Briefings in Bioinformatics, 23(1), 2022, 1–8

https://doi.org/10.1093/bib/bbab423
Problem Solving Protocol

epitope3D: a machine learning method for conformational B-cell epitope prediction

Bruna Moreira da Silva, YooChan Myung, David B. Ascher and Douglas E. V. Pires

Corresponding authors: David B. Ascher. Structural Biology and Bioinformatics, Department of Biochemistry, University of Melbourne, Melbourne, Victoria, 3052, Australia; Systems and Computational Biology, Bio21 Institute, University of Melbourne,Melbourne, Victoria, 3052, Australia; Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, 3052, Australia; Baker Department of Cardiometabolic Health, University of Melbourne, Melbourne, Victoria, 3052, Australia; Department of Biochemistry, University of Cambridge, 80 Tennis Ct Rd, Cambridge CB2 1GA, UK, Tel.: +61 90354794; E-mail: david.ascher@unimelb.edu.au; Douglas E. V. Pires. Systems and Computational Biology, Bio21 Institute, University of Melbourne,Melbourne, Victoria, 3052, Australia; Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, 3052, Australia; School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria, 3053, Australia, Tel.: +61 3 83448185; E-mail: douglas.pires@unimelb.edu.au

Abstract

The ability to identify antigenic determinants of pathogens, or epitopes, is fundamental to guide rational vaccine development and immunotherapies, which are particularly relevant for rapid pandemic response. A range of computational tools has been developed over the past two decades to assist in epitope prediction; however, they have presented limited performance and generalization, particularly for the identification of conformational B-cell epitopes. Here, we present epitope3D, a novel scalable machine learning method capable of accurately identifying conformational epitopes trained and evaluated on the largest curated epitope data set to date. Our method uses the concept of graph-based signatures to model epitope and non-epitope regions as graphs and extract distance patterns that are used as evidence to train and test predictive models. We show epitope3D outperforms available alternative approaches, achieving Mathew's Correlation Coefficient and F1-scores of 0.55 and 0.57 on cross-validation and 0.45 and 0.36 during independent blind tests, respectively.

Key words: conformational epitope; machine learning; graph-based signatures

Introduction

B-cells are an essential part of the adaptive immune system that provides long-term protection against pathogens and harmful molecules through their specific B-cell receptors, known as immunoglobulins or antibodies [1, 2]. This recognition is mediated by binding of antibodies (Ab) to a specific region of the antigen known as epitope. Most B-cell epitopes are discontinuous, which has made their identification challenging as they are often composed of residues that may be far apart in the sequence, but spatially co-located within the protein structure [3, 4].

Accurate identification of B-cell epitopes is crucial for disease control, diagnostics and vaccine development but, in general, experimental approaches are expensive, time-consuming and low throughput [5–7]. A number of sequence and structural computational approaches have been proposed, primarily to identify which residues are likely to be part of an epitope, but have been shown to be of limited predictive power [8–19]. This is, in part, a limitation of the data used to develop them, a natural imbalance

Bruna Moreira da Silva is a PhD student at The University of Melbourne. Her research interests are in digital health, machine learning and bioinformatics.
YooChan Myung is a PhD student at The University of Melbourne with interest in bioinformatics, machine learning and antibody design.
David B. Ascher is Head of the Structural Biology and Bioinformatics Group at Bio21 Institute and the Computational Biology and Clinical Informatics Group at the Baker Institute. His interests are in harnessing of medical and biological data to improve global health outcomes.
Douglas E.V. Pires is a senior lecturer in Digital Health with the School of Computing and Information Systems at the University of Melbourne. His research interests include computational biology, machine learning and the development of the next generation of bioinformatics tools.
Submitted: 7 June 2021; **Received (in revised form):** 25 August 2021

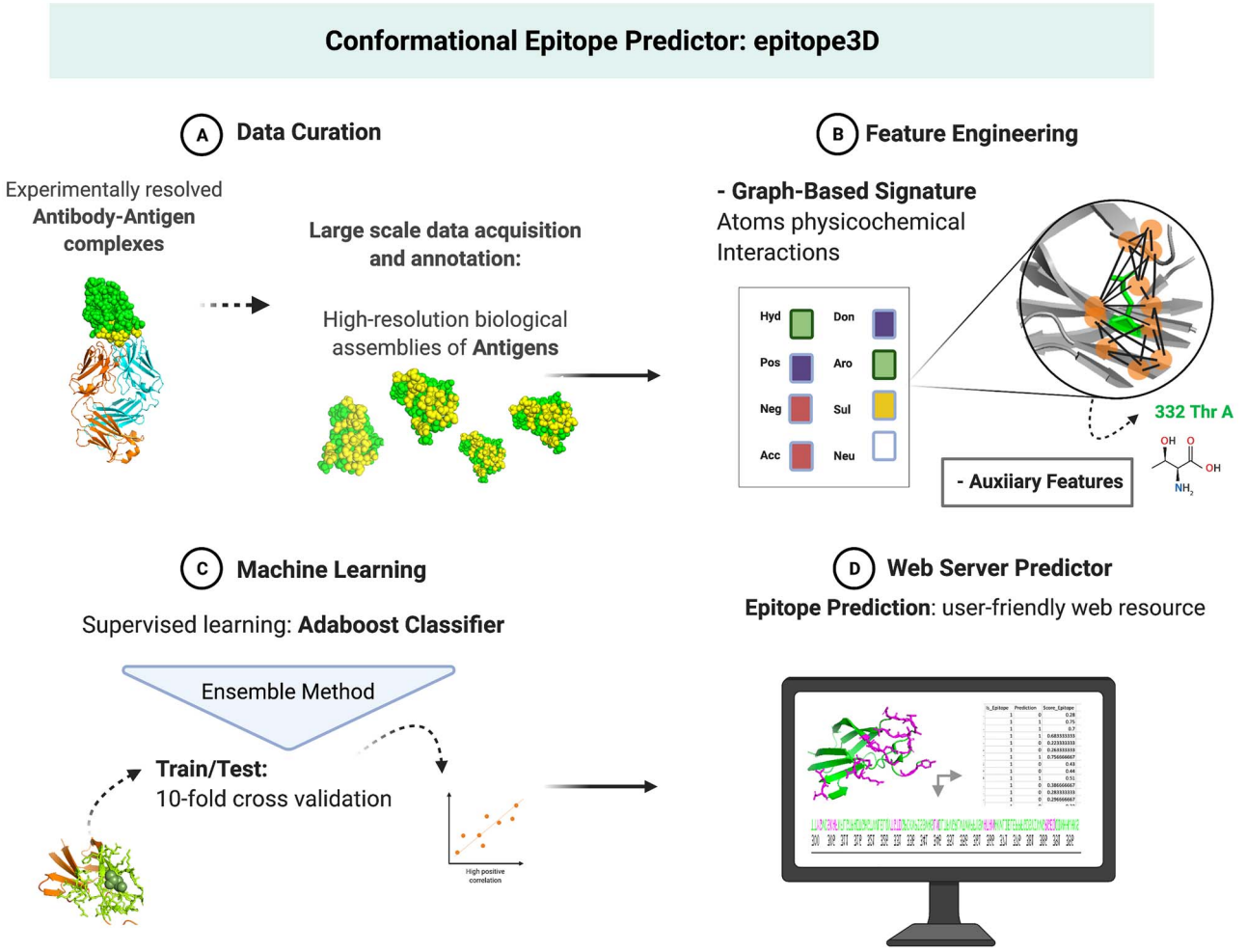

Figure 1. Overview of the epitope3D pipeline composed of four main stages. In the first stage (**A**), data acquisition and curation are performed. High-resolution biological assemblies of Ab-antigen complexes are collected and unbound antigen structure candidates selected via sequence and structure similarity, rendering a data set of 245 non-redundant structures. Two main feature classes are then calculated for the identified epitope and non-epitope residues of the data set, including graph-based signatures and complementary features describing the residue environment (**B**). Calculated features are then used to train and test predictive models via supervised learning (**C**), as a classification task and the best performing model employed to build a user-friendly web interface (**D**).

ratio of epitope and non-epitope residues present in an antigen structure and the challenges of identifying features capable of distinguishing them.

In order to fill this gap and tackle the main hurdles in epitope prediction, we developed epitope3D, a new machine learning method and user-friendly web resource trained and validated on the largest conformational epitope data set collected to date. epitope3D uses the concept of graph-based structural signatures [20–22] to better model and distinguish epitope from non-epitope regions.

Materials and methods

The development of epitope3D can be divided in four main stages: as depicted in Figure 1A, data collection and curation to identify unbound antigen structures based on experimental antibody–antigen complexes; Figure 1B feature engineering, encompassing data modelling and feature calculation using the curated data, to extract characteristics of both epitope and non-epitope residues; Figure 1C machine learning and assessment, involving qualitative data analysis of selected features, training, evaluation and optimization of predictive models and Figure 1D development of a web server and API to allow convenient access to the predictive model and seamless integration into analytical pipelines.

Data collection and curation

Data collection was performed to identify unbound-state structures of antigens, using bound structures as a reference, as done previously [23]. This approach defines epitope residues in the unbound antigen protein based on experimentally resolved antibody–antigen complexes published in Protein Data Bank (PDB) [24]. The central idea of this approach is that different epitope regions from related antigens bound to antibodies can be extracted and aggregated on the same antigen, reducing false negative annotation. A new large-scale data set of conformational epitopes was collected and curated following two major steps: bound structure identification and unbound-state structure acquisition and annotation.

The first step retrieves all biological assemblies from the PDB database with a resolution higher than or equal to 3 Å. Next, antibody–antigen complexes are identified using the ANARCI tool [25] and antigen chains with at least 25 residues are retained.

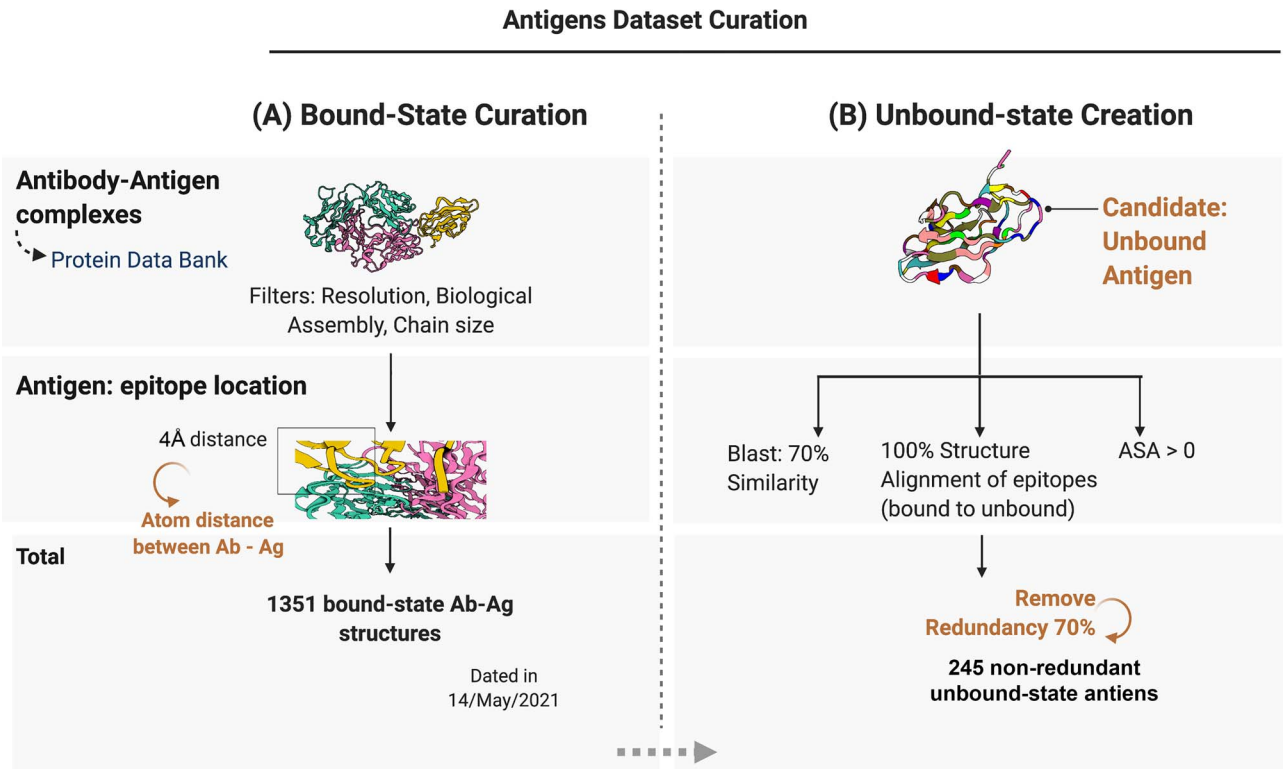

Figure 2. Two-stage dataset curation process: (**A**) Curation of a bound-state antibody–antigen followed by (**B**) unbound-state antigen candidates used to train and test the machine learning model.

Epitope residues in the antigen molecule are determined and annotated based on a cutoff distance criteria, i.e. any antigen residue with at least one heavy atom within 4 Å distance of an antibody residue will be considered an epitope residue. In this step, 1351 antibody–antigen structures were identified as of May/2021, covering 40 842 epitope residues.

After identifying epitope residues in the selected set of bound structures, these are used to propagate epitope annotation to a new set of unbound antigen structures (candidate antigens), sharing at least 70% sequence similarity with any antigen previously identified in a bound structure. In total, 443 candidate antigen structures were identified based on sequence similarity search via blastp [26]. Epitopes from bound structures are then structurally aligned with candidate antigens, using the Pymol library [27], and epitope annotation is transferred for the mapped residues only if 100% of structure alignment is achieved. To further ensure that aligned residues truly belong to epitopes, these are required to be exposed to solvent, measured by a relative solvent accessibility (RSA) larger than zero.

This step resulted in 343 antigen structures, which after being clustered with CD-HIT [28] using a 70% similarity cutoff, generated a non-redundant data set of 245 unbound antigen structures, comprising 168 739 data points, with 53.82% of surface residues and 3.56% epitope residues. The distinction between surface and buried residues considered here was based on the RSA threshold of 15% [29] and with the purpose of optimizing the epitope identification; buried residues were disregarded from the data. This comprises the largest curated conformational epitope collection to date. The overall filtering process is depicted in Figure 2.

The non-redundant epitope data set was divided into three groups: 180 structures for training, 20 for internal testing and 45 structures used as an external blind test. The training and testing set have an imbalance ratio of 1:29, imposing an extreme hurdle into the supervised learning stage. To overcome it, two approaches were assessed to achieve the optimal learning level for training: randomly under sampling the majority class (non-epitope residues) and synthetic oversampling the minority class (epitope residues) using SMOTE [30] available in the *imblearn* Python toolbox [31]. This step was repeated 10 times to guarantee the impartiality of the random selection of the non-epitope class. The optimal performance arises by applying both approaches combined: first randomly under-sampling the majority class data points till imbalance reduces to 1:8, then using SMOTE technique to synthetically create data points of the minority class, leading to 4:8 distribution, or 1:2. Therefore, the final set of 180 structures used to train the classifier contains 50 036 residues, in which 33.33% are epitopes.

As an internal blind-test to assess the machine learning classifier, the 20 structures set was employed following the same class distribution used in the training stage, 1:2, but here only a random under-sampling technique was adopted to achieve this desired imbalance ratio for a fair comparison.

Conversely, the 45-structure set used as a non-redundant external blind-test only disregards the buried residues based on RSA threshold, as described before, resulting in a 1:13 imbalance ratio. Data sets used are available as Supplementary Materials (Tables S1–S3, and online at http://biosig.unimelb.edu.au/epitope3d/data.

Feature engineering

In order to investigate properties capable of distinguishing epitopes from other protein regions, a range of different features were calculated and assessed. These can be divided into two main categories: (i) graph-based signatures and (ii) complementary properties.

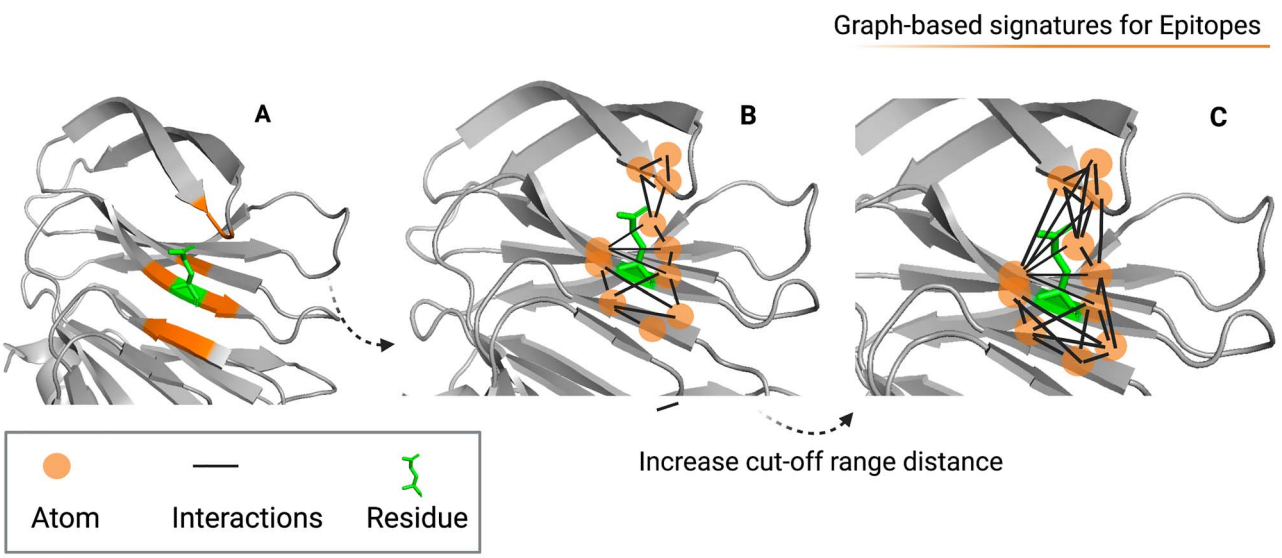

Figure 3. The concept of graph-based signatures applied for epitope and non-epitope residues. (**A**) A residue is represented as a green stick and the chosen environment that surrounds it is painted in orange colour. (**B**) The atoms within the environment are depicted as orange balls and the interaction amid them are the black edges. (**C**) Increasing the cut-off distance within the same environment, more interactions are seen between atoms.

Graph-based signatures

Graphs are versatile and powerful mathematical abstractions that have been widely and successfully used to model biological entities and their relationships. We have pioneered the concept of graph-based signatures to model physicochemical properties from small-molecules [32–37] to macromolecules [21, 22, 38]. The main idea behind this modelling technique involves modelling atoms as nodes labelled based on their physicochemical properties (or pharmacophores) and their interactions as edges. Based on this network representation, distance-based patterns between different atom types are extracted as a cumulative distribution function [21, 22, 39], which is used as evidence to train and test predictive models via supervised learning. This approach has been successfully employed in a range of applications, including the development of methods to predict the effects of mutations on protein structure, function and interactions [20, 39–44]. This concept has also been recently applied to the study of mutations on Ab-antigen interfaces [34, 45, 46]. Here, we adapted the concept of graph-based signatures to represent geometry and chemical composition of the environment surrounding epitope and non-epitope residues, as depicted in Figure 3.

Complementary features

Seven widely used additional feature classes were investigated. The first category, proposed in the present work, is the amino acid composition using a radius scanning matrix. Considering that epitope regions are enriched with particular residue types [8, 11], the ratio of each amino acid in epitope and surface regions was measured from the whole data set and stored as a propensity dictionary. This process is displayed in Figure S1. Next, considering an input structure, each residue is taken as a central point and its neighbours are scanned from a starting distance of 3–15 Å, considering an incremental step of 1 Å. The goal is to describe the residue neighbours in terms of residue composition for different distances using the dictionary described above, computing four statistical metrics: average, maximum, minimum and standard deviation. This generates a 52-value vector. Figure S2, depicts how this feature is calculated.

The remain six classes were as follows:

- Relative surface area (RSA) was calculated to measure how exposed each residue was in the protein structure.
- Secondary structure, using DSSP program [47] to designate the correspondent secondary structure annotation per residue, which were abbreviated into helix, sheet and turn.
- Disorder composition using IUPred2A and ANCHOR2 [48].
- Position-specific scoring matrix [49] to model how conserved over evolutionary time epitope and also non-epitope residues are.
- B-factor score extracted from PDB file for experimental structures. This indicates the atom's mobility. The higher its value, the more flexible a region would be.
- Physicochemical and biochemical amino acid properties were assessed from the AAindex database [50].

Machine learning methods

Different supervised learning algorithms were evaluated using the scikit-learn Python toolkit [51], including Multi-layer Perceptron, Support Vector Machines, K-Nearest Neighbour, Adaboost, Gaussian Processes, Random Forest, Gradient Boost, XGBoost and Extra Trees. Predictive models were evaluated under regular and stratified 5-, 10- and 20-fold cross-validation, with 20 repetitions and the best performing model selected based on well-established evaluation metrics including Matthew's Correlation Coefficient (MCC), Area under the ROC Curve (AUC) and F1-score. Model generalization was also assessed using a low-redundancy, independent blind test. In order to reduce model complexity, reduce noise and optimize predictive performance, an incremental stepwise greedy feature selection approach was performed [42, 45].

Results

Exploring epitope and non-epitope properties

We set out to identify properties that could help differentiate an epitope from a non-epitope residue. Exploring the complete curated data set of 245 antigen structures we analyzed, for

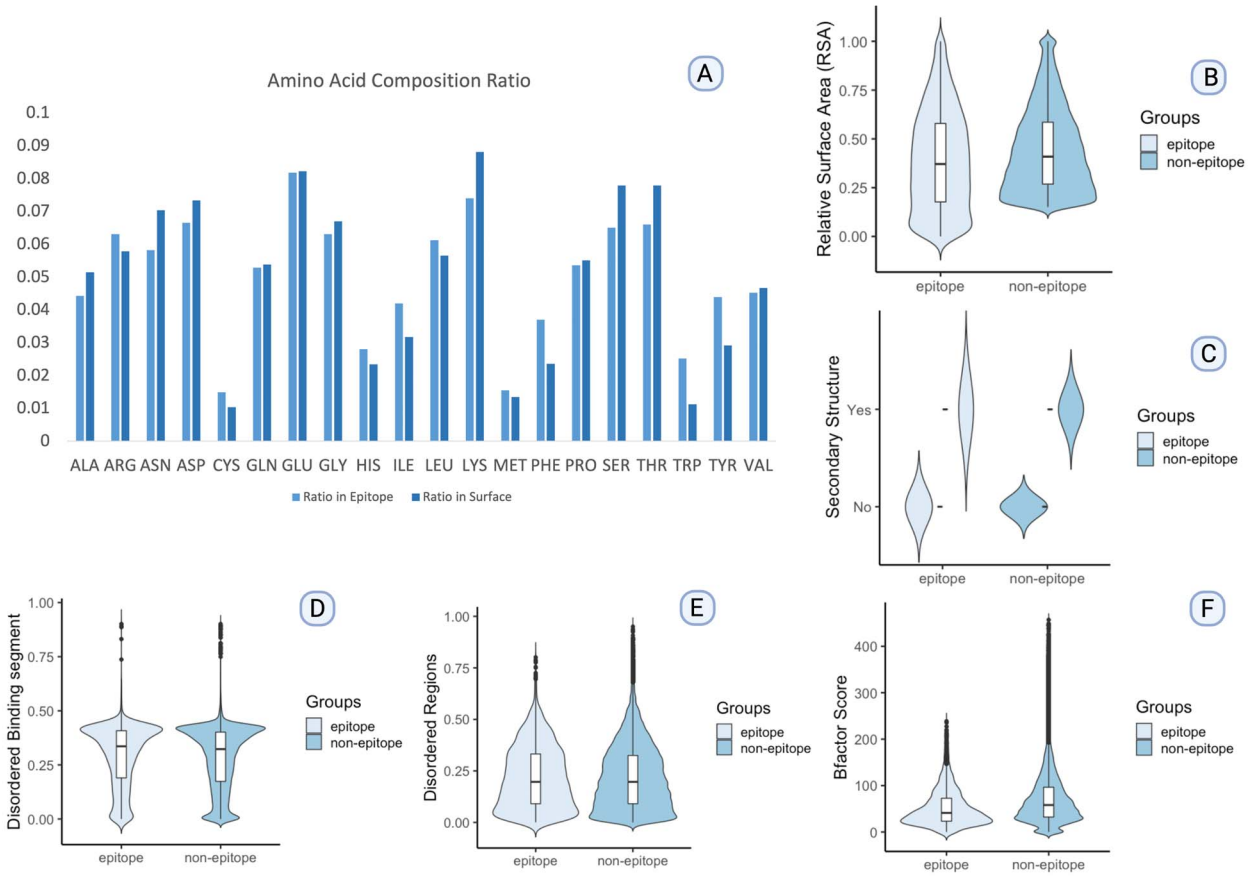

Figure 4. Analysis of the main features for epitope and non-epitope groups. In (**A**), the amino acid composition ratio depicts that Isoleucine (ILE), Phenylalanine (PHE), Tryptophan (TRP) and Tyrosine (TYR) are enriched in Epitope Regions (Surface). (**B**) The Violin plot for the RSA. In (**C**), the calculated secondary structure, if available, is mapped and grouped in Helix, Sheet or Turn. (**D**) The probability of being part of a Disordered binding segment and (**E**) the probability of belonging to a Disordered Region. To conclude, in (**F**), we present the B-factor score for both epitope and non-epitope groups. In addition to the violin plots, the non-parametric Wilcoxon Rank Sum Test was also implemented to statistically verify differences between groups from panel (**B**)–(**F**).

epitope and non-epitope residues, the distribution of the Amino Acid composition, RSA, Secondary structure, the probability of being part of a Disordered segment and a Disordered region, and B-factor.

The Shapiro–Wilk Test was applied to confirm that the set is not normally distributed (P-value < 0.001), and as independent groups, epitope and non-epitope samples were also submitted to the Wilcoxon Rank Sum Test to demonstrate the difference amongst them, resulting in P-values of $2.2e^{-16}$ (RSA), $5.7e^{-09}$ (Secondary structure), $2.3e^{-09}$ (Disordered binding segment) and $2.2e^{-16}$ (B-factor), suggesting distinctive medians, supporting that the alternative hypothesis is to be considered. Analyzing the composition of the extremes of the distributions of these properties (above or below the median $\pm$ standard deviation), showed that residues contributing to these differences, for both epitope and non-epitope groups, are distributed across all structures in the data set. On the other hand, for the Disordered region feature, a higher P-value of 0.425 demonstrates no significant difference among the two groups.

While investigating these characteristics, Figure 4, the amino acid composition ratio among epitope and surface regions shows that Lysine (LYS), Phenylalanine (PHE), Tryptophan (TRP) and Tyrosine (TYR) have the larger difference between Epitope with Surface regions on the overall data set, with the three last ones together with Isoleucine (ILE) being enriched in epitope regions.

Predicting conformational epitopes

The best performing predictive model was obtained using the Adaboost classifier, reaching an AUC and MCC of 0.78 and 0.56, respectively, under 10-fold cross-validation with only four features selected via greedy feature selection. This performance was consistent on blind tests, where epitope 3D achieved an AUC and MCC of 0.59 and 0.35, respectively. The method displayed a consistent prediction performance over multiple 10-fold cross validation repetitions, presenting a low standard deviation (<0.005). Table 1 includes further metrics describing the prediction results and in Supplementary Data; Table S4, shows the results of stratified 5-, 10- and 20-fold cross validation assessment, which were consistent with the results described above, further demonstrating the robustness of the method.

The four selected features are depicted in Figure S3. The Graph-based signature of neutral atoms within 4 Å, the Amino Acid index KARS160102 which represents the Number of edges (size of the graph-theoretic model of single point mutations), the Minimum ratio value of amino acid between Epitope-Surface regions from a 12 Å radius distance and RSA, all presenting significant differences between epitope and non-epitope classes (P-values of: 0.001, $2.5e^{-16}$, 0.01 and $2.2e^{-16}$, respectively). The relative importance of the features contributing to the Adaboost

Table 1. Performance of epitope3D on a complex-based 10-fold cross validation using the training set of 180 structures and with the test set of 20 structures with the same distribution class of 1:2. The metrics presented in the table are: MCC, F1 score (F1), Balanced accuracy (BACC) and AUC. *TP = true positives; *TN = true negatives; *FP = false positives; *FN = false negatives

	MCC	F1	BACC	AUC	TP*	TN*	FP*	FN*
10-fold cv	0.55	0.57	0.70	0.78	7152	36 015	9	10 860
Blind test	0.35	0.30	0.59	0.59	104	1184	0	488

classifier was assessed via Gini importance and is depicted in Figure S4. The radius scanning feature was identified as the most important attribute is listed as the top important, followed by the amino acid index, the graph-based signature and RSA.

Independent method evaluation with a non-redundant blind test

As more structural data and new tools become available, there is a need for an independent benchmark set to allow an impartial assessment of the predictive capacity of the methods aiming to predict conformational epitopes. Benchmark data sets used by previous approaches [8, 12, 52, 53], however, were either no longer available or were used during method development, making them unfit for independent assessment. To fill this gap, we compiled a large non-redundant unbound state antigen set composed of 45 diverse structures, which we propose to be used as a standard benchmark for future developments. This was used here to evaluate other available methods in comparison with epitope3D. The structures were pre-processed removing the buried residues based on RSA value, following the same procedure described in Methods, resulting in 12 230 data points in which 912 are epitopes (7,45%), which gives an imbalance ratio of 1:13.

In order to compare our tool in an independent blind test prediction, we have compared the performance of epitope3D with the following B-Cell epitope prediction tools: SEPPA 3.0 [54], BepiPred-2.0 [55], Discotope-2.0 [56] and ElliPro [18]. While SEPPA 3.0 explores the influence of glycosylation in antigen surface patches, inferring that antibody may prefer to bind in N-glycosylation sites, BepiPred-2.0 analyzed the residues in terms of hydrophobicity and polarity measurements, besides their volume, RSA and predicted secondary structure. Discotope-2.0 considers residue contact counts, in addition to the RSA and amino acid composition around a residue vicinity, calculating log-odds ratios between epitope and non-epitope residues. In a different structural approach, ElliPro characterized the antigen protein by approximating it to an ellipsoid and calculated the residue's protrusion index in order to cluster them. Table S5, summarizes the features used by previous methods.

The comparative results are displayed in Table 2. epitope3D significantly outperformed all alternative approaches in all presented metrics, reaching a MCC of 0.45 and F1 of 0.36, demonstrating robustness even in a scenario presenting severe class imbalance. Additionally, an ROC curve is presented in Figure S5, with epitope3D obtaining the highest amongst all selected predictors.

Conformational analysis

To assess how protein conformational states might impact epitope prediction, we have curated a second version of the blind test set, based on the same Antibody–Antigen complexes as the 45 independent set, but selecting a different unbound candidate (we were able to select 38 new structures, described in Table S6). The structural differences between the new antigen structures compared with their pair from the original 45-set were measured by Root Mean Square Deviation calculated by the align command in Pymol (average ~ 10 Å). No significant performance difference was observed, with epitope3D achieving an MCC of 0.47 and F1 score of 0.38 (in comparison with an MCC of 0.45 and F1 score of 0.36 in the original blind test), demonstrating the robustness of the method.

Table 2. Performance comparison using the independent blind test (45 structures) with 4 B-Cell epitope predictors: SEPPA 3.0, Discotope-2.0, ElliPro and BepiPred-2.0

Method	F1	MCC	BACC
SEPPA 3.0	0.14	0.02	0.52
Discotope-2.0	0.11	-0.01	0.50
ElliPro	0.11	-0.06	0.44
BepiPred-2.0	0.15	0.04	0.55
epitope3D	0.36	0.45	0.61

epitope3D web server

To facilitate method usage, a web server was developed with an intuitive interface utilizing Bootstrap version 4.1.3 as the front-end framework, taking advantage of its CSS and JavaScript elements, and Flask version 1.0.2 as the back-end framework. The user is able to input either the PDB code or upload their structure. In the Result's page, as depicted in Figure S6, a Prediction Table listing the predicted epitope residues is shown, which can also be observed via an interactive 3D viewer implemented via NGL [57]. Predictions are available to download as a comma-separated file (csv). Users can also submit jobs to epitope 3D via an Application Programming Interface (API) described at http://biosig.unimelb.edu.au/epitope3d/api.

Conclusions

In this study, we present epitope3D, a new conformational epitope prediction tool, that leverages the concept of graph-based signatures, trained on the largest curated database to date. We show that epitope3D outperforms similar methods using different independent blind tests. To further contribute to benchmarking of newly developed methods, we have also curated and released a non-redundant and independent blind test set of 45 unbound antigens which will facilitate future performance comparison between models. epitope3D is freely available in an easy-to-use web interface and API at http://biosig.unimelb.edu.au/epitope3d, and we believe that it will be an invaluable tool to assist and guide vaccine design and immunotherapy developments.

Key Points

- A novel scalable machine learning method (epitope 3D), which identifies B-cell Conformational epitopes trained and evaluated on the largest curated epitope data set to date.
- Graph-based signatures are an effective approach to model epitope and non-epitope regions and extracts distance patterns to train and test predictive models.
- epitope3D outperforms available alternative predictors and proposes a non-redundant unbound state antigen benchmark for future developments.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

Melbourne Research Scholarships, a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (MR/M026302/1); Investigator Grant from the National Health and Medical Research Council of Australia (GNT1174405); Victorian Government's OIS Program. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data availability

epitope3D and associated data sets are available through a user-friendly and freely available web interface and API at http://biosig.unimelb.edu.au/epitope3d, enabling seamless integration with bioinformatics pipelines and supporting quick assessment of epitopes to support diagnosis and vaccine design.

References

1. Delves PJ, Martin SJ, Burton DR, *et al*. *Roitt's Essential Immunology*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc, 2017.
2. Van Regenmortel MH. *What Is a B-Cell Epitope? Epitope Mapping Protocols*. Totowa, New Jersey, USA: Humana Press, 2009, 3–20.
3. Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and methods for T-and B-cell epitope prediction. *J Immunol Res* 2017;**2017**:1–14.
4. Flajnik M, DuPasquier L, Paul W. The evolution of the immune system. In: Merritt J (Acquisitions ed); Seto J (Developmental ed); Martin SP (Production ed). *Fundamental Immunology*. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins, 2012.
5. Reineke U, Sabat R. Antibody epitope mapping using SPOT™ peptide arrays. In: Reineke U and Mike Schutkowski M (eds). *Epitope Mapping Protocols*. Totowa, New Jersey, USA: Humana Press, 2009, 145–67.
6. Yasser E-M, Honavar V. Recent advances in B-cell epitope prediction methods. *Immunome Res* 2010;**6**:1–9.
7. Irving MB, Pan O, Scott JK. Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics. *Curr Opin Chem Biol* 2001;**5**:314–24.
8. Dalkas GA, Rooman M. SEPIa, a knowledge-driven algorithm for predicting conformational B-cell epitopes from the amino acid sequence. *BMC Bioinform* 2017;**18**:1–12.
9. Kulkarni-Kale U, Bhosle S, Kolaskar AS. CEP: a conformational epitope prediction server. *Nucleic Acids Res* 2005; **33**:W168–71.
10. Qi T, Qiu T, Zhang Q, *et al*. SEPPA 2.0—more refined server to predict spatial epitope considering species of immune host and subcellular localization of protein antigen. *Nucleic Acids Res* 2014;**42**:W59–63.
11. Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci* 2006;**15**:2558–67.
12. Liang S, Zheng D, Standley DM, *et al*. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. *BMC Bioinform* 2010;**11**:1–6.
13. Zhang J, Zhao X, Sun P, *et al*. Conformational B-cell epitopes prediction from sequences using cost-sensitive ensemble classifiers and spatial clustering. *Biomed Res Int* 2014;**2014**:1–12.
14. Sela-Culang I, Ashkenazi S, Peters B, *et al*. PEASE: predicting B-cell epitopes utilizing antibody sequence. *Bioinformatics* 2015;**31**:1313–5.
15. Liang S, Zheng D, Zhang C, *et al*. Prediction of antigenic epitopes on protein surfaces by consensus scoring. *BMC Bioinform* 2009;**10**:1–10.
16. Sun J, Wu D, Xu T, *et al*. SEPPA: a computational server for spatial epitope prediction of protein antigens. *Nucleic Acids Res* 2009;**37**:W612–6.
17. Sweredoski MJ, Baldi P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics* 2008;**24**:1459–60.
18. Ponomarenko J, Bui H-H, Li W, *et al*. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinform* 2008;**9**:1–8.
19. Rubinstein ND, Mayrose I, Martz E, *et al*. Epitopia: a web-server for predicting B-cell epitopes. *BMC Bioinform* 2009; **10**:1–6.
20. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 2014;**30**:335–42.
21. Pires DE, de Melo-Minardi RC, Da Silveira CH, *et al*. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics* 2013;**29**:855–61.
22. Pires DE, de Melo-Minardi RC, dos Santos MA, *et al*. Cutoff scanning matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics* 2011;**12**:S12.
23. Ren J, Liu Q, Ellis J, *et al*. Positive-unlabeled learning for the prediction of conformational B-cell epitopes. *BMC Bioinform* 2015;**16**:1–15.
24. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
25. Dunbar J, Krawczyk K, Leem J, *et al*. SAbPred: a structure-based antibody prediction server. *Nucleic Acids Res* 2016;**44**:W474–8.
26. Johnson M, Zaretskaya I, Raytselis Y, *et al*. NCBI BLAST: a better web interface. *Nucleic Acids Res* 2008;**36**:W5–9.

27. DeLano WL. Pymol: an open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography* 2002;**40**:82–92.
28. Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
29. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 1994;**20**:216–26.
30. Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321–57.
31. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 2017;**18**:559–63.
32. Pires DE, Ascher DB. mycoCSM: using graph-based signatures to identify safe potent hits against mycobacteria. *J Chem Inf Model* 2020;**60**:3450–6.
33. Pires DE, Ascher DB. CSM-lig: a web server for assessing and comparing protein–small molecule affinities. *Nucleic Acids Res* 2016;**44**:W557–61.
34. Pires DE, Ascher DB. mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res* 2016;**44**:W469–73.
35. Pires DE, Blundell TL, Ascher DB. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci Rep* 2016;**6**:1–8.
36. Pires DE, Blundell TL, Ascher DB. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J Med Chem* 2015;**58**:4066–72.
37. Pires DE, Stubbs KA, Mylne JS, *et al*. Designing safe and potent herbicides with the cropCSM online resource. *bioRxiv2020* November 02 2020. https://doi.org/10.1101/2020.11.01.364240.
38. Kaminskas LM, Pires DE, Ascher DB. dendPoint: a web resource for dendrimer pharmacokinetics investigation and prediction. *Sci Rep* 2019;**9**:1–9.
39. Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res* 2014;**42**:W314–9.
40. Pires DE, Ascher DB. mCSM–NA: predicting the effects of mutations on protein–nucleic acids interactions. *Nucleic Acids Res* 2017;**45**:W241–6.
41. Pires DE, Rodrigues CH, Ascher DB. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res* 2020;**48**:W147–53.
42. Rodrigues CH, Ascher DB, Pires DE. Kinact: a computational approach for predicting activating missense mutations in protein kinases. *Nucleic Acids Res* 2018;**46**:W127–32.
43. Rodrigues CH, Myung Y, Pires DE, *et al*. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res* 2019;**47**:W338–44.
44. Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* 2018;**46**:W350–5.
45. Myung Y, Pires DE, Ascher DB. mmCSM-AB: guiding rational antibody engineering through multiple point mutations. *Nucleic Acids Res* 2020;**48**:W125–31.
46. Myung Y, Rodrigues CH, Ascher DB, *et al*. mCSM-AB2: guiding rational antibody design using graph-based signatures. *Bioinformatics* 2020;**36**:1453–9.
47. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;**22**:2577–637.
48. Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* 2018;**46**:W329–37.
49. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
50. Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res* 1999;**27**:368–9.
51. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
52. Zhang W, Niu Y, Xiong Y, *et al*. Computational prediction of conformational B-cell epitopes from antigen primary structures by ensemble learning. *PLoS One* 2012;**7**:e43575.
53. Zheng W, Zhang C, Hanlon M, *et al*. An ensemble method for prediction of conformational B-cell epitopes from antigen sequences. *Comput Biol Chem* 2014;**49**:51–8.
54. Zhou C, Chen Z, Zhang L, *et al*. SEPPA 3.0—enhanced spatial epitope prediction enabling glycoprotein antigens. *Nucleic Acids Res* 2019;**47**:W388–94.
55. Jespersen MC, Peters B, Nielsen M, *et al*. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res* 2017;**45**:W24–9.
56. Kringelum JV, Lundegaard C, Lund O, *et al*. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol* 2012;**8**:e1002829.
57. Rose AS, Bradley AR, Valasatava Y, *et al*. NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* 2018;**34**:3755–8.