OXFORD

*Briefings in Bioinformatics*, 2023, **24(3)**, 1–8
https://doi.org/10.1093/bib/bbad114
Advance access publication date 10 April 2023
**Problem Solving Protocol**

# epitope1D: accurate taxonomy-aware B-cell linear epitope prediction

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

Corresponding authors: David B. Ascher, Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia. E-mail: d.ascher@uq.edu.au; Douglas E.V. Pires, Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia. E-mail: douglas.pires@unimelb.edu.au

## Abstract

The ability to identify B-cell epitopes is an essential step in vaccine design, immunodiagnostic tests and antibody production. Several computational approaches have been proposed to identify, from an antigen protein or peptide sequence, which residues are more likely to be part of an epitope, but have limited performance on relatively homogeneous data sets and lack interpretability, limiting biological insights that could otherwise be obtained. To address these limitations, we have developed epitope1D, an explainable machine learning method capable of accurately identifying linear B-cell epitopes, leveraging two new descriptors: a graph-based signature representation of protein sequences, based on our well-established Cutoff Scanning Matrix algorithm and Organism Ontology information. Our model achieved Areas Under the ROC curve of up to 0.935 on cross-validation and blind tests, demonstrating robust performance. A comprehensive comparison to alternative methods using distinct benchmark data sets was also employed, with our model outperforming state-of-the-art tools. epitope1D represents not only a significant advance in predictive performance, but also allows biologically meaningful features to be combined and used for model interpretation. epitope1D has been made available as a user-friendly web server interface and application programming interface at https://biosig.lab.uq.edu.au/epitope1d/.

**Keywords:** immunoinformatics, linear epitopes, machine learning, B-cell epitopes

## INTRODUCTION

B-cell epitopes encompass a class of antigenic determinants that are dependent on the amino acid arrangement on the surface of the antigen protein structure. Contiguous stretches of residues along the primary sequence form linear epitopes, whereas non-adjacent residues, though nearby placed due to protein folding, form the discontinuous (or conformational) epitopes. Both forms impose a significant role upon the binding with its counterparts, the Immunoglobulins, that could be either in the form of membrane bound receptors or as antibodies, which are versatile macromolecules capable of recognizing foreign threats [1, 2].

Identifying and selecting the appropriate epitope that could elicit an effective immune reaction in the host, and thus creating a protective memory immunity, is the fundamental basis of vaccine development [3, 4]. Being an extremely complex and multifactorial process, the amount of time spent in vaccine development is on average a decade, and it can cost over 2 billion dollars for it to reach the market [5, 6]. Consequently, effectively aiding the selection of epitope candidates with computational techniques holds a promising role in this field in terms of substantial decreasing development time and cost.

Linear B-cell epitopes account for only 10% among the two classes and although *in silico* prediction methods have significantly evolved over the past decades, varying from amino acid propensity scale scores [7–10], to combining physicochemical attributes and more robust machine learning techniques [11–15], their performance is still biased toward specific data sets, leading to limited generalization capabilities. A recently published approach [16] attempted to address these gaps by systematically cross-testing several previous benchmark data sets on their machine learning model and thus proposing two final models: a generalist and other specifically tailored for viral antigens. However, the general model was trained on data predominantly from HIV epitopes, which could be potentially non-representative, with the virus-specific model still performing modestly, reaching a maximum Matthew's Correlation Coefficient (MCC) of 0.26 on blind tests.

To fill these gaps, here we propose an explainable machine learning classifier based on the largest experimentally curated linear epitope data set so far, covering a large span of organisms, presenting robust performance with different validation techniques, in addition to two new feature representation approaches:

**Bruna Moreira da Silva** is a PhD student at the University of Melbourne. Her research interests are in bioinformatics, immunoinformatics and machine learning to advance Global Health.
**David B. Ascher** is the deputy director of biotechnology at the University of Queensland and head of Computational Biology and Clinical Informatics at the Baker Institute and Systems and Computational Biology at Bio21 Institute. He is interested in developing and applying computational tools to assist leveraging clinical and omics data for drug discovery and personalized medicine.
**Douglas E.V. Pires** is a senior lecturer in digital health at the School of Computing and Information Systems at the University of Melbourne and group leader at the Bio21 Institute. He is a computer scientist and bioinformatician specializing in machine learning and AI and the development of the next generation of tools to analyze omics data, and guide drug discovery and personalized medicine.
**Received:** November 23, 2022. **Revised:** January 30, 2023. **Accepted:** March 7, 2023
© The Author(s) 2023. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

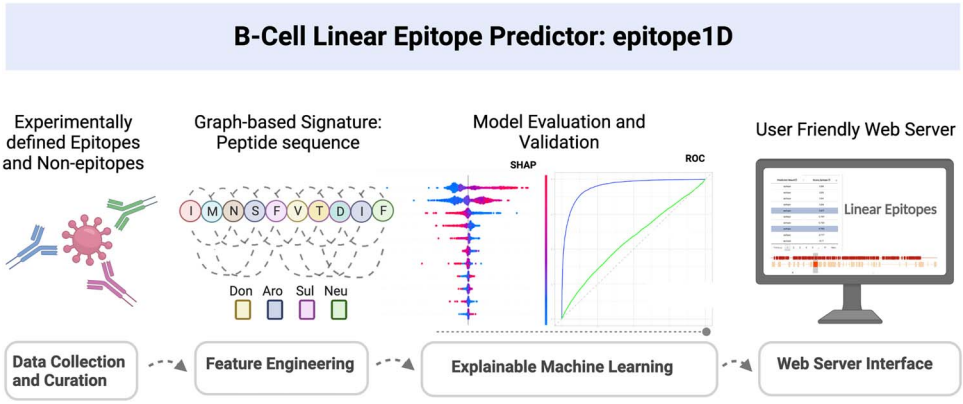


**Figure 1.** epitope1D workflow with four main stages: (1) Data Collection and Curation, which includes the selection of benchmarks and also the curation of an updated large-scale data set; (2) Feature Engineering, representing the step where all descriptors were calculated; (3) Explainable Machine Learning, in which the supervised machine learning classifiers were analyzed in terms of their predictive power, explainability, and assessed via cross-validation and blind-test approaches; (4) Web Server Interface, where epitope1D is made publicly available as a user-friendly web interface and API.

graph-based signatures of protein sequences labeled with physicochemical properties and organism ontology identification of each input peptide, leveraging the classifier distinction between epitopes and non-epitopes.

## MATERIALS AND METHODS

epitope1D has four general steps, as can be seen in Figure 1: (1) Data collection and curation; (2) Feature Engineering: evaluation of currently used features and newly proposed features to represent peptide sequences; (3) Explainable Machine Learning: assessment of several supervised learning algorithms and further comparison among previous work using different data sources and explainability resources and (4) Web server interface: development of an easy-to-use platform for end users.

### Data collection

Well-established reference data sets were used to train, test and validate supervised learning algorithms, as a classification task. A comprehensive list of benchmark data sets, derived from ABCPred method [11], BCPred [17], AAP [18], LBtope [13] and iBCE-EL [12], was further employed to impartially evaluate the predictive power of our selected model against the original ones and state-of-the-art tools such as EpitopeVec [16], BepiPred [19] and BepiPred-2.0 [14]. A detailed review of them is located in Supplementary Materials available online at http://bib.oxfordjournals.org/. Subsequently, given that most data sets were outdated (dated from 5 to 16 years ago), we have curated a newly updated data set derived from IEDB database [20], which will be used as the final basis for our model. Table 1 summarizes the information of all data used in this work, and further analysis can be read below.

### Curating a new experimental benchmark data set

We have curated a new set derived from the IEDB database aiming to reflect data availability on experimentally confirmed linear B-cell epitopes and also non-epitopes. Our main motivation arises because most of the previously mentioned data sets were collected more than 10 years ago, and their negative class sets (non-epitope sequences) were not empirically proven. In addition, epitope sequences derived from the Bcipep database [21] can impose an obstacle to model generalization, given that around 80% of them refer only to HIV. Therefore, organism information from each sequence in the new data is taken into account to assess whether the taxonomy can aid distinguish epitopes from non-epitopes, in a variety of subspecies.

The curation process of the new experimental data set comprises the following steps: (1) Download all possible (any host and disease) linear peptides of B-cell epitopes and non-epitopes as of June, 2022; (2) Keep only the epitopes and non-epitopes confirmed in two or more different assays; (3) Consider solely the peptides with length between 6 and 25 amino acids, since 99% of linear epitopes range within these lengths [1, 12, 14]; (3) Remove sequences that were present in both classes simultaneously; (4) Exclusively retain entries that contain information about the source organism; (5) Perform a systematic sequence redundancy removal step using CD-HIT [22], at different thresholds (95%, 90%, 80% and 70%) to assess the overall learning efficiency within high to medium similarity. Considering that a single antigen may contain several different epitope stretches that lead to distinct antibody bindings, it is worth accommodating the majority of available epitope sequences belonging to each antigen protein [2].

The final set, with a maximum of 95% similarity, is composed of 154,899 data points, in which 25,902 are epitopes, encompassing 1,192 sub-species that were aggregated into a higher taxonomy parent organism lineage of 20 classes, each belonging to the superkingdom of Virus, Eukaryota or Bacteria. The final set was randomly divided into a training set with 123,919 data points, in which 20,638 are epitopes (ratio 1:6) and correspond to 80% of the data, and the remaining 20% as an independent test set with 30,980 data points with the same epitope/non-epitope proportion.

### Feature engineering

To better characterize peptide sequences that might compose an epitope, previously used descriptors as well as novel features were evaluated. To reduce model complexity, a forward stepwise greedy feature selection algorithm was applied [23] to retain only the most representative set. The description of new proposed features is introduced here, while auxiliary features are described in Supplementary Materials available online at http://bib.oxfordjournals.org/. An overall summary is also provided in Table S1 available online at http://bib.oxfordjournals.org/.

#### *Graph-based signatures*

We have designed a new graph-based feature, tailored for modeling linear epitopes of flexible length, inspired by the Cutoff

**Table 1.** Description of the data sets applied to train and evaluate epitope1D

| Data Set | Original Method | #Epitopes | #Non Epitopes | Experimentally Defined | Data Source | Peptide Length | Validation |
|---|---|---|---|---|---|---|---|
| General benchmark data sets | | | | | | | |
| BCPred | BCPred | 701 | 701 | Only Epitopes | Bcipep and SwissProt | 20-mer | Cross-validation |
| ABCPred-1 | ABCPred | 700 | 700 | Only Epitopes | Bcipep and SwissProt | 20-mer | Blind |
| ABCPred-2 | ABCPred | 187 | 200 | Only Epitopes | Bcipep/SDAP and SwissProt | Assorted | Blind |
| AAP | AAP | 872 | 872 | Only Epitopes | Bcipep and SwissPro | 20-mer | Blind |
| LBtope | LBTope | 7,824 | 7,853 | Yes | IEDB | 20-mer | Blind |
| iBCE-EL-1 | iBCE-EL | 4,440 | 5,485 | Yes | IEDB | Assorted | Blind |
| iBCE-EL-2 | iBCE-EL | 1,110 | 1,408 | Yes | IEDB | Assorted | Blind |
| New curated large-scale benchmark data set | | | | | | | |
| epitope1D Training | epitope1D | 20,638 | 103,281 | Yes | IEDB | Assorted | Cross-validation |
| epitope1D Testing | epitope1D | 5,264 | 25,716 | Yes | IEDB | Assorted | Blind |

The first column, named "Data Set", is the name we are referring them to throughout the text; "Original Method" is where the set originally is derived from; "Epitopes" and "Non Epitopes" correspond to the total amount of labeled data within the set; "Experimentally Defined" indicates if the data from the two classes were experimentally assessed; "Data Source" specifies the database from which the set was extracted; "Peptide Length" indicates the size of the peptides within the data set (specifies the length—if fixed; or Assorted); "Validation" column designates if we apply the set for cross-validation or blind-testing purposes.

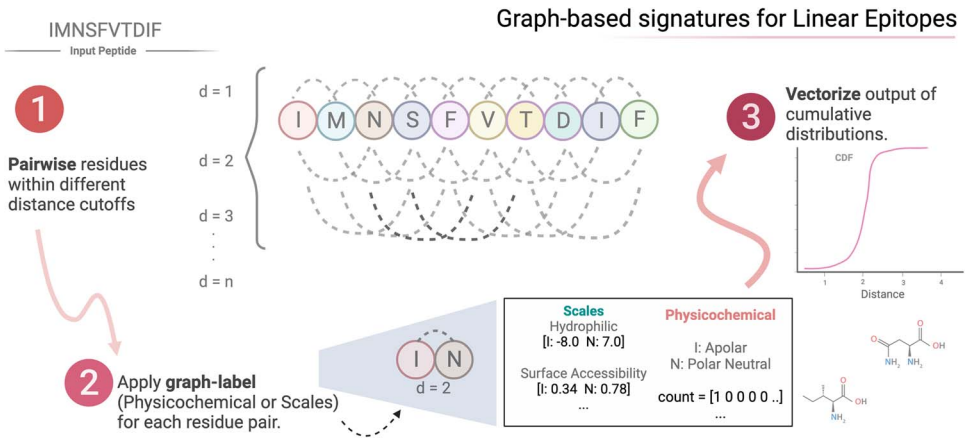


**Figure 2.** Modeling linear epitopes using graph-based signatures. The first step comprises the selection of residue pairs at incremental sequence distances and applies different types of labeling approaches on them. The third and final step is to vectorize the output as cumulative distances between different label pairs.

Scanning Matrix (CSM) algorithm [23–26]. The key idea was to model distance patterns among residues (nodes) at different distance cutoffs (each distance inducing the edges of a graph), which are summarized in two approaches: cumulative and non-cumulative distributions. Sequence graphs were labeled in two ways: (1) the corresponding scales of hydrophilicity prediction [8], beta turn prediction [27], surface accessibility [9] and antigenicity [10] and (2) the amino acid physicochemical properties, such as Apolar, Aromatic, Polar Neutral, Acid or Basic, as done previously [23, 28, 29]. Figure 2 shows the steps comprising the new graph-based feature.

### *Organism identification*

The organism source information, extracted from the IEDB database together with each peptide sequence, is expressed by its ontology identifier deriving from two sources: (1) The Ontobee data server [30] for the NCBI organismal taxonomy and (2) The Ontology of Immune Epitopes, which is an internal web resource from IEDB that was then converted back to the corresponding NCBI taxonomy term for standardization. This information was used aiming to contribute with epitope identification addressing the pain point in the machine learning process that arises from high heterogeneity in organism classes [16, 31]. To transform the 20 ontological terms, described in Table S1 available online at http://bib.oxfordjournals.org/, from categorical data into numerical, a one-hot encoding process was imposed. This descriptor was applied in the new curated benchmark data set only.

## Machine learning methods

As epitope identification could be described as a binary classification task, various supervised learning algorithms were assessed using the Scikit Learn Python library [32]. These included Support Vector Machine (SVM), Adaptive Boosting, Gradient Boosting, Random Forest (RF), Extreme Gradient Boosting, Extra Trees, K-nearest neighbor, Gaussian Processes and Multi-Layer Perceptron. In addition, an inherently interpretable method named Explainable Boosting Machine (EBM), a type of generalised additive model and considered as a glassbox model, was assessed via the open-source Python module InterpretML [33]. The goal of interpretable machine learning models is to provide a rationale behind prediction that would allow for meaningful biological insights to be derived, also assisting in the possible biases and errors as well as highly predictive features.

Performance evaluation for each model was done based on MCC, which is a robust statistical measure appropriate for imbalanced data sets [34]. Complementary performance metrics were

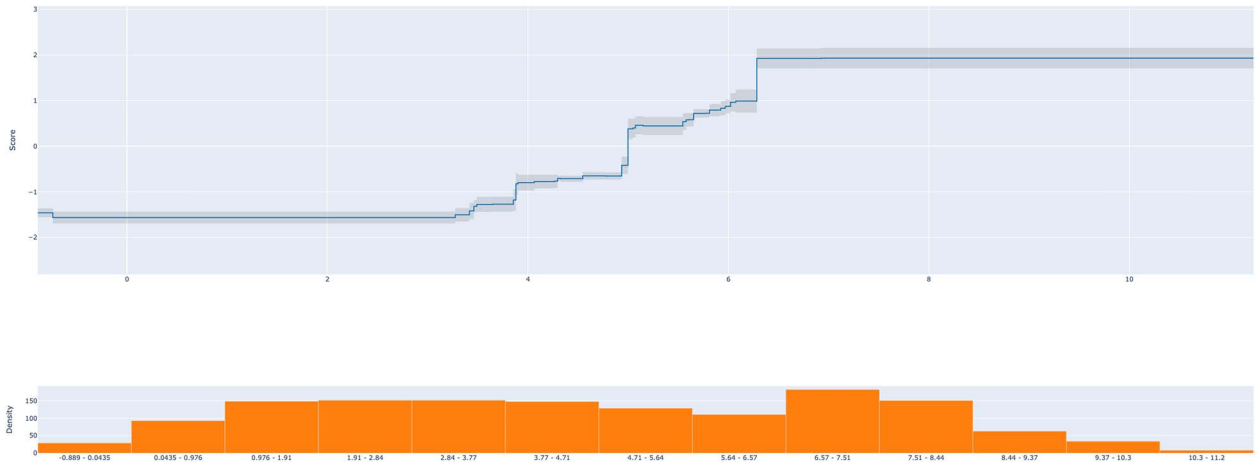


**Figure 3.** Behavior of the feature AAT_max, which represents the maximum rate of Antigenicity for AAT, over its possible values ranging from −0.889 to 11.2. The epitope class probability increases when the AAT_max becomes larger than 5 (score above 0). The bottom chart depicts the distribution of the corresponding data points in each feature interval.

also used including F1-score, Balanced Accuracy and Area Under the ROC Curve (AUC). Performance between internal validation (*k*-fold cross validation) and external validation (blind tests) was contrasted to infer generalization capabilities.

## RESULTS

Two scenarios of evaluation were considered to assess the ability of epitope1D to accurately identify linear epitopes, based on the data set employed and direct comparison with previous methods. The first comprises the use of well-established benchmarks: the BCPred set, assessed under 10-fold cross-validation, followed by external validation with different independent blind-test sets, as previously described (ABCPred-1, ABCPred-2, AAP, LBtope, iBCE-EL-1 and iBCE-EL-2). This scenario impartially compares the performance of our models with recent developments and identifies feature importance and their current limitations.

The second scenario consists of our newly curated, large-scale data set extracted from the IEDB database which includes organism information. Data sets filtered at different sequence similarity levels were employed: with internal validation evaluated using 10-fold cross-validation, and models assessed externally via blind tests. Furthermore, recently development methods, including BepiPred-3.0 [35], EpitopeVEC and EpiDope, also had their performance assessed on the same blind test to determine differences in performance.

### Comparison with alternative methods using the BCPred data set

#### *Feature representation: What makes up a linear epitope?*

The data set extracted from BCPred, composed of 1402 peptide sequences with a balanced class ratio of 1:1, was used to train and test several supervised learning models as a classification task. Their ability to distinguish between epitopes and non-epitopes was assessed and most predictive features identified. Outstanding features identified in this scenario include: (i) the maximum and minimum value of Antigenicity ratios in terms of amino acid triplets (AAT, measuring how overrepresented some amino acids are in the epitope class of this data set); (ii) the Composition pattern (Composition, Transition and Distribution - CTD) of physicochemical and structural properties (hydrophobicity, normalized van der Waals volume, polarity, secondary structure and solvent accessibility) and (iii) the Graph-based signatures using both types of labeling: physicochemical properties (Acidic, Apolar, Polar Neutral, Basic and Aromatic) and Parker hydrophilicity prediction scale.

Using the interpretable classifier, EBM, to understand feature importance (Figure S1 of Supplementary Materials available online at http://bib.oxfordjournals.org/), we observed that the antigenicity ratio features were in the top three most relevant: the maximum value within a peptide sequence (AAT_max), the interaction amid the maximum and the minimum (AAT_max x AAT_min) and the minimum value (AAT_min); followed by its interaction with specific Graph-based and Composition physicochemical descriptors, such as Apolar:Aromatic-8 (pairs of apolar and aromatic amino acids within a sequence distance cutoff of 8) and the amino acid composition in terms of Hydrophobicity (G1: Polar, G2: Neutral, G3: Hydrophobicity).

Further exploring interpretability, Figure 3 depicts the rationale behind the model's decision considering only the topmost significant feature, the maximum AAT value, the cumulative sum of the antigenicity ratio scale for all possible AAT within a peptide sequence. In the top chart, the horizontal axis details the feature range values, while the vertical axis shows the class, with the decision mark between the two classes set to 0 (above zero a higher probability of being an epitope and non-epitope otherwise). A clear decision point has been learned when *AAT_max* ranges from 4 to 6 (more precisely, larger than 5), which is also the average value of this feature for this data set. A possible interpretation of this result, in terms of the data set that includes 20-mer peptides only, can be that if at least five combinations of AAT are overrepresented in the sequence, there is a higher chance of it being an epitope. The bottom chart of the figure depicts the feature value distribution.

### Machine learning models

Under 10-fold cross-validation, the best performing models include RF and EBM, both reaching a MCC of 0.72 and AUC of 0.92 and 0.93, respectively. Similar performance was observed using 5-fold cross validation. Table 2 provides a comparison of epitope1D with the previous methods that employed this data set including BCPred and EpitopeVec. Both methods used an SVM approach.

**Table 2.** Performance comparison of epitope1D using two different algorithms (EBM and RF) with BCPred and EpitopeVec methods under 10-fold cross-validation using the BCPred data set

| METHOD | MCC | ROC-AUC | F1 |
|---|---|---|---|
| BCPred | 0.360 | 0.758 | –[a] |
| EpitopeVec | 0.620 | 0.880 | 0.810 |
| epitope1D (RF) | 0.720 | 0.920 | 0.850 |
| epitope1D (EBM) | 0.720 | 0.930 | 0.860 |

[a]Metric unavailable in original publication.

**Table 3.** Performance comparison with previous methods using three distinct blind-test sets: AAP, ABCPred-2 and LBTope

| METHOD | MCC | ROC-AUC | F1 | ACCURACY |
|---|---|---|---|---|
| *Data set: AAP* | | | | |
| iBCE-EL | −0.036 | 0.528 | 0.350 | 0.494 |
| BepiPred | 0.217 | 0.665 | 0.600 | 0.604 |
| BepiPred-2.0 | −0.021 | 0.424 | 0.400 | 0.493 |
| EpiDope | 0.061 | 0.559 | 0.350 | 0.507 |
| EpitopeVec | 0.770 | **0.958** | 0.880 | 0.883 |
| epitope1D | **0.815** | 0.907 | **0.909** | **0.907** |
| *Data set: ABCPred-2* | | | | |
| ABCPred | –[a] | –[a] | –[a] | 0.664 |
| AAP | 0.292 | 0.689 | –[a] | 0.646 |
| iBCE-EL | −0.227 | 0.501 | 0.320 | 0.434 |
| BepiPred | 0.132 | 0.627 | 0.570 | 0.566 |
| BepiPred-2.0 | 0.181 | 0.62 | 0.480 | 0.555 |
| EpiDope | 0.091 | 0.541 | 0.360 | 0.508 |
| EpitopeVec | 0.445 | 0.778 | 0.720 | 0.718 |
| epitope1D | **0.543** | **0.841** | **0.848** | **0.781** |
| *Data set: LBTope* | | | | |
| iBCE-EL | **0.135** | **0.619** | 0.39 | 52.20 |
| BepiPred | 0.092 | 0.566 | **0.55** | **54.57** |
| BepiPred-2.0 | −0.001 | 0.476 | 0.42 | 49.95 |
| EpiDope | 0.036 | 0.559 | 0.35 | 50.34 |
| EpitopeVec | 0.065 | 0.548 | 0.51 | 52.98 |
| epitope1D | 0.067 | 0.527 | 0.333 | 52.80 |

[a]Metric unavailable in original publication. Highlighted in bold are the highest performing values for each dataset.

In order to externally validate our model and assess its generalization capabilities, different blind-test sets were presented to the epitope1D (EBM) model as detailed in Table 3 and in Table S2 of Supplementary Materials available online at http://bib.oxfordjournals.org/, where we can also examine the performance achieved by previous methods such as BepiPred, BepiPred-2.0, EpiDope and EpitopeVec. Significant performance differences were observed for the methods when trained and tested using different data sources (i.e. Bcipep and IEDB databases). For instance, with the AAP data set (originally derived from Bcipep) in the first part of Table 3, epitope1D and EpitopeVec (both trained using data from Bcipep database) achieved higher performances (MCC of 0.815 and 0.770, respectively) compared to the other methods that were trained using data derived from IEDB database (iBCE-EL, BepiPred and EpiDope). Alternatively, when applying the iBCE-EL testing data set (extracted from IEDB), our model and EpitopeVec achieved lower values of MCC, 0.092 and 0.095, compared to the model trained on data from this source (which reached a MCC of 0.454). However, in this scenario, some models trained using the same data source (e.g. BepiPred, BepiPred-2.0 and EpiDope) did not perform well either.

Regarding the machine learning process, this behavior raises concerns of potential biases in these data sets or lack of representativeness. Both cases lead to a lack of generalization and can occur due to a variety of reasons, some of which we might conjecture: (i) all previous benchmark data sets listed here were adjusted to a highly balanced ratio for epitope and non-epitope classes, which do not represent the biological truth; (ii) Bcipep databases, as originally stated, are predominantly composed of viruses (HIV predominantly), which induces an underrepresentation of other organisms; (iii) Truncation/Extension approaches, adopted by part of the methods to define a fixed peptide length, change the originally validated epitope sequence and may impose a learning bias toward an artificial set; (iv) The use of non-experimentally validated sequences to populate the non-epitope class, strategy adopted by some of the benchmarks, could lead the machine learning model to learn from imprecise or even erroneous data.

## Performance on a newly curated benchmark data set from the IEDB database

### *Feature importance and organism-aware predictions*

To address the potentially unrepresentative nature of the data, we curated an experimentally validated, large-scale data set, integrated with high-level taxonomy organism information that incorporates the three main superkingdoms: Virus, accounting for 83% of the data and enclosed in eight classes (Riboviria, Duplodnaviria, Monodnaviria, Varidnaviria, Ribozyviria, Anelloviridae, Naldaviricetes, Adnaviria), followed by 15% of Eukaryota with five classes (Metamonada, Discoba, Sar, Viridiplantae, Opisthokonta); and 2% of Bacteria with seven classes (Terrabacteria group, Proteobacteria, PVC group, Spirochaetes, FCB group, Thermodesulfobacteria, Fusobacteria), totalising 20 binary categories. Organism taxonomy information was included in the set of features previously used (detailed in Table S3 of Supplementary Materials available online at http://bib.oxfordjournals.org/), composed of four main categories: Graph-based signatures, AAT Antigenicity ratio, Composition features and Organism taxonomy.

To better understand individual feature contributions to model outcomes, a post-hoc analysis using the SHAP [36] was employed using the RF model. The importance order of each descriptor in this scenario can be understood as a ranked summary depicted in Figure S2 of Supplementary Materials available online at http://bib.oxfordjournals.org/. The Antigenicity ratio group, with AAT maximum and minimum values, are very predictive features with higher values strongly correlating with the epitope class. The next most important feature is part of the Composition group, charge.G3, denoting a higher number of negatively charged amino acids in the epitope class. The fourth most important feature was organism taxonomy, particularly the Riboviria, potentially showing what the model learned as a consequence of the class imbalanced data, where 87% of the Riboviria sequences belong to the non-epitopes, thus correlating the epitope class with organisms other than Riboviria. Graph-based signature features also play an important role to the model decision (e.g. Apolar:PolarNeutral-9 feature), denoting pairs of residues (polar and apolar) far apart from each other in sequence, though contributing to the epitope class (particularly for Riboviria sequence, Figure S3 available online at http://bib.oxfordjournals.org/). To complement the SHAP summary visualization, the impurity-based value (Gini importance) is also being presented in Table S3 available online at http://bib.oxfordjournals.org/, besides each feature description, as well as the performance of each feature category (Table S4 of Supplementary Materials available online at http://bib.oxfordjournals.org/). The Gini importance of all

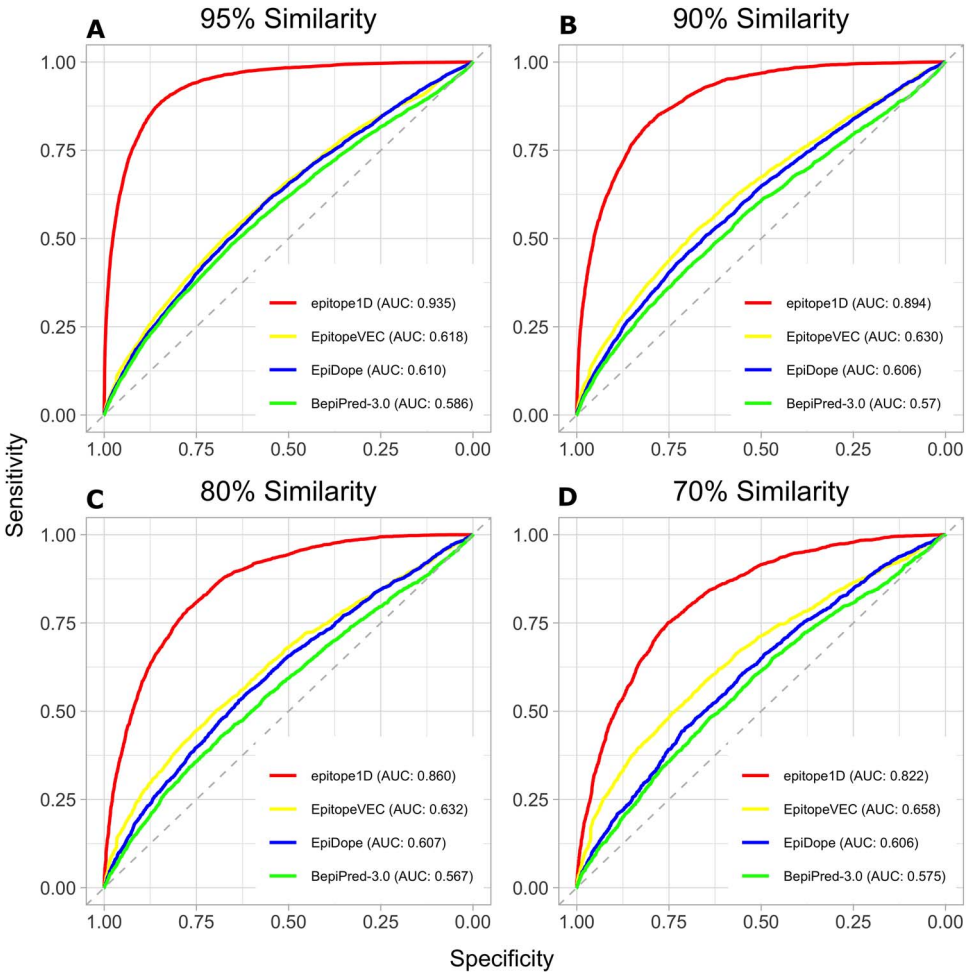


**Figure 4.** Performance comparison via ROC curves using the epitope1D test set at different similarity levels (95%, 90%, 80% and 70% for panels, **A**, **B**, **C** and **D**, respectively). epitope1D achieved significantly higher AUC values of up to 0.935 (curve located closer to the upper left axes in red), followed by EpitopeVEC (in yellow), EpiDope (in blue) and BepiPred-3.0 (in green).

descriptors sums to 1, thus the higher the value the stronger the contribution to the model decision. Although the sorted descending order presents some slight differences compared with the SHAP ranking, we can again perceive that the top 10 features comprise the Antigenicity group, with around 20% and 17% importance to the AAT_max and AAT_min, followed by 11% of the Graph-based named Apolar:PolarNeutral-9, the Composition group and two of the Organisms, the Riboviria with 3.5% and the Opisthokonta with 1.2%.

### *Machine learning models*

In this second analysis, EBM and RF classifiers were assessed and performed equally, with RF presenting a slightly faster training time, the reason why it has been chosen. Ten-fold cross-validation was performed using the epitope1D training set, followed by the blind-test evaluation with the independent blind test. Performances of state-of-the-art methods BepiPred-3.0, EpitopeVec and EpiDope on the same blind test were compared. Table 4 outlines the performance metrics for the cross-validation and blind tests, with epitope1D reaching a MCC of 0.613 and 0.608, respectively, in contrast with the best performing alternative method, EpitopeVEC, only achieving up to 0.139 MCC and BepiPred-3.0 achieving −0.007. Although the EpiDope method was trained using data from the same source, IEDB, we did not perform a homology removal check on the test set to guarantee direct comparison and avoid data contamination.

**Table 4.** Performance metrics on cross-validation (CV) and blind test, using epitope1D data set. State-of-the-art methods for linear B-cell epitope prediction were also appraised using the blind test set: BepiPred-3.0, EpitopeVEC and EpiDope.

| Method | MCC | ROC-AUC | F1 |
|---|---|---|---|
| epitope1D (CV) | **0.613** | **0.935** | **0.658** |
| epitope1D (blind test) | **0.608** | **0.935** | **0.654** |
| EpitopeVec (blind test) | 0.139 | 0.618 | 0.306 |
| EpiDope (blind test) | 0.051 | 0.610 | 0.024 |
| BepiPred-3.0 (blind test) | -0.007 | 0.586 | 0.290 |

Highlighted in bold are the highest performing values in cross validation and blind test.

To better visualize model performance and Sensitivity/Specificity tradeoff for all the methods on the blind test, ROC curves were created (Figure 4), for the data set filtered at different similarity level cutoffs. The epitope1D curve, displayed closer to the top-left axes in red, reached a significantly better ROC-AUC value of 0.935, compared to 0.618 from EpitopeVEC, plotted just below it in yellow, followed by 0.610 from EpiDope in blue and 0.586 from BepiPred-3.0 in green, nearer the dashed diagonal line (for a 95% similarity cutoff—Figure 4A). Further analysis with different cutoffs of 90%, 80% and 70% (Figure 4B–D, respectively) applied

in the independent testing set are likewise depicted, demonstrating that epitope1D consistently and significantly outperforms all alternative methods, with a small decrease in performance, further highlighting its robustness. In addition, we evaluated the iBCE-EL method using the 70% similarity blind-test set, since this method was originally trained and tested using data from IEDB database with a homology threshold of 70%, which resulted in a MCC value of 0.120 and ROC-AUC of 0.630, compared to 0.440 and 0.822 for epitope1D, respectively, emphasizing the considerable gain in performance due to incremental data and proposed new features.

The architecture of both EpiDope and BepiPred-3.0 models is based on deep learning approaches with language models; the former applies a pre-trained Embeddings from Language Model, ELMo, that is a character-level CNN to encode the amino acid input sequences followed by LSTM layers, while the later employed a similar strategy, but combining Feed Forward with CNN and LSTM and additionally assessing the model with BLOSUM62 and sparse encodings.

### *Web server and application programming interface*

epitope1D was made available as a user-friendly web server interface, where the user can input the protein or peptide sequence in fasta format and select from a drop-down menu the equivalent organism taxonomy representation (Figure S4 available online at http://bib.oxfordjournals.org/). In addition, an application programming interface (API) enables for batch submissions and integration to standard analytical pipelines, contributing to reproducibility and usability of the resource.

## CONCLUSIONS

Linear B-cell epitope prediction is yet an extremely challenging task in which the sophisticated biological mechanism underlying the binding among Antibody–Antigen poses challenges to computational and experimental methods. The majority of previous benchmark data sets dated from up to 15 years ago, also suggesting a potential bias and lack of organism representativeness, leading to weak to poor generalization capabilities.

epitope1D fills these gaps via an explainable machine learning method, built on the largest non-redundant experimentally validated data set to date, composed of over 150,000 data points, consisting of a diverse set of organisms within Virus, Eukaryota and Bacteria superkingdoms. epitope1D leverages well-established as well as novel features engineered to model epitopes, including a new graph-based signature to train and test taxonomy-aware and accurate predictors.

A comprehensive comparison of our method with state-of-the-art tools showed robust performance across distinct blind-test sets, with epitope1D significantly outperforming all methods, thus highlighting its generalization capabilities. We believe epitope1D will be an invaluable tool assisting vaccine and immunotherapy development and have made it freely available to the community as an easy-to-use web interface and API at https://biosig.lab.uq.edu.au/epitope1d/.

**Key Points**

- epitope1D is a novel and accurate linear B-cell epitope predictor based on an explainable machine learning.
- We have curated the largest, non-redundant experimentally derived epitope data set to date, covering a large number of organisms.
- epitope1D demonstrates robust performance and generalization capabilities, being validated under internal and external validation procedures and outperforming state-of-the-art approaches.
- epitope1D is available as an API to allow programmatic integration with bioinformatics pipelines as well as democratizing access to users with no computational background via a user-friendly web interface.

## SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

## FUNDING

Melbourne Research Scholarship; Investigator Grant from the National Health and Medical Research Council of Australia (GNT1174405); Victorian Government's OIS Program; D.E.V.P. received funding from an Oracle for Research Grant.

## DATA AVAILABILITY

epitope1D and associated data sets are available through a user-friendly and freely available web interface and API at https://biosig.lab.uq.edu.au/epitope1d/, enabling seamless integration with bioinformatics pipelines and supporting quick assessment of sequences to support diagnosis and vaccine design. In addition, a standalone version is also accessible at https://github.com/munamomo/epitope1D.

## REFERENCES

1. Ponomarenko JV, Regenmortel MHVV. *B-Cell Epitope Prediction*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc. Vol. **2**. Structural Bioinformatics, 2009:849–79.
2. W. E. Paul, *Fundamental Immunology*. Philadelphia, USA: Wolters Kluwer, 2012. [Online] http://ebookcentral.proquest.com/lib/unimelb/detail.action?docID=3417830 (20 May 2022, date last accessed).
3. Takahashi H. Antigen presentation in vaccine development. *Comp Immunol Microbiol Infect Dis* 2003;**26**(5):309–28.
4. Hoft DF, Brusic V, Sakala IG. Optimizing vaccine development. *Cell Microbiol* 2011;**13**(7):934–42. https://doi.org/10.1111/j.1462-5822.2011.01609.x.
5. Gouglas D, Thanh le T, Henderson K, *et al.* Estimating the cost of vaccine development against epidemic infectious diseases: a cost minimisation study. *Lancet Glob Health* 2018;**6**(12):e1386–96.
6. Plotkin S, Robinson JM, Cunningham G, *et al.* The complexity and cost of vaccine manufacturing—an overview. *Vaccine* 2017;**35**(33):4064–71.
7. Welling GW, Weijer WJ, van der Zee, Welling-Wester S. Prediction of sequential antigenic regions in proteins. *FEBS Lett* 1985;**188**(2):215–8.
8. Parker JMR, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography

peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. *Biochemistry* 1986;**25**(19):5425–32.
9. Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol* 1985;**55**(3):836–9.
10. Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett* 1990;**276**(1):172–4.
11. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 2006;**65**(1): 40–8.
12. Manavalan B, Govindaraj RG, Shin TH, *et al.* iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. *Front Immunol* 2018;**9**. [Online]. https://www.frontiersin.org/article/10.3389/fimmu.2018.01695 (19 May 2022, date last accessed).
13. Singh H, Ansari HR, Raghava GPS. Improved method for linear B-cell epitope prediction using Antigen's primary sequence. *PloS One* 2013;**8**(5):e62216.
14. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res* 2017;**45**(W1):W24–9.
15. Collatz M, Mock F, Barth E, *et al.* EpiDope: a deep neural network for linear B-cell epitope prediction. *Bioinformatics* 2021;**37**(4): 448–55.
16. Bahai A, Asgari E, Mofrad MRK, *et al.* EpitopeVec: linear epitope prediction using deep protein sequence embeddings. *Bioinformatics* 2021;**37**(23):4517–25.
17. EL-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels. *J Mol Recognit* 2008;**21**(4):243–55. https://doi.org/10.1002/jmr.893.
18. Chen J, Liu H, Yang J, Chou K-C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* 2007;**33**(3):423–8.
19. Larsen JEP, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. *Immunome Res* 2006;**2**(1):2.
20. Vita R, Mahajan S, Overton JA, *et al.* The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**(D1): D339–43.
21. Saha S, Bhasin M, Raghava GP. Bcipep: a database of B-cell epitopes. *BMC Genomics* 2005;**6**(1):79.
22. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**(23):3150–52.
23. da Silva, Myung Y, Ascher DB, Pires DEV. epitope3D: a machine learning method for conformational B-cell epitope prediction. *Brief Bioinform* 2022;**23**:bbab423.
24. Pires DE, de Melo-Minardi, dos Santos, *et al.* Cutoff scanning matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics* 2011;**12**(4):S12.
25. da Silveira, Pires DEV, Minardi RC, *et al.* Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins* 2009;**74**(3): 727–43.
26. Pires DEV, de Melo-Minardi, da Silveira, *et al.* aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics* 2013;**29**(7):855–61.
27. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* 1978;**47**:45–148.
28. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 2014;**30**(3):335–42.
29. Pires DEV, 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**(W1): W314–9.
30. Xiang Z, Mungall C, Ruttenberg A, He Y. Ontobee: A linked data server and browser for ontology terms. *In ICBO* 2011.
31. Ashford J, Reis-Cunha J, Lobo I, *et al.* Organism-specific training improves performance of linear B-cell epitope prediction. *Bioinformatics* 2021;**37**(24):4826–34.
32. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *The Journal of machine Learning research* 2011;**12**:2825–30.
33. Nori H, Jenkins S, Koch P, Caruana R. Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv 2019:1909.09223.
34. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; **21**(1):6.
35. Clifford JN, Høie MH, Deleuran S, Peters B, et al. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Science* 2022;**31**(12):e4497.
36. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;**30**.