EpitopeTransfer: a Phylogeny-aware Transfer Learning Framework for Taxon-specific Linear B-cell Epitope Prediction

Lindeberg Pessoa Leite[1,2,a], Teófilo Emidio de Campos[1,3,b], Francisco Pereira Lobo[4,5,c], Felipe Campelo[6,*,d]

[1] *Department of Computer Science, University of Brasília, Brasília DF 70910-900, Brazil.*

[2] *Directorate of Information Technology and Innovation, Federal Police, Brasília DF 70610-902, Brazil.*

[3] *Applied Sciences Group, Microsoft, Reading RG6 1WG, UK.*

[4] *Department of Genetics, Ecology and Evolution, Federal University of Minas Gerais, Belo Horizonte MG 31270-901, Brazil.*

[5] *NCI-Frederick, National Cancer Institute, Frederick MD 21702, USA.*

[6] *School of Engineering Mathematics and Technology, University of Bristol, Bristol BS8 1TW, UK.*

[*] Corresponding author.

E-mail: f.campelo@bristol.ac.uk (Campelo F).

Running title: *Leite LP et al / EpitopeTransfer*

[a] ORCID: 0000-0002-9728-0955.

[b] ORCID: 0000-0001-6172-0229.

[c] ORCID: 0000-0002-9671-9313.

[d] ORCID: 0000-0001-8432-4325.

Total word count: 3693

Total figures: 2

Total tables: 0

Total supplementary figures: 2

Total supplementary tables: 5

Total supplementary files: 1

Abstract

The identification of linear B-cell epitopes (LBCEs) plays a pivotal role in the development of immunological products such as immunodiagnostic, vaccines and therapeutic antibodies. This criticality has led to the development of computational approaches for the prediction of LBCEs, aiming at accelerating discovery and prioritising targets for experimental assessment. Existing LBCE predictors rely on having access to large volumes of data from a wide range of pathogens to train generalist predictive models, which can result in biases towards widely studied organisms and compromise performance when focusing on neglected or emerging pathogens. We introduce EpitopeTransfer, a phylogeny-aware transfer learning strategy that results in noticeable gains in expected predictive performance. This is achieved by fine tuning large protein language models using abundant data from higher-level taxa and applying this refined feature embedder for fitting pathogen- or lower taxon-specific models. We report substantially increased performance in comparison to state-of-the-art approaches for LBCE prediction across a broad spectrum of pathogens (viruses, bacteria and eukaryotes), and show that these gains can be directly attributed to the use of phylogeny-aware fine tuning of the feature embedder coupled with pathogen- or taxon-optimized modelling.

KEYWORDS: Phylogeny-aware methods; Linear B-cell epitopes; Deep learning; Transfer learning.

Introduction

The humoral immune response of mammals depends on antibodies, which play a crucial role in identifying and attaching to antigens [1]. B-cell epitopes (BCEs) are parts of an antigenic molecule that bind to an antibody [2]. BCEs can emerge from contiguous sequences of amino acid residues within the primary protein structure or from disconnected regions that are brought together by protein folding. The former are commonly referred to as *linear B-cell epitopes* (LBCEs) whereas the latter are *conformational* BCEs. Identifying epitopes is a crucial step in a broad range of medical and immunological applications, including vaccines [3], therapeutic antibodies [4], and immunodiagnostics [5].

Deep learning techniques are becoming increasingly common for modelling and analysing protein data [6–10], with transfer learning emerging as a notable approach [11–13]. Several methods involve using latent space vector representations of amino acid residues that are extracted from large, pretrained protein language models. These representations encode complex biological properties of proteins in a context-dependent manner [14–16], making them a compelling option for the development of new models capable of capturing the immunogenic properties of peptides. As a particular example, the Evolutionary Scale Models (ESM) for protein representation [10,14,17] employ self-supervised learning and are based on the Transformer architecture, which has proven to be a powerful general-purpose framework for representation learning and generative modelling, outperforming recurrent and convolutional architectures in natural language domains [18]. These models provide powerful representations of protein properties that encode useful information for a variety of downstream modelling tasks, including the prediction of B-cell epitopes [19].

Most LBCE prediction models, including recent ones leveraging large protein language models [19,20], are trained using datasets containing labelled peptide sequences from a phylogenetically diverse range of organisms. As we have previously discussed [21,22], the main goal of epitope prediction models developed using such

heterogeneous datasets is to provide general-purpose models that can be readily used without referencing the source organism of the proteins or peptides queried. There is, however, mounting evidence that pathogen- or taxon-specific models can result in improved performance in the usual scenario where predictions are made with a specific target pathogen or taxon of interest [21–26], e.g., for the development of immunodiagnostic tests [27] or vaccines.

Here we present EpitopeTransfer, a deep learning embedder-based framework for instantiating taxon-specific models for epitope prediction. We propose the fine tuning of ESM models on relatively abundant LBCE data at higher taxonomic levels to enhance downstream LBCE prediction tasks at lower levels, with models then fit and optimised specifically to maximise predictive performance for a specific target pathogen or taxon of interest.

Method

We developed a general and modular phylogeny-aware modelling approach called *EpitopeTransfer* for the generation of taxon-specific linear B-cell epitope predictors. The data sources and modelling steps are outlined in Figure 1. The performance of this approach was assessed using a phylogenetically diverse set of nineteen target taxa of interest including viruses, bacteria, and eukaryotic pathogens. Details regarding these datasets including data retrieval, quality control, data consolidation, the total volume of data and the criteria for deriving the datasets used in this work are provided in Supplementary File 1 (Section “Data”) and Supplementary Tables 1-2.

Modelling

The framework for instantiating taxon-specific predictors consists of three main components (see Figure 1). Given a pair of datasets for matched {higher, lower} level taxa, the model development tasks are as follows.

The first step is *Embedder development*, consisting of the fine tuning of the general-purpose ESM protein language model to the task of LBCE prediction, as detailed in Supplementary File 1, section “Embedder Development Details”. We tested both the ESM2 and ESM-1b versions, which yielded similar results (see section “Results”). The ESM embedder is re-trained using the entries from the higher-level dataset, resulting in a protein embedder fine-tuned for that particular phylogenetic group. Through this fine-tuning step, the structure learned as part of the ESM model's primary development [14] is augmented with knowledge about the representation of LBCEs from that particular higher-level taxon. In this work we used the 650M parameter versions of ESM-1b and ESM2 as base models; however, adapting the framework to use different embedder model sizes is a straightforward task.

The fine-tuned embedder is deployed to extract features (*Feature calculation*) for the sequences from the target (lower-level) taxon. In the feature generation process, the full lower-level protein sequences are fed to the tuned embedder rather than just the labelled peptides, which enables the fine-tuned ESM embedder to capture richer, non-local contextual information. This results in an enhanced feature representation for each residue position. The output of this feature calculation model is then reduced so that only the labelled peptide regions are selected for later classifier training and performance assessment.

The last step is the *Predictive model training*, in which the features generated in the previous step are used to fit and optimise the hyperparameters of a classification model, resulting in a bespoke LBCE prediction model for the target taxon. In this work, a Random Forest classifier [33] was used to develop the final predictor.

Performance assessment and comparison

To assess the predictive performance of the models and ensure comparability with existing literature, we focus on the Area Under the ROC curve (AUC) [34] as our main performance indicator for the evaluation of all predictors. The definition of this metric as well as several other commonly used performance indicators is provided in Supplementary File 1, section “Performance Indicators”. Besides AUC, we report

eight additional performance measures in Supplementary Tables 3–5 for broader comparability of results with the literature: Accuracy, Balanced accuracy, F_1 score, Matthews Correlation Coefficient (MCC), Positive and Negative Predictive Values (PPV and NPV, respectively), Sensitivity and Specificity.

We created two sets of internal baseline models to investigate the effect of the proposed approach on predictive performance. The first one used the general-purpose ESM embedder, without any fine tuning, as the feature calculation engine (*ESM baseline*). Models based on raw ESM features were developed to allow us to estimate the effect of the fine-tuning step by contrasting their predictive performance against models created using the full EpitopeTransfer procedure, which employs fine-tuned embedders.

The second set of internal baseline models was created using precisely the same methodology as EpitopeTransfer but breaking the phylogenetic link between the higher-level dataset (used to fine-tune the ESM embedder) and the target taxon. This was achieved by using data of pathogens from superkingdoms distinct from the target taxon, i.e., with only a very remote phylogenetic link (Supplementary Figure 1). These non-phylogenetic transfer baselines (*NPTransfer*) were created to allow the estimation of the effect of using phylogeny-aware filter when selecting data for the fine-tuning of the ESM embedder.

In addition to the internal baselines listed above, four state-of-the-art LBCE predictors were used as external comparison baselines: BepiPred 3.0 [19], EpiDope [6] and EpitopeVec [7], which are deep learning-based generalist approaches using sophisticated feature spaces for LBCE prediction; and epitope1D [23], a recent approach which provides taxon-specific models, commonly stratified at the phylum or kingdom level. These external baselines were selected as representative of the current state-of-the-art in LBCE prediction which provide online interfaces or easily reproducible source code.

All performance values presented in the Results section correspond to results estimated using out-of-sample predictions. In the case of EpitopeTransfer and our internal baselines, this represents the observed performance on an independently

extracted final test set for each target taxon (referred to in the text as “EpitopeTransfer holdout”), which was not used during any part of our embedder or model development. Supplementary File 1 (section “Training and Validation Procedures for EpitopeTransfer”) provides details on the data splitting and other aspects of model development and assessment for EpitopeTransfer and the internal baselines.

Since the peptides in the EpitopeTransfer holdout sets are not necessarily outside the training data of the different external baselines used, we assessed these tools primarily using observations of each target taxon that are not found in these methods’ published training datasets. This approach is more likely to provide unbiased estimates of their generalization performance for each target taxa used in our experiments. We have additionally calculated the performance of the external baseline methods on the EpitopeTransfer holdout set and report those for reference in the Results section.

To account for the effects of different methods, embedders, and datasets, we compared the performance of EpitopeTransfer with the internal baselines using a factorial analysis of variance (ANOVA) design with *Method* and *Embedder* as experimental factors, and *Dataset* as a blocking factor [35]. Post-hoc pairwise tests for the differences of mean performance of different methods (accounting for *Embedder* and *Dataset* effects) were conducted using Dunnet's all-vs-one contrasts [35], with EpitopeTransfer as the reference method for comparison.

For the comparison of the selected methodology (EpitopeTransfer with the ESM2 embedder) against the external baseline methods, we employed a similar ANOVA design, with *Method* as the experimental factor and *Dataset* as a blocking variable. The same type of post-hoc comparison was used to estimate the pairwise performance differences and p-values. In all comparisons, inspection of the model residuals indicated good adherence to the statistical assumptions of ANOVA.

Results

Phylogeny-aware transfer learning improves LBCE prediction

The central objective of this work was to investigate the effect of phylogeny-aware transfer learning on the performance of taxon-specific LBCE predictors. This was assessed by contrasting the performance of EpitopeTransfer against baseline models developed following an identical procedure, changing only the fine-tuning process of the feature embedder in two manners, as detailed in section "Method": without embedder fine tuning (*ESM baseline*) and with embedder fine tuning but without a phylogenetic data filtering (*NPTransfer*).

Using AUC as our main performance indicator, the results of the ANOVA test indicated a significant difference of Methods ($p = 0.0035$) and did not detect any significant difference between Embedders ($p = 0.3183$). In the post-hoc comparisons between different methods, we observed statistically significant AUC gains of the EpitopeTransfer methodology in relation to both internal baselines. When compared against the ESM baseline, EpitopeTransfer yields significant gains ($p = 0.004$; $\delta = 0.032$; $CI_{95\%} = (0.009,0.055)$; Figure 2(a)).[1] We also observed reasonable effect sizes in terms of other performance indicators (Supplementary Table 3), most notably the Matthews Correlation Coefficient (MCC; $p = 0.0016$; $\delta = 0.095$), which is a robust, class imbalance-resistant performance indicator [37]. These results indicate that the fine-tuning step induced relevant performance gains, well balanced across positive and negative predictions.

One of the core hypotheses underlying the development of EpitopeTransfer is that LBCE data from phylogenetically closer pathogens carries more relevant information for training taxon-specific predictors than similar data from more distantly related organisms. We tested this hypothesis by contrasting the performance of EpitopeTransfer against the NPTransfer baseline, consisting of models created using the same procedure as EpitopeTransfer but with the embedder fine tuning performed

[1] The effect size δ refers to the mean difference of performance after removal of the embedder and dataset effects.

using exclusively data from the most distantly related taxa (subsection "Performance assessment and comparison" under section "Methods"; Supplementary Figure 1).

The results obtained in the comparison between EpitopeTransfer and NPTransfer baseline indicate a clear gain in performance resulting from the incorporation of a phylogeny-aware data filtering strategy for the fine-tuning step. We observed a statistically significant increase in AUC ($p = 0.0105$; $\delta = 0.029$; $CI_{95\%} = (0.006,0.052)$; Figure 2(a)). Gains in terms of other metrics were also relevant (Supplementary Table 3), with substantial gains of, e.g., MCC ($p = 0.0004$; $\delta = 0.107$). Taken in combination with the observed improvement over the ESM baseline, these results provide compelling evidence that the superior performance of EpitopeTransfer due to the fine-tuning of the feature embedder using specifically data from pathogens that are phylogenetically close to the target taxon, rather than being the effect of simply refining the protein language model for the general epitope prediction task.

EpitopeTransfer yields better predictive models than state-of-the-art approaches

Given the absence of an Embedder effect in the initial comparison, any of the two EpitopeTransfer variations – EpitopeTransfer (ESM-1b) and EpitopeTransfer (ESM2) could have been selected for further comparisons against the external baselines. We decided to use the ESM2 version, since it represents a more recent version of the ESM embedder which is likely to be maintained for longer than ESM-1b. For notational simplicity, we will refer to this ESM2-based version as simply *EpitopeTransfer* from now on.

We compared the results obtained by EpitopeTransfer against the four state-of-the-art LBCE predictors. Three of those are deep learning-based, generalist LCBE predictors: BepiPred 3.0 [19], which also employs the ESM2 protein embedder as part of its pipeline, augmented by surface accessibility and protein length features;[2] Epidope [6], which utilizes a deep neural network sequence embedder known as ElMo [32]; and EpitopeVec [7], which uses a combination of amino acid features, modified

[2] For all comparisons, we used BepiPred 3.0's *linear epitope score* as that model's output.

antigenicity scales and neural network-based protein embeddings. The fourth external comparison baseline, epitope1D [23], implements a rough taxon-specific modelling approach, with models provided for major phyla or kingdoms of bacterial, viral, and eukaryotic pathogens. epitope1D is based on a bespoke graph-based feature space extracted from a set of physicochemical and statistical amino acid descriptors.

The main purpose of these comparisons was to establish whether the specific combination of components used in the EpitopeTransfer modelling pipeline is sufficient to produce taxon-specific predictors with an expected performance comparable or superior to the best current predictive methods. As can be seen in Figure 2(b), we found statistically significant differences in mean AUC between EpitopeTransfer and all four external baselines, with effect sizes noticeably greater than the ones observed in the comparisons against the ESM2 and NPTransfer baselines. Supplementary Table 4 provides full numerical details for the comparisons of AUC and other performance indicators.

The observed performance gains of EpitopeTransfer were substantial in relation to the three generalist baselines: Epidope ($p = 0.0017$; $\delta = 0.130$; $CI_{95\%} = (0.069,0.190)$), BepiPred 3.0 ($p = 1.10 \times 10^{-5}$; $\delta = 0.090$; $CI_{95\%} = (0.029,0.150)$) and EpitopeVec ($p = 1.86 \times 10^{-6}$; $\delta = 0.120$; $CI_{95\%} = (0.038,0.202)$). Similarly large AUC gains were found in the comparison against epitope1D, which also uses taxon-specific models albeit at a coarser granularity than EpitopeTransfer ($p = 3.70 \times 10^{-6}$; $\delta = 0.132$; $CI_{95\%} = (0.071,0.193)$).

Figure 2(c) shows the distribution of the magnitude of paired differences of AUC for each dataset, stratified in terms of performance gains of EpitopeTransfer. Consistently with the results reported above, we observed performance gains in most comparisons, with occasional instances of performance degradation against BepiPred 3.0 and Epidope. The average magnitudes of gains are also considerably greater than those of performance decreases, as shown in Figure 2(c). For twelve of the nineteen datasets used in this experiment, EpitopeTransfer exhibited positive AUC gains simultaneously against all external baselines; in nine of these, the AUC gains were all greater than our minimally relevant effect size of +0.05. These cases are explicitly

labelled in Figure 2(c). Importantly, these include models across the three domains explored: viral, bacterial, and eukaryotic pathogens. Five of these models exhibiting relevant AUC gains correspond to specific pathogenic species; the other four were derived for different taxonomic levels.

EpitopeTransfer provides high-performance LBCE predictors across taxonomically diverse pathogens

Some of the models instantiated in this study present particularly high predictive performance, making them potential models of choice for researchers investigating certain pathogens.

The model for the Filoviridae family, which contains highly virulent pathogens such as the Ebola and Marburg viruses [39], is the one where observed performance differences are most prominent. As shown in Figure 2(c), this model resulted in absolute gains in AUC greater than 0.4 in relation to BepiPred 3.0, EpitopeVec and Epidope, and greater than 0.35 in comparison to Epitope1D. Supplementary Table 5 reports the full performance values observed for this model, which displays an out-of-sample AUC of 0.96 and an MCC of 0.61, suggesting it as a robust LBCE predictor for these viruses. Other EpitopeTransfer models where high performance gains were observed include the ones derived for *Escherichia coli* (AUC = 0.91; MCC = 0.44), *Chlamydia trachomatis* (AUC = 0.83; MCC = 0.57) and *Plasmodium falciparum* (AUC = 0.79; MCC = 0.41). These models, as well as the others nominally labelled in Figure 3(c), were shown to provide LBCE predictions with considerably better results for their target taxa than all current state-of-the-art.

On the other extreme, we observed a few cases for which we would not recommend the EpitopeTransfer models derived in this work, at least in comparison to existing approaches. Only a single dataset, *Mycobacterium tuberculosis*, resulted in EpitopeTransfer exhibiting inferior AUC performance against all external baselines. This appears to be a particularly challenging dataset, with all methods tested exhibiting poor AUC: 0.52 for EpitopeTransfer, 0.54 for Epitope1D, 0.55 for EpitopeVec, and 0.59 for Epidope. The best model for this specific dataset, BepiPred

3.0, still had a modest AUC of 0.61. Other specific datasets for which EpitopeTransfer did not perform well were *Bordetella pertussis* (AUC = 0.53) and *Schistosoma mansoni* (AUC = 0.54). For the two other datasets explicitly labelled in the negative gains part of Figure 3(c), the comparatively low performance against BepiPred 3.0 stems mostly from that method's solid performance for those datasets rather than to poor performance by EpitopeTransfer, as documented in Supplementary Table 5.

Discussion

In this work we presented EpitopeTransfer, a general modelling framework to instantiate taxon-specific predictors of LBCEs using a phylogeny-aware fine-tuning of the ESM2 protein embedder, coupled with a model fitting approach which fine-tunes the predictive model to any target taxon of interest. The fact that fine-tuning of protein language models improves performance in a variety of downstream tasks is well-established [13]. However, as far as we are aware, no comprehensive evaluation has been conducted on the impact of explicitly incorporating phylogenetic information into such models. By contrasting the performance of EpitopeTransfer with similar models developed without the fine-tuning step (*ESM baseline*) and with the tuning step executed using data from more distantly related pathogens (*NPTransfer baseline*) we were able to establish that the strong predictive performance observed by models instantiated from the EpitopeTransfer framework was mainly due to the phylogeny-aware data selection for the refinement of the ESM protein language model.

We also established that EpitopeTransfer models provide substantial mean performance gains in AUC (between +0.09 and +0.123) over both generalist LBCE predictors [6,7,19] and a recent taxon-specific predictor [23]. In its current version, the EpitopeTransfer framework can serve as the basis to generate bespoke models for the prediction of LBCEs in a wide range of pathogens, including viruses, bacteria, and eukaryotic pathogens. By leveraging evolutionary relationships, it is possible to create optimised models for both well-studied pathogens and for emerging or neglected

infectious diseases, through the selection of the most appropriate lower-level taxon from a data availability perspective. Possibly the major exception are fungal pathogens, for which there appears to be a scarcity of well-curated LBCE data.

One potential limitation of our method is the occurrence of horizontal gene transfer (HGT), as horizontally acquired genes are not vertically shared. In this scenario, LBCEs located within such genes may escape detection by our phylogeny-based modelling strategy. This limitation is expected to become more pronounced as phylogenetic distance increases. However, recent large-scale genomic studies have shown that HGT is substantially more frequent among closely related taxa, with its frequency declining sharply as phylogenetic distance increases [41]. This phenomenon is likely driven by many factors, including genetic compatibility, ecological overlap, and shared cellular machinery. These observations suggest that, while HGT could theoretically limit the effectiveness of phylogeny-aware models, its practical impact on EpitopeTransfer is likely to be minor. Since most HGT events occur among phylogenetically close species and, therefore, likely within the same broader taxonomic groups used for model training, our framework should remain well-suited to capture epitope-relevant features in such cases.

The EpitopeTransfer framework presented here is completely modular, making it easily adaptable to other similar predictive tasks, such as predicting T-cell epitopes or identifying conformational BCEs. By adapting the data sources or the feature embedding block (e.g., incorporating protein embedders capable of calculating conformational or functional features into the pipeline) phylogeny-aware predictive frameworks such as EpitopeTransfer present the potential of accelerating discovery in the context of immunodiagnostics, vaccines and immunotherapeutics against current and future infectious diseases. This modularity also facilitates the exploration of modelling alternatives, e.g., using different classification approaches such as gradient boosting or feedforward neural networks to build the predictors. Although preliminary testing (not shown in this paper) did not point to systematic gains of these more sophisticated modelling approaches over the comparatively simple Random Forest classifiers, it is possible that they may provide performance advantages in other

potential applications of our phylogeny-aware transfer learning approach. Similarly, further refinements in the data consolidation and preprocessing steps (including, e.g., an investigation on alternative approaches for peptide merging and label conflict resolution) could potentially yield performance gains for LBCE prediction.

The models developed using EpitopeTransfer are adapted from broader to more specific domains using partially labelled protein data from the source domain (full protein sequences from the higher-level taxon containing some labelled peptides) to fine-tune the protein language model; and using the less abundant labelled samples from the target domain to train a pathogen- or taxon-specific classifier. Our framework may be categorised as a semi-supervised domain adaptation approach, because a relatively small set of labels are used in the target domains [42]. In principle, this could be done even in the absence of labelled data in the target domain, which would make EpitopeTransfer extensible to understudied but nonetheless critical cases such as fungal pathogens.[3] However, the performance of the framework for such cases remains to be investigated.

Conclusion

The accurate prediction of linear B-cell epitopes is critical for vaccine design, therapeutic antibody development, and immunodiagnostics. Nevertheless, as we demonstrated in this work, existing general-purpose computational methods often present sub-optimal performance for certain pathogens. To address this, we introduced EpitopeTransfer, a general and modular phylogeny-aware framework that significantly improves linear B-cell epitope prediction by fine-tuning protein language models on distinct taxa of pathogens. Our approach consistently outperformed current state-of-the-art predictors across diverse viral, bacteria and eukaryotic pathogens. By integrating evolutionary information into predictive models, we demonstrate the importance of considering phylogenetic relationships in the analysis of cross-species

[3] WHO fungal priority pathogens list to guide research, development and public health action: https://www.who.int/publications/i/item/9789240060241

biological data. Future directions of improvement include addressing challenges related to data scarcity for specific taxa, such as fungal pathogens and exploring alternative embedding strategies and classifiers. EpitopeTransfer provides a flexible and powerful framework to advance computational approaches in immunological research for the development of pathogen-specific diagnostics and therapeutics.

Authors' contributions

Lindeberg Pessoa Leite: Software, Validation, Methodology, Writing (original draft). **Teófilo Emidio de Campos**: Supervision, Resources, Writing (review & editing); **Francisco Pereira Lobo**: Conceptualization, Writing (Review & Editing); **Felipe Campelo**: Conceptualization, Methodology, Data curation, Visualization, Supervision, Writing (original draft), Writing (Review & Editing). All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Code availability

The source code, data and analysis scripts can be retrieved from https://github.com/fcampelo/epitopetransfer-GPB. A permanent archival version is also available from https://doi.org/10.5281/zenodo.15743750.

References

[1] Janeway C. Immunobiology. Garland Science, New York, 8th edition, 2012.

[2] Selvan SR, Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and methods for T- and B-cell epitope prediction. J Immunol Res 2017; 2680160.

[3] Hamley IW. Peptides for vaccine development. ACS Appl Bio Mater 2022; 5(3):905–44.

[4] Sun R, Qian MG, Zhang X. T and B cell epitope analysis for the immunogenicity evaluation and mitigation of antibody-based therapeutics. mAbs 2024; 16:2324836.

[5] Mucci J, Carmona SJ, Volcovich R, Altcheh J, Bracamonte E, Marco JD et al. Next-generation ELISA diagnostic assay for Chagas disease based on the combination of short peptidic epitopes. PLoS Negl Trop Dis 2017; 11(10):e0005972.

[6] Collatz M, Mock F, Barth E, Holzer M, Sachse K, Marz M. EpiDope: a deep neural network for linear B-cell epitope prediction. Bioinformatics 2020; 37(4):448–55.

[7] Bahai A, Asgari E, Mofrad MRK, Kloetgen A, McHardy AC. EpitopeVec: linear epitope prediction using deep protein sequence embeddings. Bioinformatics 2021; 37(23):4517–25.

[8] Croce G, Bobisse S, Moreno DL, Schmidt J, Guillame P, Harari A et al. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. Nat Commun 2024; 15:3211.

[9] Liu F, Yuan C, Chen H, Yang F. Prediction of linear B-cell epitopes based on protein sequence features and BERT embeddings. Sci Rep 2024; 14:2464.

[10] Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z et al. Simulating 500 million years of evolution with a language model. Science 2025; 387(6736):850–8.

[11] Fenoy E, Edera AA, Stegmayer G. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. Brief Bioinform 2022; 23(4):bbac232

[12] Liu Y, Li D, Zhang X, Xia S, Qu Y, Ling X et al. A protein sequence-based deep transfer learning framework for identifying human proteome-wide deubiquitinase-substrate interactions. Nat Commun 2024; 15:4519.

[13] Schmirler R, Heinzinger M, Rost B. Fine-tuning protein language models boosts predictions across diverse tasks. Nat Commun 2024, 15:7407.

[14] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS 2021; 118:e2016239118.

[15] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang L, Jones L. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2021; 44(10):7112–27.

[16] Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K et al. Single-sequence protein structure prediction using a language model and deep learning. Nat Biotechnol 2022; 40(11):1617–23.

[17] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023; 379(6637):1123–30.

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al. Attention is all you need. Adv Neural Inf Process Syst 2017; 6000–10 (December).

[19] Clifford JN, Høie MH, Deleuran S, Peters B, Nielsen M, Marcatili P. BepiPred-3.0: improved B-cell epitope prediction using protein language models. Protein Sci 2022; 31(12):e4497.

[20] Høie MH, Gade FS, Johansen JM, Wurtzen C, Winther O, Nielsen M et al. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. Front Immunol 2024; 15:1322712.

[21] Ashford J, Reis-Cunha J, Lobo I, Lobo FP, Campelo F. Organism-specific training improves performance of linear B-cell epitope prediction. Bioinformatics 2021; 37:btab536.

[22] Campelo F, Lobo FP. The rise of taxon-specific epitope predictors. Brief Bioinform 2024; 25(2):bbae092.

[23] da Silva B, Ascher DB, Pires DEV. epitope1D: accurate taxonomy-aware B-cell linear epitope prediction. Brief Bioinform 2023; 24(3):bbad114.

[24] Liu R, Hu YF, Du J, Zhang BZ, Yau T et al. Family-specific training improves linear B cell epitope prediction for emerging viruses. IEEE/ACM Trans Comput Biol Bioinform 2023; 20(6):3669–80.

[25] Yin R, Zhu X, Zeng M, Wu P, Li M, Kwoh CK. A framework for predicting variable-length epitopes of human-adapted viruses using machine learning methods. Brief Bioinform 2022; 23(5):bbac281.

[26] Lim LC, Lim YY, Choong YS. Data curation to improve the pattern recognition performance of B-cell epitope prediction by support vector machine. Pure and Applied Chemistry 2021; 93(5):571–77.

[27] Campelo F, de Oliveira ALG, Reis-Cunha J, Fraga VG, Bastos PH, Ashford J et al. Phylogeny-aware linear B-cell epitope predictor detects targets associated with immune response to orthopoxviruses. Brief Bioinform 2024; 25(6):bbae527.

[28] Breiman L. Random forests. Mach Learn 2001; 45:5–32.

[29] Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. 1st ed. Addison-Wesley Longman Publishing Co.; 2005.

[30] Sheskin DJ. Handbook of Parametric and Nonparametric Statistical Procedures. 3rd ed. Chapman and Hall/CRC; 2003.

[31] Chicco D, Jurman G. The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020; 21:1–13.

[32] Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 2019; 20:723.

[33] Kuhn JH, Amarasinghe GK, Basler CF, Bavari S, Bukreyev A, Chandran K et al. ICTV virus taxonomy profile: Filoviridae. J Gen Virol 2019; 100(6):91112

[34] Dmitrijeva M, Tackmann J, Rodrigues JFM, Huerta-Cepas J, Coelho LP, von Mering C. A global survey of prokaryotic genomes reveals the eco-evolutionary pressures driving horizontal gene transfer. Nat Ecol Evol 2024; 8(5):986–98

[35] Csurka G. A Comprehensive Survey on Domain Adaptation for Visual Applications. In: Csurka G (editor). Domain Adaptation in Computer Vision Applications. Paris; Springer Cham; 2017, p. 1–35.

Figure legends

Figure 1 Overview of the EpitopeTransfer framework for building taxon-specific LBCE predictors.

(**A**) Labelled peptides (i.e., peptides known to be either epitope-containing regions or non-epitopes) are retrieved from the IEDB export and their source protein sequences are downloaded. Protein sequences are clustered based on normalised alignment scores, using single-linkage hierarchical clustering with a local similarity threshold of 30%. The resulting clusters are treated as allocation units when splitting data for model development/training and testing. (**B**) The consolidated data is filtered according to the target taxon of interest. All non-target entries associated with a higher taxonomic level (e.g., kingdom or phylum) are used to fine tune the ESM feature embedder, which is then used to extract features for entries belonging to the target taxon. (**C**) Data related to the target taxon (e.g., species or genus) is then used to fit a classifier optimised for predicting epitopes at that lower taxonomic level. The effects of the transfer learning strategy are evaluated by benchmarking the final models on sets of held-out observations and their performance is compared against models trained using the baseline ESM features (without fine-tuning); and using an ESM variant tuned using epitope data from pathogens without any close phylogenetic link; as well as against external benchmarks.

Figure 2 Comparison of EpitopeTransfer against internal and external baselines.

(**A**) EpitopeTransfer provides significant AUC gains when compared to internal baselines. Smaller dots indicate AUC values observed for individual datasets. Larger points with standard error bars represent estimates of mean performance for each combination of method and embedder. p-values correspond to post-ANOVA pairwise comparisons between EpitopeTransfer and the two baselines, after removing the effects of embedder and dataset. (**B**) Comparison of the ESM2 version of EpitopeTransfer against the external baselines. Small black dots indicate performance estimates on individual datasets. The larger square markers represent mean

performances obtained considering the “no leakage” sets (as described in Section “Methods”, subsection “Performance assessment and comparison”) with associated standard errors. For the external baselines, the mean and standard errors calculated using the “EpitopeTransfer holdout” set are also displayed as hollow circles, for reference. (**C**): Distribution of EpitopeTransfer’s per-dataset performance gains, illustrating the gap between observed performance gains (green) and losses (red). A region of practical equivalence comprising differences of ±0.05 in AUC is shown as a grey band. Mean gains and mean losses are displayed as grey squares with associated standard errors. EpitopeTransfer models that presented gains greater than +0.05 simultaneously against all baselines are explicitly labelled, as well as any cases of performance loss outside the region of practical equivalence.

Supplementary material

Supplementary Figure 1: NPTransfer fine-tuning regimen

Set of unrelated groups used to establish the non-phylogenetic baselines. These pairs of datasets are used to perform domain adaptation across unrelated biological groups, to test our hypothesis about the effectiveness of using phylogenetically related data for the embedder fine tuning. If the non-phylogenetic baseline models produced equivalent of favourable results, the enhancements observed in epitope prediction tasks could be exclusively attributed to the effect of fine tuning the representation with LBCE data, which would indicate that considering phylogeny is unnecessary. The observed favourable results for the EpitopeTransfer models instead indicated the presence of a positive effect of the phylogeny-aware data filtering step.

Supplementary Figure 2: Data consolidation step.

In this illustrative example of the data consolidation step, labelled peptides with overlapping coordinates on the protein were aggregated in a residue-by-residue manner. Labels were attributed based on the mode (most common value), with ties being removed. Row “Label summary” summarises the total number of

positive/negative labels on each region of the protein for this mock example, which is used to determine the final labels.

Supplementary File: Supplementary File 1.pdf

This supplementary file contains details about the data retrieval, filtering, and consolidation; The embedder development details; Model training and validation procedures; And a comment on the use of alternative classification approaches.

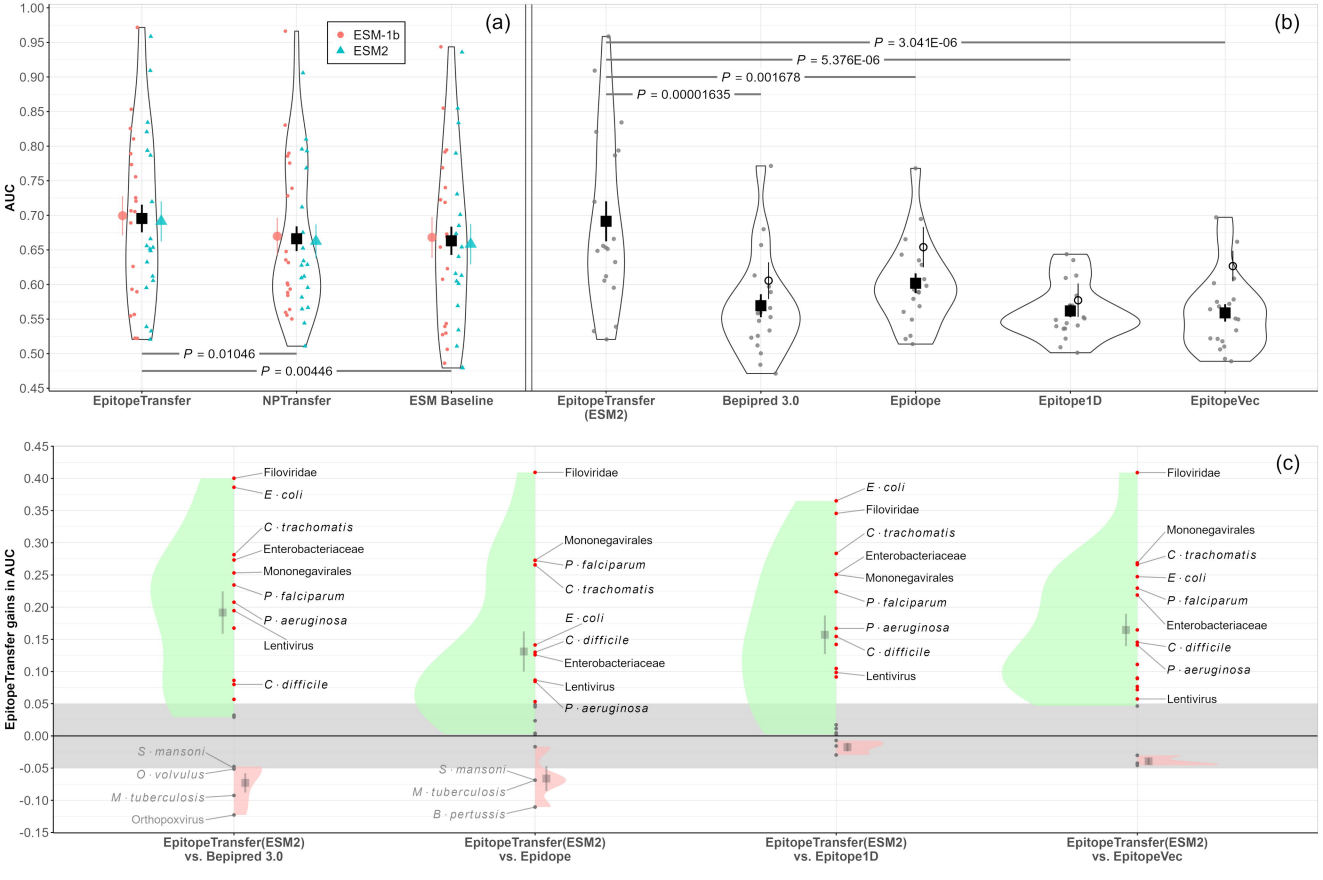
Supplementary Table 1: Paired higher- and lower-level datasets used in this work.

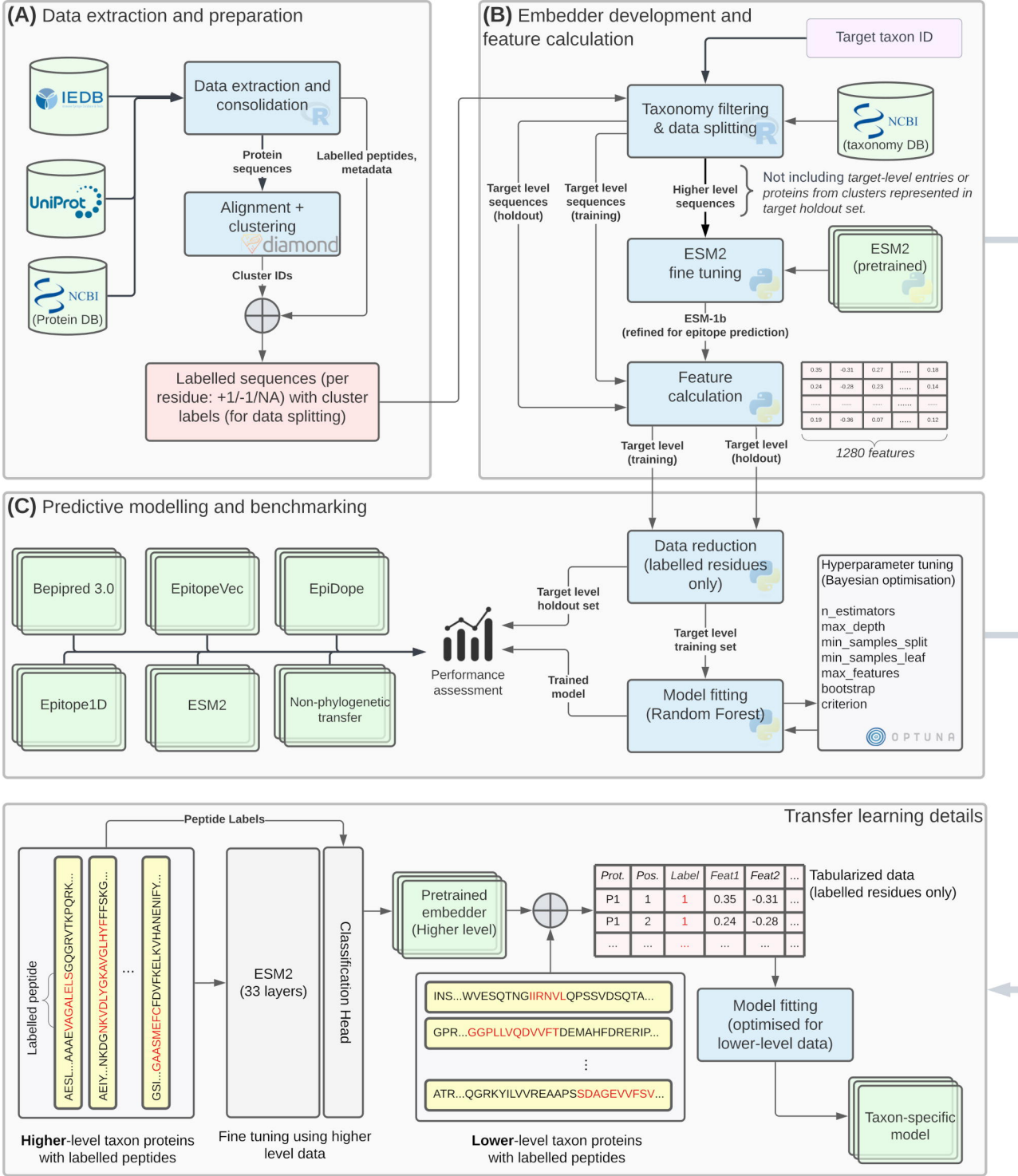
Supplementary Table 2: Data volume in terms of labelled peptides and residues for each dataset used, showing the effect of the filtering and consolidation steps.

Supplementary Table 3: Statistical comparisons of mean values for each performance metric for the internal baselines. The p-values and effect sizes reported refer to the post-hoc tests for pairwise differences of means between the different methods, after removing the effects of Embedder and Dataset. Dunnett contrasts (all-vs-one) were used with EpitopeTransfer as the reference level.

Supplementary Table 4: Statistical comparisons of mean values for each performance metric between EpitopeTransfer (ESM2) and the external baselines. The p-values and effect sizes reported refer to the post-hoc tests for pairwise differences of means between the different methods, after removing the Dataset effects. Dunnett contrasts (all-vs-one) were used with EpitopeTransfer as the reference level.

Supplementary Table 5: Performance details of all methods on all datasets.


(a)
ESM-1b
ESM2
AUC
P = 0.01046
P = 0.00446
EpitopeTransfer
NPTransfer
ESM Baseline
(b)
P = 3.041E-06
P = 5.376E-06
P = 0.001678
P = 0.00001635
EpitopeTransfer (ESM2)
Bepipred 3.0
Epidope
Epitope1D
EpitopeVec
(c)
EpitopeTransfer gains in AUC
Filoviridae
E · coli
C · trachomatis
Enterobacteriaceae
Mononegavirales
P · falciparum
P · aeruginosa
Lentivirus
C · difficile
S · mansoni
O · volvulus
M · tuberculosis
Orthopoxvirus
B · pertussis
EpitopeTransfer(ESM2) vs. Bepipred 3.0
EpitopeTransfer(ESM2) vs. Epidope
EpitopeTransfer(ESM2) vs. Epitope1D
EpitopeTransfer(ESM2) vs. EpitopeVec


(A) Data extraction and preparation
IEDB
UniProt
NCBI
(Protein DB)
Data extraction and consolidation
Protein sequences
Labelled peptides, metadata
Alignment + clustering
diamond
Cluster IDs
Labelled sequences (per residue: +1/-1/NA) with cluster labels (for data splitting)
(B) Embedder development and feature calculation
Target taxon ID
Taxonomy filtering & data splitting
NCBI
(taxonomy DB)
Target level sequences (holdout)
Target level sequences (training)
Higher level sequences
Not including target-level entries or proteins from clusters represented in target holdout set.
ESM2 fine tuning
ESM2 (pretrained)
ESM-1b (refined for epitope prediction)
Feature calculation
1280 features
Target level (training)
Target level (holdout)
(C) Predictive modelling and benchmarking
Bepipred 3.0
EpitopeVec
EpiDope
Epitope1D
ESM2
Non-phylogenetic transfer
Performance assessment
Target level holdout set
Trained model
Data reduction (labelled residues only)
Target level training set
Model fitting (Random Forest)
Hyperparameter tuning (Bayesian optimisation)
n_estimators
max_depth
min_samples_split
min_samples_leaf
max_features
bootstrap
criterion
OPTUNA
Transfer learning details
Peptide Labels
Labelled peptide
AESL...AAAEVAGALELSGQGRVTKPQRK...
AEIY...NKDGNKVDLYGKAVGLHYFFFSKG...
GSI...GAASMEFCFDVFKELKVHANENIFY...
ESM2 (33 layers)
Classification Head
Higher-level taxon proteins with labelled peptides
Fine tuning using higher level data
Pretrained embedder (Higher level)
INS...WVESQTNGIIRNVLQPSSVDSQTA...
GPR...GGPLLVQDVVFTDEMAHFDRERIP...
ATR...QGRKYILVVREAAPSSDAGEVVFSV...
Lower-level taxon proteins with labelled peptides
Prot. Pos. Label Feat1 Feat2
P1 1 1 0.35 -0.31
P1 2 1 0.24 -0.28
Tabularized data (labelled residues only)
Model fitting (optimised for lower-level data)
Taxon-specific model