Li. 2012. B-cell epitope prediction through a graph model. *BMC Bioinformatics 13*(S17). https://doi.org/10.1186/1471-2105-13-s17-s20 .

Supplementary Figures

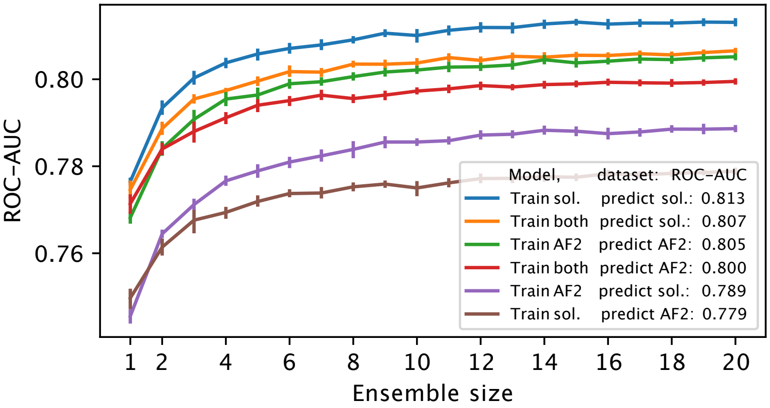


Fig. S1 Effect of ensemble size. Validation set gain in AUC-ROC from ensembling the full-feature model. Performance graphs are shown for training on either experimentally solved, AlphaFold predicted or both structures, and then evaluated on either the solved or predicted structure validation set.

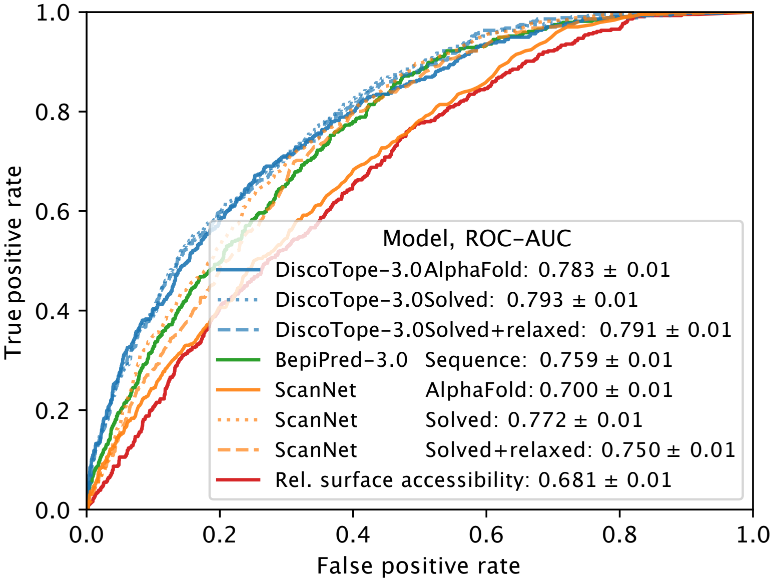


Fig. S2 External test set AUC-ROC. Test set AUC-ROC, as evaluated on 24 antigens modeled with AlphaFold. For PR-AUC see Figure 3.

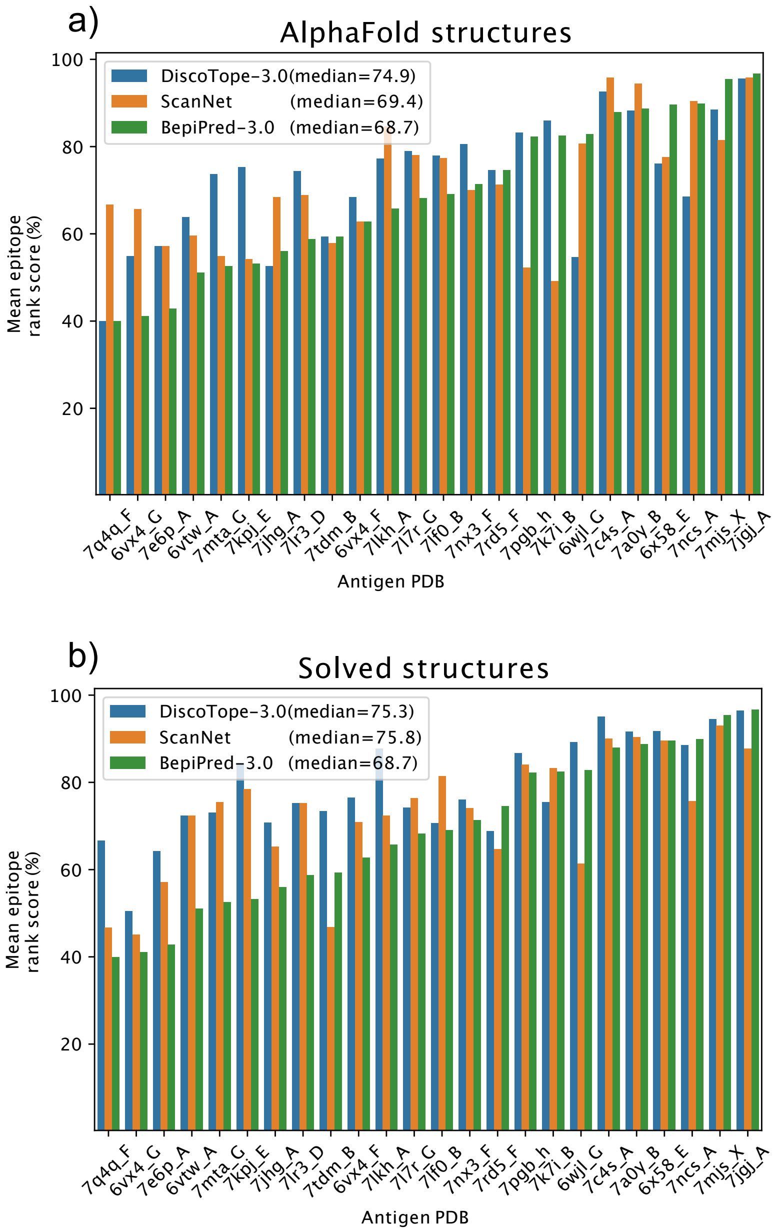


Fig. S3 External test set PDB performances. Evaluation on 24 antigens modeled with AlphaFold (left) or experimentally solved structures (right). BepiPred-3.0 performances on antigen sequences only.

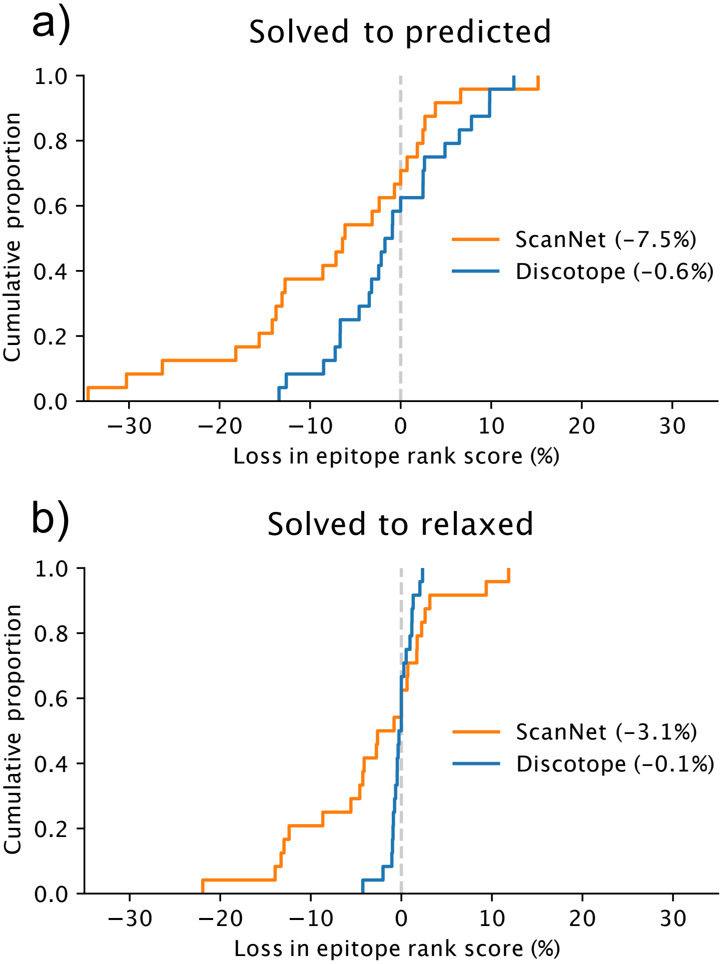


Fig. S4 DiscoTope-3.0 is robust towards modeling and relaxation. External test set change in mean epitope rank scores across PDBs, when (a) swapping predicted structures with their original solved structure or (b) solved structures with the same structure after FoldX relaxation (see Methods). Mean performance loss shown in percent.

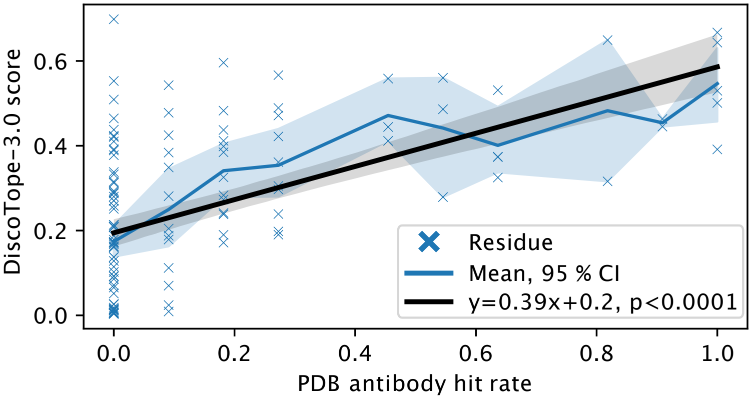


Fig. S5 DiscoTope-3.0 score significantly correlates with antibody hit rate. Lysozyme case study on 1a2y_C, showing PDB antibody hit rate versus DiscoTope-3.0 score. Model is trained excluding all lysozyme structures from training (see Methods).