Estimating the Limits of Organism-Specific Training for Epitope Prediction

Jodie Ashford, Anikó Ekárt, Felipe Campelo
Aston Centre for Artificial Intelligence Research and Application
Aston University
B4 7ET, Birmingham, UK
Email: f.campelo@aston.ac.uk

***Abstract*—The identification of linear B-cell epitopes is an important task in the development of vaccines, therapeutic antibodies and several diagnostic tests. Recently, organism-specific training has been shown to improve prediction performance for data-rich organisms. This article investigates the limits of organism-specific training for epitope prediction, by systematically quantifying the effect of the amount of training data on the performance of the models developed. The results obtained indicate that even models trained on small organism-specific data sets can outperform similar models trained on much larger heterogeneous and mixed data sets, as well as widely-used predictors from the literature, which are trained on heterogeneous data. These results suggest the potential for a much broader applicability of pathogen-specific models, which can be used to accelerate the development of diagnostic tests and vaccines in the context of emerging pathogens and to support faster responses in future disease outbreaks.**

I. Introduction

The immune system is a complex network of processes designed to protect the body against pathogens. One vital aspect of human immunity is humoral, or antibody-mediated, immunity. In humoral immunity, B-lymphocytes, also known as B-cells, are activated when their B-cell receptor (BCR) binds with an antigen. Activated B-cells then produce antibodies, which are released into the circulatory system to find and bind with their specific antigens [1]. This antigen-antibody recognition is a vital process in protecting the body against pathogens and B-cells are key cells in this process.

A B-cell epitope (or antigenic determinant) is the exact portion of an antigen that the antigen-binding site of a B-cell receptor recognises and binds to [2], [3]. B-cell epitope identification is an essential process in a number of medical processes; it can help with therapeutic antibody production, vaccine development and in developing diagnostic tools [4]–[6]. There are two categories of epitopes: linear and conformational. Linear or continuous epitopes correspond to contiguous sequences of amino acid (AA) residues; these epitopes are recognised by antibodies by their primary structure/linear sequence of amino acids. Conformational or discontinuous epitopes are formed by AAs that, although separated in the primary sequence, are brought together by protein folding [7, Chapter 3].

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

Most current epitope prediction methods are designed to predict linear epitopes [8]–[18], though the vast majority of epitopes are thought to be conformational [19], [20]. There are multiple reasons for this: due to their nature, linear epitopes can be predicted from protein sequence data alone, which are readily available in numerous public databases [21]–[24]; conformational epitopes, on the other hand, require structural protein data for prediction which, historically, has not been as readily available. Predicting conformational epitopes also takes more time as it is more computationally expensive than linear epitope prediction [25] and these epitopes are more difficult to synthesise in the laboratory [26, Chapter 1]. For these reasons most epitope prediction studies, including ours, focus on linear B-cell epitope prediction.

Traditionally, experimental methods were used for B-cell epitope identification, for example: X-ray crystallography, peptide arrays, enzyme-linked immunosorbent assay (ELISA) and phage display [26]–[28]. However, these methods are time consuming, resource intensive and technically difficult to execute [6], [26]. Because of this and the current availability of protein sequence data, the focus is now on computational methods for epitope prediction. Machine learning algorithms for epitope prediction are trained to be able to distinguish B-cell epitopes from non-epitopes. Numerous machine learning (ML) methods exist for B-cell epitope prediction and these methods have been shown to generally outperform early epitope prediction methods based solely on simple amino acid propensity scale calculations [3], [29].

Examples of machine learning approaches for epitope prediction include: neural network-based methods such as ABCpred [12], which uses a recurrent neural network (RNN) to predict B-cell epitopes from antigen sequences using fixed length patterns and other amino acid composition-based features as input. Other popular ML methods for epitope prediction include Support Vector Machines (SVM) [30] which have been used in many epitope prediction pipelines [13], [31]–[39]. One example of this is BCPred [13], which uses SVM classifiers with string kernels [13]. Random Forest Classifiers [40] have also been used in multiple epitope prediction pipelines [17], [41], [42]. Saravanan and Gautham described an amino acid composition-based feature descriptor, Dipeptide Deviation from Expected Mean (DDE), and evaluated it using a support vector machine and an AdaBoost-Random Forest,

with the latter exhibiting the best performance [42].

ML methods like the ones mentioned above help to bypass some of the difficulties (e.g. time and resources) usually encountered by traditional experimental epitope prediction methods [25], [43], [44]. However, many prediction methods still exhibit relatively low prediction performance [25]. Currently, most epitope prediction models are trained on large heterogeneous data sets made up of observations from multiple organisms including: prokaryotes, viruses, fungi, protozoan, humans and other eukaryotes. However, we have recently shown that training models on smaller organism-specific data sets can help improve predictive performance [45]. In that work, organism-specific models were developed for three different organisms, selected due to the availability of a large volume of observations – both validated epitopes and non-immunogenic peptides – in the Immune Epitope Database (IEDB) [46]. The results obtained showed that, for these data-rich organisms, organism-specific models outperformed models trained on much larger heterogeneous data sets as well as several of the best epitope prediction tools from the literature, across multiple performance measures.

Unfortunately, large volumes of validated epitope data are not available for most organisms, which is particularly exacerbated in the case of emerging pathogens that may represent pandemic risk. As an example, at the start of the 2022 global monkeypox outbreak only five LBCEs were listed on the IEDB for the MPX virus, with no negative examples [47], a common scenario for emerging zoonotic pathogens which could preclude the training of models using exclusively organism-specific data. The aim of this study is therefore to investigate the limits of organism-specific training, by focusing on two main questions: (i) How does the number of available organism-specific training peptides affect prediction performance?; and (ii) What is the smallest volume of organism-specific data that produces models surpassing the performance of those trained on large, heterogeneous data sets? To answer these questions, we calculate and compare the predictive performance of models trained on reduced training sets against models trained on mixed data as well as on large, heterogeneous data sets. We also contrast the observed performances with four predictors from the literature trained as generalist (as opposed to organism-specific) models – Bepipred2.0 [17], LBtope [39], iBCE-EL [48] and ABCpred [6]. We hope that, by clarifying this critical aspect in the training of tailored models for specific pathogens, this study can further support the investigation of better modelling practices for the development of higher-performance epitope predictors in the context of emerging pathogens of pandemic potential. Finally, it is relevant to highlight that, although non-explainable ML approaches often encounter challenges for adoption in medical domains [49], the task of epitope prediction is sufficiently upstream from direct clinical application to not suffer from this challenge. Computationally discovered targets always need further experimental validation, at which point biochemical domain expertise takes over and interpretability of results becomes straightforward.

II. Methods

A. Data Sets

Data from three pathogens specific to the organisms: *Onchocerca volvulus* (taxonomy ID: 6282), Epstein-Barr Virus (taxonomy ID: 10376) and Hepatitis C Virus (taxonomy ID: 11102) were used [45]. These data sets were generated based on the full XML export of the IEDB retrieved on the 10th of October 2020, and filtered according to the criteria listed by Ashford et al. [45] (section 2.1, "Data sets"). The available data were split at the protein level, with entries coming from the same protein, or from proteins exhibiting sequence coverage and similarity greater than 80%, always placed in the same split. Two base sets were derived from the data available for each organism: a *Hold-out* set containing approximately 25% of the data; and a second set containing the remaining observations to be used for all model development activities. A set of *Heterogeneous* data was also extracted for each organism, by randomly sampling observations, grouped by taxonomy ID, from the full IEDB export (excluding any observations related to the specific organism). These heterogeneous sets contain around 6000 labeled peptides, with a 50% class balance.

We set out to investigate the effect of the size of organism-specific data sets on prediction performance, and try to estimate rough lower bounds of the required amount of data for organism-specific training to still represent a good alternative to models developed on larger, heterogeneous data sets. For these we extracted several *reduced* organism-specific and heterogeneous/hybrid training sets for each organism, based on the available model development data described above. For each organism and each desired training set size, we split the full model development data into smaller non-overlapping *Organism-specific* data sets, each containing data from between 20 and 500 peptides (see figure 1). The same class balance as the full organism-specific data set was maintained in all subsets.

Table I details the information on the reduced organism-specific data sets generated for each pathogen. Based on these variable-sized organism-specific training sets, we assembled two groups of *hybrid* data sets:

- *Hybrid-A*, composed of the organism-specific peptides plus an equal amount of peptides sampled from other pathogens. Consequently, *Hybrid-A* data sets were always composed of twice as many peptides as their corresponding organism-specific ones, and the balance between organism-specific and "other" peptides was always 50-50%.
- *Hybrid-B*, composed of the organism-specific peptides plus the required amount of peptides sampled from other pathogens to complete a data set size of $1,000$ training peptides (e.g., 20 organism specific + 980 "other" peptides, 40+960, etc.). *Hybrid-B* data sets had a fixed size, but a varying level of balance of data from the target pathogen vs. other organisms.

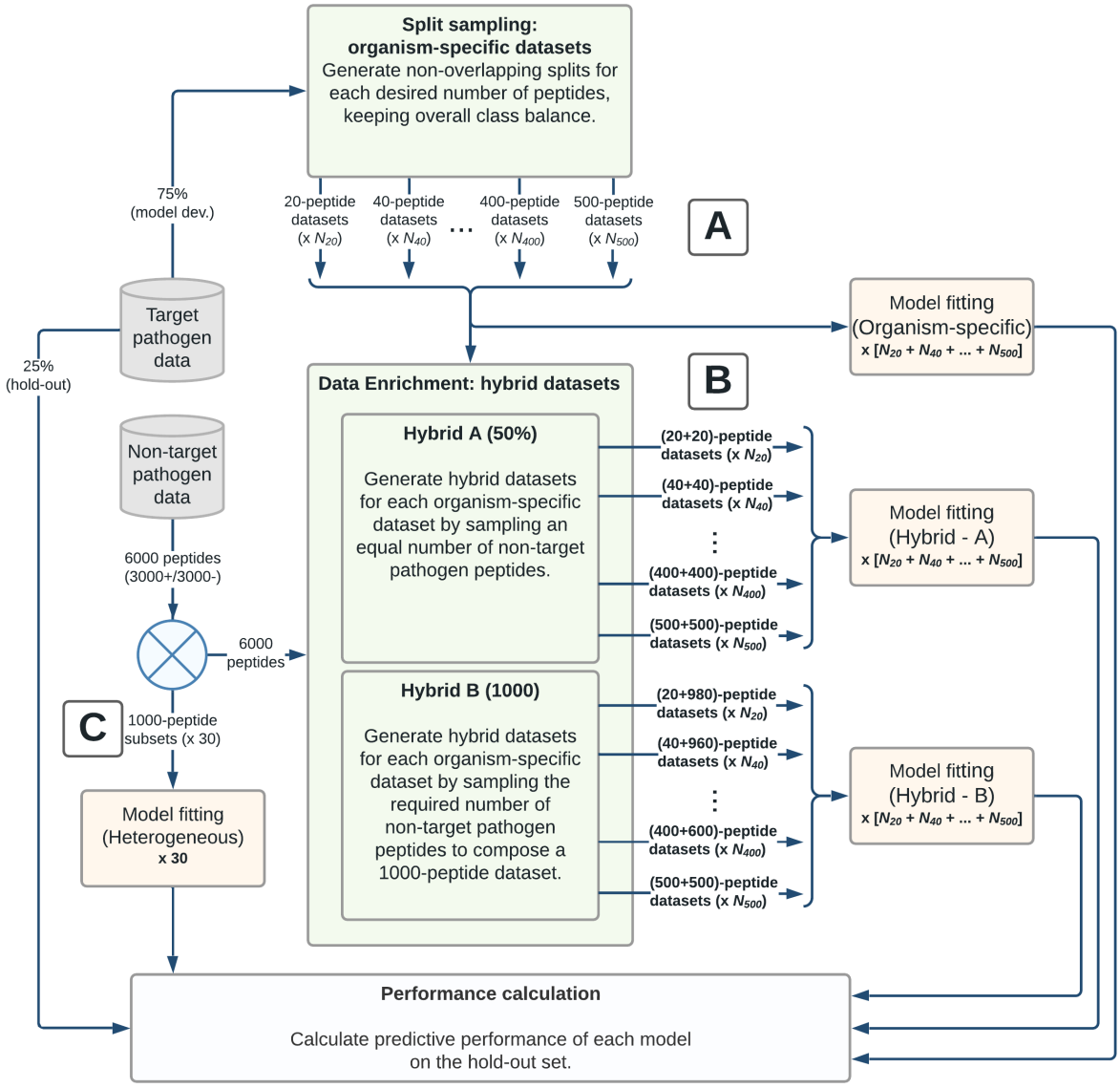

Fig. 1. Experimental protocol for testing the limits of organism-specific model training for linear B-cell epitope prediction. (A) For each pathogen and each desired data size (in terms of number of peptides from the target pathogen), the model development data set is split into non-overlapping subsets of the desired size, each maintaining the original class balance of the data. (B) Two sets of hybrid data sets are composed based on the organism-specific reduced-data replicates: *Hybrid-A* maintains a fixed 50-50 balance between organism-specific and heterogeneous data at all data set sizes; *Hybrid-B* adds the required number of non-target organism observations to complete a data set of 1,000 peptides, and therefore results in sets with variable proportions of organism-specific peptides. (C) Baseline data sets composed of 1,000 exclusively non-target pathogen peptides are also generated based on different sub-samplings (without replacement) from the heterogeneous data. All data sets are used to train Random Forest models, which then have their performance assessed on organism-specific hold-out data.

For each data size tested (defined in this experiment as the number of organism-specific peptides in the data sets) both the *Hybrid-A* and *Hybrid-B* groups had the same number of replicates as the organism-specific sets of that size. The number of replicates at each size is documented in Table I.

Besides the hybrid data sets, for each target pathogen we also fit models on 30 samples of $1,000$ peptides from "other" organisms. In the results this is analysed as the limit case of the *Hybrid-B* data sets (as a "0+1000"-peptide set). Figure 1 illustrates the full experimental pipeline, including the generation of all relevant data sets.

B. Modelling and Performance Assessment

Epitope prediction models were developed by training Random Forest (RF) predictors on each of the training data sets outlined above, using Scikit-learn version 0.24.1 [50] under standard hyper-parameter values. The choice of Random Forest was based on preliminary experimentation, as documented [45], and also to make this work more directly comparable

TABLE I

SUMMARY OF ORGANISM-SPECIFIC DATA SETS: NUMBER OF POSITIVE / NEGATIVE PEPTIDES IN EACH SET, AND NUMBER OF REPLICATES FOR EACH SET SIZE (SET SIZE = NUMBER OF ORGANISM-SPECIFIC PEPTIDES IN THE SET). HYBRID-A AND HYBRID-B SETS WERE GENERATED BASED ON THE SAME SUBSETS OF ORGANISM-SPECIFIC PEPTIDES, AND THEREFORE HAVE THE SAME NUMBER OF REPLICATES AT EACH SIZE. HETEROGENEOUS SETS WERE GENERATED SEPARATELY, WITH 30 REPLICATES OF 1,000 NON-TARGET PATHOGEN PEPTIDES USED IN THE EXPERIMENTS.

	O. volvulus	Hepatitis C virus	Epstein-Barr virus
Hold-out peptides	(832+ / 777−)	(218+ / 358−)	(625+ / 315−)
Model dev. peptides	(2441+ / 2378−)	(919+ / 783−)	(1746+ / 811−)
20-peptide sets (N_{20})	237	83	124
40-peptide sets (N_{40})	118	41	62
60-peptide sets (N_{60})	79	27	42
80-peptide sets (N_{80})	59	21	31
100-peptide sets (N_{100})	47	17	25
150-peptide sets (N_{150})	32	11	16
200-peptide sets (N_{200})	24	8	12
250-peptide sets (N_{250})	19	6	10
300-peptide sets (N_{300})	16	5	8
400-peptide sets (N_{400})	12	4	6
500-peptide sets (N_{500})	9	4	5

with the results reported in that earlier one. The trained models were then used to generate predictions for the organism-specific hold-out data sets and prediction performance was assessed using multiple different performance measures, namely: Balanced Accuracy (BAL.ACC), Matthew's Correlation Coefficient (MCC), Area Under the Curve (AUC), Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Sensitivity (SENS). As these measures were calculated on the hold-out data sets (which were not seen by the models at any point other than testing) it can be assumed that these values represent a reasonable estimate of the generalisation performance of the models used for epitope prediction on proteins coming from each of the pathogens. The estimated mean performance and standard errors for each quality indicator were calculated from the replicates at each pathogen and data set size.

The new results are compared to a series of baselines [45]: the observed performance of Bepipred2.0 [17], LBtope [39], iBCE-EL [48] and ABCpred [6], on the hold-out set of each pathogen; the results obtained by Random Forest models trained on the full model development data and on a set of 6000 non-target pathogen peptides, for each organism.

III. RESULTS

Figures 2 and 3 display the mean performance results from each set of models on the hold-out data set of each pathogen.[1] Each figure plots the number of organism-specific peptides in the training data set *versus* the estimated mean performance according to different indicators.

For *Onchocerca volvulus* (the largest data set in this study), the highest scores on the hold-out set are from the full organism-specific model (except for sensitivity), as documented earlier [45]. The next highest scores are from the split-sampling organism-specific models (for data sizes $\geq$ 40 peptides), which approach the full organism-specific performance after about 150 peptides. A clear pattern can be seen across all performance measures: models trained on the organism-specific data sets consistently and uniformly outperform those trained on *Hybrid-A* (double size) data sets, which in turn outperform the models trained on *Hybrid-B* (1,000-peptides) data sets.[2] Within each group of tested models (trained on organism-specific, *Hybrid-A* and *Hybrid-B*) the pattern of performance improvement as the training set becomes larger was observed, as expected. The small-sample organism-specific models outperform those trained on the large heterogeneous (*Heter 6k*) and large hybrid (*OS-full+6k*) models (for $\geq$ 40 peptides), and also all models from the literature across all performance measures – except sensitivity, where Bepipred2.0 had the highest score; and NPV, where Bepipred2.0 outperformed the models trained with ≤ 40 organism-specific peptides.

A similar pattern can be noticed for the Epstein-Barr Virus data. Figure(s) 2 and 3 again show that, across all performance measures, the highest scores on the hold-out set are from the full organism-specific model, with the exceptions of positive predictive value (where LBtope has the highest score) and sensitivity, where the full organism-specific model and the reduced split-sampling organism-specific models have very similar scores across all training data sizes. The second highest scores across all performance measures are almost always the split-sampling organism-specific models (down to the smallest size: 20 peptides) apart from for AUC, where LBtope approaches the performance of the full organism-specific model. The overall pattern of our reduced data set EBV models is the same as that of the *Onchocerca volvulus* models: organism-specific $>$ *Hybrid-A* $>$ *Hybrid-B*. The performance of the models also generally decreases as the training data become more scarce, as expected. The EBV small-sample organism-specific models outperform all the tested models from the literature, as well as the *OS-full+6k* & *Heter 6k* results, across all performance measures except AUC and PPV, where LBtope yields better performane values.

The results for the Hepatitis C Virus reinforce the performance patterns observed for the other two pathogens. As documented [45], the apparently excellent performance of LBtope for this pathogen across all performance indicators can be partially attributed to the fact that several of the hold-out peptides used in this work are also part of LBtope's training data[3]. With the exception of LBtope's results, the pattern we observe for the Hepatitis C models closely mirrors the results on the other two pathogens, with organism-specific models generally outperforming the literature predictors tested even when trained with a very modest amount of peptides between 40 and 100, depending on the performance indicator.

[1]Tables containing the numerical performance estimates and standard error, as well as all experimental and analysis code, the data sets and results are available at https://github.com/fcampelo/orgspec_LBCE_limits.

[2]The largest data sets from both *Hybrid-A* and *Hybrid-B* always have very similar scores, which is expected, as in both cases the sets contain 500 peptides coming from the target pathogen data and 500 coming from the non-target pathogen data.

[3]https://webs.iiitd.edu.in/raghava/lbtope/data/LBtope_Variable_Positive_epitopes.txt

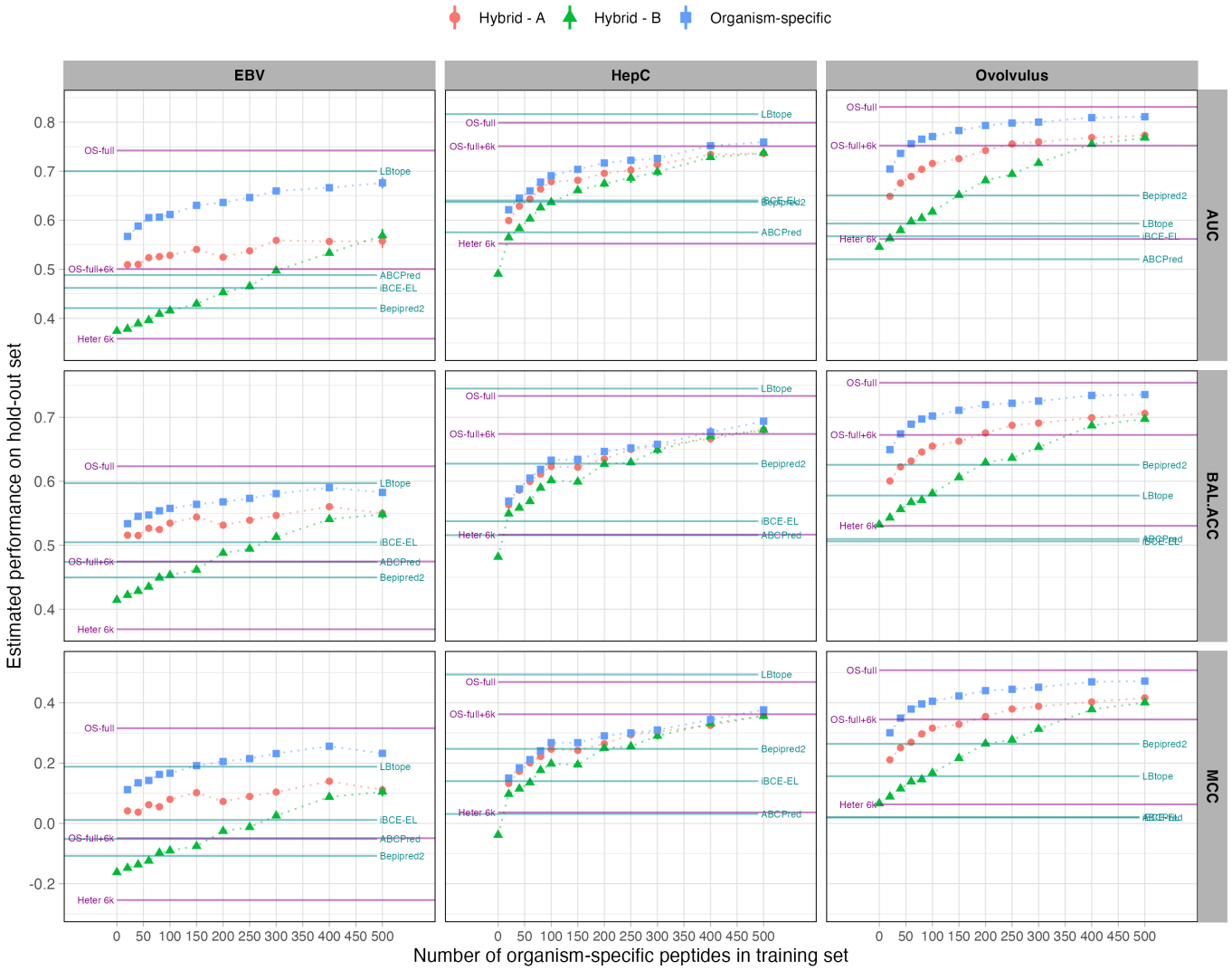

Fig. 2. Mean performance (AUC, balanced accuracy, MCC) and standard errors for all models tested. Blue triangles indicate the scores from models trained on organism-specific data sets, red crosses are from those trained on the *Hybrid-A* ("doubled data") data sets, and green squares refer to models trained on the *Hybrid-B* ("1000 peptides") data sets. Horizontal lines indicate reference values extracted from [45]: models trained on the full training set (*'OS-full'*), on a large heterogeneous set (*'Heter 6K'*), and on a large hybrid set (*'OS-full+6K'*), as well as the scores of several predictors from the literature on the same hold-out sets. For all pathogens tested, organism-specific training resulted in uniformly better performance across all data sizes when compared to models trained on hybrid or purely heterogeneous data, even when as few as 20 organism-specific peptides are used in the training set. Notice also how the performance of organism-specific models quickly surpasses that of most of the comparison predictors tested, even when very few organism-specific peptides are available to fit the models. (Note: standard error bars are in most cases shorter than the size of the point estimate markers)

For this pathogen, the performance difference within each group is considerably smaller than the differences that can be seen for the other organisms tested, albeit still with a clear trend of organism-specific models presenting the best performances and those trained with *Hybrid-B* the worst, for all performance indicators except PPV, where the three training regimens generally overlap across all data sizes.

When comparing all organism-specific reduced-data models scores to the heterogeneous model scores, across all organisms and for all performance measures, Figures 2-3 clearly show that almost all organism-specific models score considerably higher than the purely heterogeneous models (left-most point in the *Hybrid-B* group), as well as the hybrid models (*Hybrid-A & Hybrid-B*); the larger hybrid model from the previous study (*OS-full+6k*) and the generalist predictors from the literature, even when the organism-specific models are trained on modest-sized data sets. In all cases, prediction performance decreases as the number of organism-specific peptides used is reduced, even if the total number of peptides in the training set is kept fixed (*Hybrid-B*). For the organisms in this study, the organism-specific models also appear to be the most robust, with smaller performance decreases as the amount of organism-specific data is reduced when compared to *Hybrid-A* and, in particular, *Hybrid-B*.

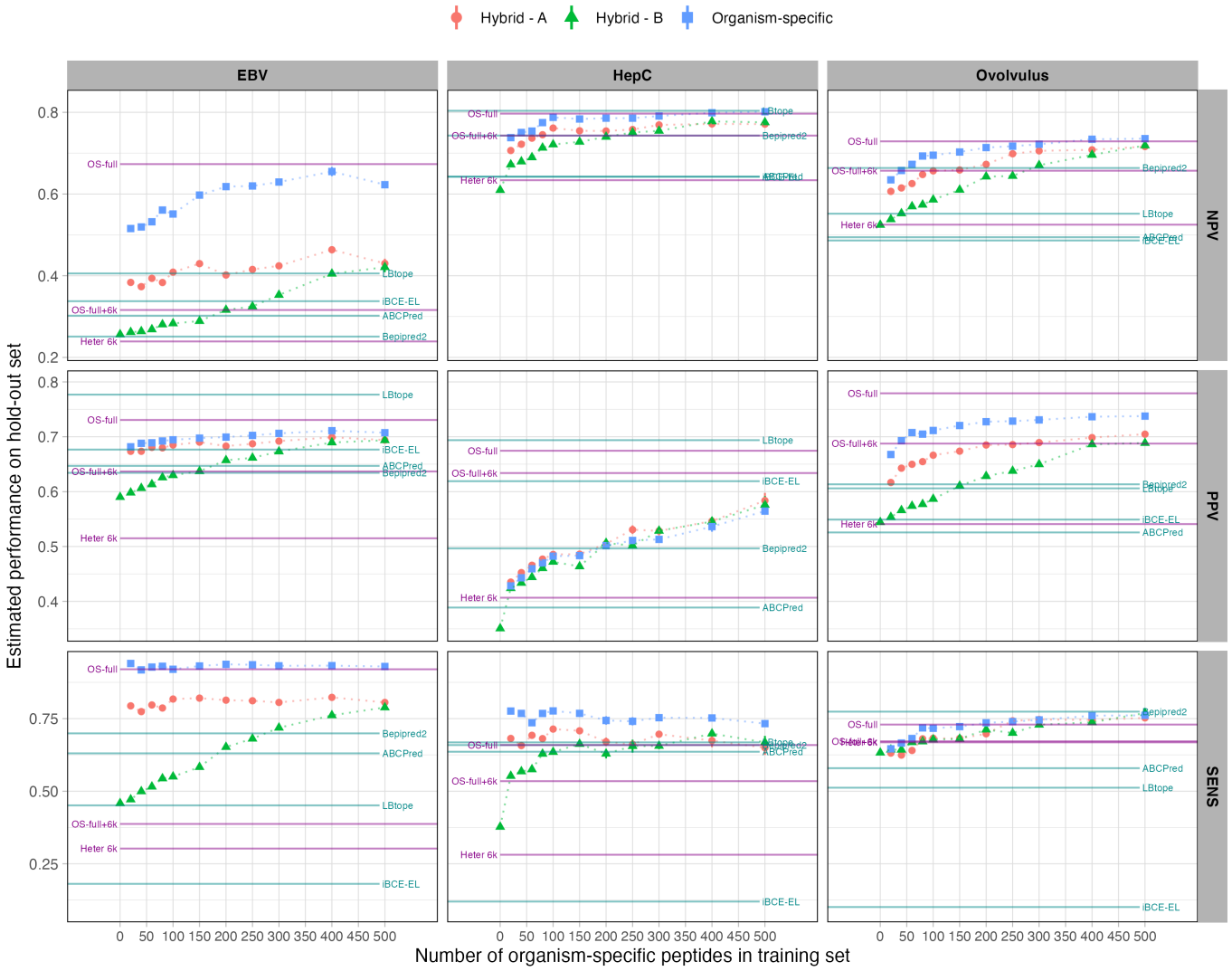

Fig. 3. (Continuing from Figure 2) Mean performance (PPV, NPV, Sensitivity) and standard errors for all models tested. The same pattern observed in figure 2 – uniform superiority of organism-specific training, when compared to models trained on bigger, but hybrid/heterogeneous, data sets – is also observed in the case of the three performance metrics shown here. As documented in [45], the apparently excellent performance of LBtope on the Hepatitis C data across all quality indicators can be partially attributed to the fact that several of the hold-out peptides used in this work are also found in LBtope's training data [39].

IV. Discussion

The results from this study indicate that, when compared to heterogeneous and hybrid training, organism-specific training produces higher linear B-cell epitope prediction performance scores, *even for very small data set sizes*. The number of organism-specific peptides in the training set is shown to strongly affect the predictive performance of organism-specific models across multiple performance indicators, particularly up to about 100 peptides, after which performance continues to increase with more data but with diminishing returns, asymptotically approaching that of models trained on the full available training data for each pathogen [45]. The results also show that organism-specific training outperforms generalist training (predictors from the literature, trained on peptides from a wide variety of pathogens) even when very small organism-specific data sets are available. The only systematic exception was the high observed performance of LBtope for the Hepatitis C Virus; However, as mentioned earlier, "*part of the hold-out examples used to asses the performance of the models is present in the training data of LBtope (9.59% of the Hep C hold-out sequences are present in the LBtope training data set)*" [45], which in the case of our experiments would result in some level of information leakage and an artificial inflation of that predictor's estimated performance. In addition to showing that organism-specific training outperforms heterogeneous and hybrid training, this work shows that adding unrelated data to organism-specific training sets *decreases* the generalisation performance of the resulting model when tasked with predicting epitopes for the target pathogen. It is also apparent that the more heterogeneous data is added to the

training set, the poorer the prediction performance becomes, which can be clearly seen from the comparison between *Hybrid-A* and *Hybrid-B* results in Figures 2-3. This suggests that, when training models for organism-specific predictions, the training data sets should be as specific (containing only labelled peptides from that organism) as possible.

Taken together, the results presented here provide a strong indication that organism-specific models trained on data sets beyond around 100 peptides provide very competitive predictive performance when compared to the generalist predictors tested. Additionally, the point at which organism-specific models start to outperform generalist predictors depends on the organism. For *O. volvulus* and Epstein-Barr Virus models the performance of organism-specific models compared favourably to that of generalist models down to the smallest organism-specific data set tested (20 peptides), while for Hepatitis C more peptides were required for the organism-specific training to become competitive.

This highlights the strengths of organism-specific training and extends the conclusions and scope of application of the methods described in our previous study [45], which were limited to data-rich organisms. In contrast, this study has shown that organism-specific training improves epitope prediction performance for data-poor organisms as well. As a comparison, the number of labelled peptide examples in the full training sets used in [45] were: 8,819 for *O. volvulus*, 2,557 for Epstein-Barr Virus, and 1,702 for Hepatitis C Virus. These are three of the most data-rich organisms on the IEDB. Currently, most organisms have far fewer labelled epitope examples available to them, and this work has shown that, for many if not most of these organisms, organism-specific training can provide significant improvements in prediction performance.

V. Conclusions

In a previous work, we showed that organism-specific training improves linear B-cell epitope prediction performance for data-abundant organisms. This work extends the scope of organism-specific modelling by showing that, contrary to our initial assumptions, organism-specific training is also a viable option for relatively data-poor organisms. However, it is clear that there are limits to organism-specific training for epitope prediction. The results documented in this study suggest that organism-specific models trained with more than about 100 labelled peptides will generally compare favourably to generalist predictors trained on substantially larger, but heterogeneous, data sets. It also confirms that predictive performance, across a wide variety of indicators, tends to increase monotonically with the number of organism-specific peptides included in the training data. It should be noted, however, that the results documented in this work have only been validated for reasonably class-balanced data sets. We have not tested models trained on strongly imbalanced data - the worst case among the pathogens tested was the Epstein-Barr virus data with a 2:1 balance of classes, which does not configure extreme class imbalance. While a further investigation of imbalanced classification approaches for epitope prediction would potentially help extend the scope of the organism-specific training framework even further, the results presented here, coupled with the increasingly cheap availability of computing power, already indicate a promising new direction for the development of bespoke predictors for pathogens under study, even for relatively data-poor organisms such as neglected pathogens or emerging health threats. Although synthetic data generation approaches such as SMOTE [51] could be potentially used to yield more balanced datasets, row dependencies emerging from the sequential nature of the data would require adaptations to enable SMOTE to perform better than simple minority oversampling, as indicated in preliminary computational experiments. Another promising line of research is to perform a more comprehensive investigation of the approach utilized in this work, not only in terms of testing on more pathogens but also investigating the effect of training models using data from phylogenetically-related (and potentially more well-studied) pathogens, which carries the potential of further extending the scope and applicability of tailored models to detect epitopes of understudied pathogens.

Acknowledgments

We would like to thank our collaborators Dr. Francisco Lobo (UFMG, Brazil) and Dr. João Reis-Cunha (University of York, UK) for excellent discussions and insights that contributed to the ideas explored in this paper. J.A was supported by the Engineering and Physical Sciences Research Council (EPSRC DTP grant EP/R512989/1). Experiments were run using Aston EPS Machine Learning Server (EPSRC Core Equipment Fund, Grant EP/V036106/1).

References

[1] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology*, 4th ed. New York: W.H.Freeman & Co Ltd, 2000.

[2] W. Paul, *Fundamental immunology*, 7th ed. London: Lippincott Williams & Wilkins, 2012.

[3] J. L. Sanchez-Trincado, M. Gomez-Perosanz, and P. A. Reche, "Fundamentals and methods for t- and b-cell epitope prediction," *Journal of Immunology Research*, vol. 2017, pp. 1–14, 2017. [Online]. Available: https://doi.org/10.1155/2017/2680160

[4] P. Leinikki, M. Lehtinen, H. Hyöty, P. Parkkonen, M.-L. Kantanen, and J. Hakulinen, "Synthetic peptides as diagnostic tools in virology," *Advances in virus research*, vol. 42, pp. 149–186, 1993.

[5] N. L Dudek, P. Perlmutter, I. Aguilar, N. P Croft, A. W Purcell *et al.*, "Epitope discovery and their use in peptide based vaccines," *Current pharmaceutical design*, vol. 16, no. 28, pp. 3149–3157, 2010.

[6] L. Potocnakova, M. Bhide, and L. B. Pulzova, "An introduction to b-cell epitope mapping and in silico epitope prediction," *Journal of Immunology Research*, vol. 2016, pp. 1–11, 2016. [Online]. Available: https://doi.org/10.1155/2016/6760830

[7] T. J. Kindt, R. A. Goldsby, B. A. Osborne, and J. Kuby, *Kuby immunology*, 7th ed. ,: Macmillan, 2007.

[8] A. Kolaskar and P. C. Tongaonkar, "A semi-empirical method for prediction of antigenic determinants on protein antigens," *FEBS letters*, vol. 276, no. 1-2, pp. 172–174, 1990.

[9] M. Odorico and J.-L. Pellequer, "Bepitope: predicting the location of continuous epitopes and patterns in proteins," *Journal of Molecular Recognition*, vol. 16, no. 1, pp. 20–22, 2003.

[10] J. E. P. Larsen, O. Lund, and M. Nielsen, "Improved method for predicting linear b-cell epitopes," *Immunome research*, vol. 2, no. 1, p. 2, 2006.

[11] S. Saha and G. P. S. Raghava, "Bcepred: prediction of continuous b-cell epitopes in antigenic sequences using physico-chemical properties," in *Proc. International Conference on Artificial Immune Systems*. Catania, Italy: Springer, 2004, pp. 197–204.

[12] ——, "Prediction of continuous b-cell epitopes in an antigen using recurrent neural network," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 1, pp. 40–48, 2006.

[13] Y. EL-Manzalawy, D. Dobbs, and V. Honavar, "Predicting linear b-cell epitopes using string kernels," *Journal of Molecular Recognition: An Interdisciplinary Journal*, vol. 21, no. 4, pp. 243–255, 2008.

[14] M. J. Blythe and D. R. Flower, "Benchmarking b cell epitope prediction: underperformance of existing methods," *Protein Science*, vol. 14, no. 1, pp. 246–248, 2005.

[15] H.-W. Wang, Y.-C. Lin, T.-W. Pai, and H.-T. Chang, "Prediction of b-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification," *Journal of Biomedicine and Biotechnology*, vol. 2011, pp. 1–12, 2011. [Online]. Available: https://doi.org/10.1155/2011/432830

[16] B. Yao, D. Zheng, S. Liang, and C. Zhang, "Conformational b-cell epitope prediction on antigen protein structures: A review of current algorithms and comparison with common binding site prediction methods," *PLoS ONE*, vol. 8, no. 4, p. e62249, Apr. 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0062249

[17] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, "Bepipred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes," *Nucleic acids research*, vol. 45, no. W1, pp. W24–W29, 2017.

[18] M. Collatz, F. Mock, E. Barth, M. Hölzer, K. Sachse, and M. Marz, "Epidope: A deep neural network for linear b-cell epitope prediction," *Bioinformatics*, vol. 37, no. 4, pp. 448–455, 2021.

[19] M. H. Van Regenmortel, "Mapping epitope structure and activity: from one-dimensional prediction to four-dimensional description of antigenic specificity," *Methods*, vol. 9, no. 3, pp. 465–472, 1996.

[20] Y.-T. Lo, T.-W. Pai, W.-K. Wu, and H.-T. Chang, "Prediction of conformational epitopes with the use of a knowledge-based energy function and geometrically related neighboring residue characteristics," *BMC bioinformatics*, vol. 14, no. S4, p. S3, 2013.

[21] R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, and B. Peters, "The immune epitope database (iedb): 2018 update," *Nucleic acids research*, vol. 47, no. D1, pp. D339–D343, 2019.

[22] S. Saha, M. Bhasin, and G. P. Raghava, "Bcipep: a database of b-cell epitopes," *BMC genomics*, vol. 6, no. 1, pp. 1–7, 2005.

[23] C. P. Toseland, D. J. Clayton, H. McSparron, S. L. Hemsley, M. J. Blythe, K. Paine, I. A. Doytchinova, P. Guan, C. K. Hattotuwagama, and D. R. Flower, "Antijen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data," *Immunome research*, vol. 1, no. 1, pp. 1–12, 2005.

[24] B. T. Foley, T. K. Leitner, C. Apetrei, B. Hahn, I. Mizrachi, J. Mullins, A. Rambaut, S. Wolinsky, and B. T. M. Korber, "Hiv sequence compendium 2021," Los Alamos National Laboratory, Los Alamos, NM (United States), Tech. Rep., 2021.

[25] X. Yang and X. Yu, "An introduction to epitope prediction methods and software," *Reviews in medical virology*, vol. 19, no. 2, pp. 77–96, 2009.

[26] U. Reineke and M. Schutkowski, *Epitope mapping protocols*. New York, NY: Springer, 2009, vol. 1.

[27] B. F. Arnold, H. M. Scobie, J. W. Priest, and P. J. Lammie, "Integrated serologic surveillance of population immunity and disease transmission," *Emerging infectious diseases*, vol. 24, no. 7, p. 1188, 2018.

[28] M. C. Jespersen, S. Mahajan, B. Peters, M. Nielsen, and P. Marcatili, "Antibody specific b-cell epitope predictions: leveraging information from antibody-antigen protein complexes," *Frontiers in immunology*, vol. 10, p. 298, 2019.

[29] J. A. Greenbaum, P. H. Andersen, M. Blythe, H.-H. Bui, R. E. Cachau, J. Crowe, M. Davies, A. Kolaskar, O. Lund, S. Morrison *et al.*, "Towards a consensus on datasets and evaluation metrics for developing b-cell epitope prediction tools," *Journal of Molecular Recognition: An Interdisciplinary Journal*, vol. 20, no. 2, pp. 75–82, 2007.

[30] I. Steinwart and A. Christmann, *Support vector machines*. New York, NY: Springer Science & Business Media, 2008.

[31] J. Chen, H. Liu, J. Yang, and K.-C. Chou, "Prediction of linear b-cell epitopes using amino acid pair antigenicity scale," *Amino acids*, vol. 33, no. 3, pp. 423–428, 2007.

[32] M. J. Sweredoski and P. Baldi, "Cobepro: a novel system for predicting continuous b-cell epitopes," *Protein Engineering, Design & Selection*, vol. 22, no. 3, pp. 113–120, 2009.

[33] L. J. Wee, D. Simarmata, Y.-W. Kam, L. F. Ng, and J. C. Tong, "SVM-based prediction of linear b-cell epitopes using bayes feature extraction," *BMC Genomics*, vol. 11, no. S4, p. S21, Dec. 2010. [Online]. Available: https://doi.org/10.1186/1471-2164-11-s4-s21

[34] Y. Wang, W. Wu, N. N. Negre, K. P. White, L. Cheng, and P. K. Shah, "Determinants of antigenicity and specificity in immune response for protein sequences," *BMC Bioinformatics*, vol. 12, p. 251, 2011.

[35] J. Gao, E. Faraggi, Y. Zhou, J. Ruan, and L. Kurgan, "Best: Improved prediction of b-cell epitopes from antigen sequences," *Plos One*, vol. 7, no. 6, p. e40104, 2012.

[36] S. Y.-H. Lin, C.-W. Cheng, and E. C.-Y. Su, "Prediction of b-cell epitopes using evolutionary information and propensity scales," *BMC bioinformatics*, vol. 14, no. 2, pp. 1–9, 2013.

[37] W. Shen, Y. Cao, L. Cha, X. Zhang, X. Ying, W. Zhang, K. Ge, W. Li, and L. Zhong, "Predicting linear b-cell epitopes using amino acid anchoring pair composition," *BioData mining*, vol. 8, no. 1, pp. 1–12, 2015.

[38] B. Yao, L. Zhang, S. Liang, and C. Zhang, "Svmtrip: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity," *PLoS One*, vol. 7, p. e45152, 2012.

[39] H. Singh, H. R. Ansari, and G. P. S. Raghava, "Improved method for linear b-cell epitope prediction using antigen's primary sequence," *PLoS ONE*, vol. 8, no. 5, p. e62216, May 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0062216

[40] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] Y. EL-Manzalawy and V. Honavar, "Building classifier ensembles for b-cell epitope prediction," *Methods in Molecular Biology*, vol. 1184, pp. 285–294, 2014. [Online]. Available: https://doi.org/10.1007/978-1-4939-1115-8_15

[42] V. Saravanan and N. Gautham, "Harnessing computational biology for exact linear b-cell epitope prediction: a novel amino acid composition-based feature descriptor," *Omics: a journal of integrative biology*, vol. 19, no. 10, pp. 648–658, 2015.

[43] P. Haste Andersen, M. Nielsen, and O. Lund, "Prediction of residues in discontinuous b-cell epitopes using protein 3d structures," *Protein Science*, vol. 15, no. 11, pp. 2558–2567, 2006.

[44] E.-M. Yasser and V. Honavar, "Recent advances in b-cell epitope prediction methods," *Immunome research*, vol. 6, no. 2, pp. 1–9, 2010.

[45] J. Ashford, J. Reis-Cunha, I. Lobo, F. Lobo, and F. Campelo, "Organism-specific training improves performance of linear b-cell epitope prediction," *Bioinformatics*, vol. 37, p. 4826–4834, 2021.

[46] R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, and B. Peters, "The immune epitope database (iedb): 2018 update," *Nucleic acids research*, vol. 47, no. D1, pp. D339–D343, 2019.

[47] F. Campelo, J. Reis-Cunha, J. Ashford, A. Ekárt, and F. P. Lobo, "Phylogeny-aware linear b-cell epitope predictor detects candidate targets for specific immune responses to monkeypox virus," *bioRxiv preprint*, 2022. [Online]. Available: https://doi.org/10.1101%2F2022.09.08.507179

[48] B. Manavalan, R. G. Govindaraj, T. H. Shin, M. O. Kim, and G. Lee, "ibce-el: a new ensemble learning framework for improved linear b-cell epitope prediction," *Frontiers in immunology*, vol. 9, p. 1695, 2018.

[49] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[51] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.