EPITOPETRANSFER: A PHYLOGENY-AWARE TRANSFER LEARNING FRAMEWORK FOR TAXON-SPECIFIC LINEAR B-CELL EPITOPE PREDICTION

A PREPRINT

Lindeberg P. Leite
Department of Computer Science
University of Brasília, Brazil
and
Directorate of IT and Innovation
Federal Police, Brazil

Teófilo E. de Campos
Applied Sciences Group
Microsoft, UK
and
Department of Computer Science
University of Brasília, Brazil

Francisco P. Lobo
NCI-Frederick
National Cancer Institute, USA
and
Institute of Biological Sciences
UFMG, Brazil

Felipe Campelo*
Intelligent Systems Laboratory
University of Bristol
Bristol BS8 1TW, UK
f.campelo@bristol.ac.uk

First submitted for peer review on 5 October 2024

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 for training predictive models, which can result in biases towards widely studied organisms and compromise performance when focusing on neglected or emerging pathogens. We introduce 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, 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.

1 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 molucule 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 analyzing protein data [6, 7, 8, 9, 10], with transfer learning emerging as a notable approach [11, 12, 13]. Several methods involve using latent space vector representations of amino acid residues that are extracted from large, pre-trained protein language models. These representations have the ability to encode biological properties of proteins in a context-dependent manner [14, 15, 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 [14, 17, 10] 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 modeling, 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 labeled 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, 22, 23, 24, 25, 26], e.g., for the development of immunodiagnostic tests[27] or vaccines.

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

2 Materials and Methods

We developed a phylogeny-aware modelling framework 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 twenty target taxa of interest including viruses, bacteria and eukaryotic pathogens (Supplementary Tables 1 – 2).

2.1 Data extraction and preparation

Data extraction, filtering, and consolidation were performed using the *epitopes* R package [28], according to the procedures outlined below.

Epitope data was retrieved from the complete XML export of the Immune Epitopes Database, IEDB [29]. All entries classified as LBCEs from organisms within the taxa Viruses (NCBI:txid10239), Bacteria (NCBI:txid2), and Eukaryota (NCBI:txi2759) were extracted from the IEDB export. The proteins associated with each entry were retrieved from either the NCBI protein database [30] or UniprotKB [31], based on the protein IDs provided in the metadata of each IEDB record. Overlapping peptides of the same class were merged into a single entry to prevent partial data duplication, while residues with conflicting labels were labelled as the mode of the labels for each residue, and removed in case of ties. Positive-labelled peptides of length greater than 30 residues were removed to prevent long "Epitope-containing regions" from adding excessive noise to the training data, and labelled peptides shorter than 5 residues were also treated as noise and removed from the datasets.

Entries were clustered based on source protein dissimilarity, calculated using DIAMOND [32], using agglomerative clustering with single linkage and a 30% similarity threshold. Clusters were considered as the basic splitting unit when isolating the final test sets (referred to as "EpitopeTransfer holdout" in the text), as well as for cross validation and hyperparameter tuning, to minimize data leakage due to similarity/homology.

To investigate the transfer learning approach proposed in this work, we instantiated twenty pairs of datasets (higher level and lower level) for a diverse range of pathogens, including bacterial, viral and eukaryotic pathogens. Supplementary Table 1 provides the details of the datasets used, including the number of positively- and negatively-labelled peptides at each level, and the taxonomic levels used for each case. Supplementary Table 2 provides the full list of NCBI taxonomy IDs encompassed by each dataset.

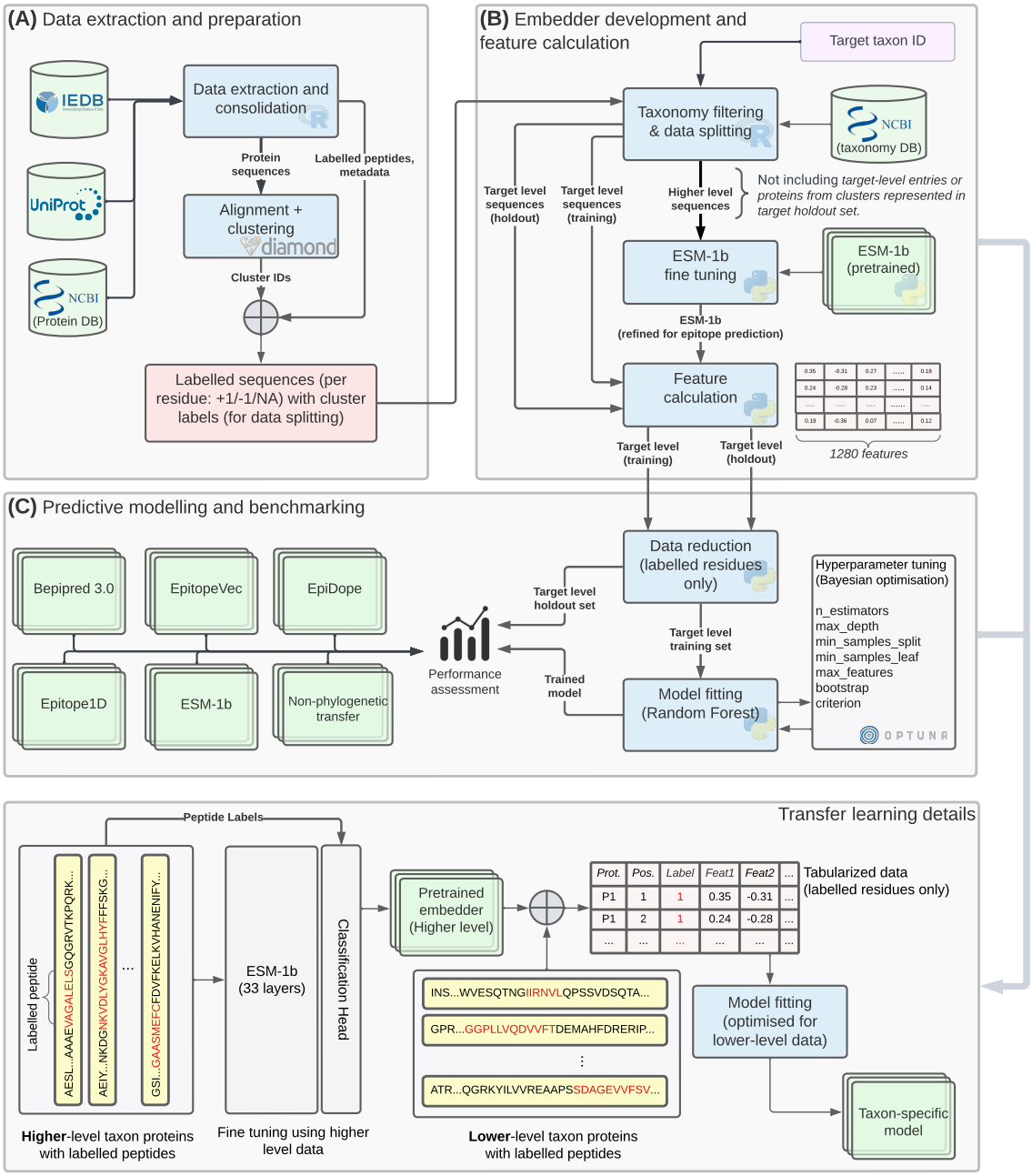

Figure 1: Overview of the EpitopeTransfer framework for building taxon-specific LBCE predictors. (A) Labelled peptides (i.e., peptides known to be either epitope-containing regions or non-epitopes) are retrieved from the IEDB export and their source proteins are retrieved either from the NCBI Protein database or UniprotKB, based on the protein ID provided in the IEDB entry. Protein sequences are clustered based on normalised alignment scores, using single-linkage hierarchical clustering with a 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-1b 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 comparing their performance against models trained (i) using the baseline ESM-1b features (without fine tuning on epitope data) and (ii) using an ESM-1b variant tuned using epitope data from pathogens outside the target taxon's superkingdom, i.e., without a direct phylogenetic link; as well as against external benchmarks.

2.2 Modelling

We developed a framework for instantiating taxon-specific predictors, consisting of three main components (see Figure 1). Given a pair of datasets for matched higher level – lower level taxa, the modelling tasks were as follows:

1. *Embedder development*: fine tuning of the general-purpose ESM protein language model to the task of LBCE prediction, as detailed in the Supplementary Materials, section "Embedder Development Details". The ESM model 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 model 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 650 million parameter version of ESM-1b as our base model, but adapting the framework to use later or larger embedder versions is a straightforward task.
2. *Feature calculation*: the fine-tuned embedder is deployed to extract features 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 model to capture richer 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 labeled peptide regions are selected for later classifier training and performance assessment.
3. *Predictive model training*: in the final step, the data generated in the previous step are used to fit and optimize the hyperparameters a Random Forest classifier [33], resulting in a bespoke LBCE prediction model for the target taxon.

2.3 Performance assessment and comparison

To assess the predictive performance of the models and ensure comparability with existing literature, we focus mainly 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 the Supplementary Materials, 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 (PPV) and Negative (NPV) predictive values, Sensitivity and Specificity.

We created two internal baselines to investigate of the effect of the proposed approach on predictive performance. The first used the general purpose 650 million parameter ESM-1b, without any fine tuning, as the feature calculation engine. Models based on raw ESM-1b 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 were developed 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., without a direct phylogenetic link (see Supplementary Figure 1). These *non-phylogenetic transfer* (NPTransfer) baselines were created to allow the estimation of the effect of using phylogeny-aware filter when selecting data for the fine-tuning of the ESM-1b embedder.

In addition to the internal baselines listed above, four state-of-the-art generalist LBCE predictors were used as external comparison baselines: BepiPred 3.0 [19], EpiDope [6] and EpitopeVec [7], which are deep learning-based 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, and 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 the independently extracted final test set for each target taxon (which is referred to in the text as "EpitopeTransfer holdout"). Section "Training and Validation Procedures for EpitopeTransfer" of the Supplementary Materials provides details on the training and validation procedures 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 have assessed these tools primarily using the observations of each target taxon that are not found in their published training datasets. This approach is more likely to provide unbiased estimates of their

generalization performance for each target taxa in Supplementary Tables 1 – 2. Additionally, we also calculated the performance of the external baseline methods on the EpitopeTransfer holdout set, and reported these in the results section as well as in Supplementary Tables 3 – 5.

We compared the performance estimates of EpitopeTransfer with the internal and external baselines using Wilcoxon Signed-Ranks tests [35] to assess the statistical significance of the median of the paired differences of performance, as measured by each different indicator. To correct for the simultaneous test of multiple hypotheses (EpitopeTransfer vs. each of the five baselines), the p-values returned by each comparison were adjusted using the Bonferroni-Hochberg correction [36], which we used to control the false discovery rate (FDR) of each family of hypotheses at the adjusted significance threshold of 0.05. Unless otherwise stated, all p-values reported in this text correspond to FDR-corrected values.

3 Results

3.1 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 mainly by contrasting the performance of EpitopeTransfer against baseline models developed following an identical procedure, removing only the phylogeny-aware fine tuning of the feature embedder (*ESM-1b baseline*).

We observed significant performance gains of EpitopeTransfer in relation to the ESM-1b baseline across all performance measures (see Supplementary Tables 3–5). In particular, the Area Under the ROC Curve (AUC) values obtained by the two approaches (Figure 2, top) indicate statistically significant gains of EpitopeTransfer in comparison to the ESM-1b baseline (Δmedians $= 0.043$; median of paired differences $= 0.031$; Wilcoxon Signed Rank $p_{adj} = 0.005$). We also observed particularly large effect sizes in terms of the median of paired differences in terms of MCC ($+0.089$) and Specificity ($+0.082$), indicating that the fine tuning step induced practically relevant performance gains, particularly in terms of the correct detection of negative examples.

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 non-Phylogenetic transfer learning baselines (NPTransfer). These baselines consist of models created using the same procedure as EpitopeTransfer, with the exception that the embedder fine tuning was performed using epitope data from higher level taxa that not only do not contain the target taxon, but also consist only of data from very distantly related taxa (see Methods section, "Performance assessment and comparison" and Supplementary Figure 1).

The results obtained very clearly indicate a gain in performance resulting from the incorporation of a phylogeny-aware data filtering strategy for the fine tuning step across almost every performance criteria, with the exception of a minor numerical decrease in Sensitivity (-0.029, not statistically significant; see Supplementary Tables 3–5 for details). In terms of AUC, we observed a statistically significant increase in performance (Δmedians $= 0.079$; median of paired differences $= 0.038$; Wilcoxon Signed Rank $p_{adj} = 0.033$), which can be seen in the top panel of Figure 2. We noticed particularly large gains in terms of Accuracy ($+0.108$), MCC ($+0.082$), NPV ($+0.073$) and Specificity ($+0.179$). Taken in combination with the observed improvement over the ESM-1b baseline, these results suggest that the superior performance of EpitopeTransfer is at least in part due to the fine tuning of the feature embedder specifically on 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 without taking phylogeny into consideration.

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

We compared the results obtained by EpitopeTransfer models against four state-of-the-art methods for LBCE prediction. Three of those are deep-learning based, generalist LCBE predictors: Bepipred 3.0 [19], which also employs an ESM protein embedder as part of its pipeline, augmented by surface accessibility and protein length features; Epidope [6], which utilizes a deep neural network sequence embedder known as ElMo [37]; and EpitopeVec [7], which uses a combination of amino acid features, modified antigenicity scales and neural network-based protein embeddings. The fourth external comparison baseline, epitope1D [23], implements a taxon-specific modelling approach, with models provided for major phyla or kingdoms of bacterial, viral, and eukaryotic pathogens, 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 the top panel of Figure 2, we observed statistically significant differences in median performance between EpitopeTransfer and the four external baselines, with effect sizes noticeably greater than the ones observed in the comparisons against the ESM-1b and NPTransfer baselines (Supplementary Tables 3 – 5 provide full numerical details for AUC as well as the other performance indicators).

Of the three generalist baselines, Epidope presented the best median AUC across the 20 datasets used in this study. Contrasting the observed performance of EpitopeTransfer (median AUC = 0.707) against Epidope, we see a considerable improvement in median predictive performance (Δmedians $= 0.113$; median of paired differences $= 0.107$; Wilcoxon Signed Rank $p_{adj} = 0.002$). Similarly or even larger performance gains are observed in comparisons against all generalist predictors across all performance indicators except Specificity, for which EpitopeTransfer was observed to result in a statistically significant performance decrease in relation to Epidope, and a non statistically significant decrease in comparison to Bepipred 3.0.

Examining the comparison against epitope1D, which also uses taxon-specific models (albeit with a coarser granularity that EpitopeTransfer) we observe similarly large performance differences across most indicators. In terms of AUC, we have Δmedians $= 0.133$, the median of paired differences $= 0.143$, and the Wilcoxon Signed Rank $p_{adj} = 0.0005$. Similar or larger performance increases are also seen across all performance indicators, again with the exception of Specificity.

The bottom panel of Figure 2 shows the distribution of the magnitude of paired differences of AUC for each dataset, stratified in terms of performance gains and performance losses. We observed significantly more instances of performance gains in comparison with cases of performance degradation in all comparisons (Binomial test p-values of $0.041,\ 0.012,\ 0.003$ and 0.012 for the comparisons against Bepipred 3.0, Epidope, EpitopeVec and Epitope1D, respectively). Moreover, we obtained substantially larger average gains in AUC than the observed performance decreases. Thirteen EpitopeTransfer models exhibited positive AUC gains in all comparisons, with twelve of the datasets had AUC gains greater than our (arbitrarily defined) minimally relevant effect size of $+0.05$ in all cases. Gains larger than 0.1 on all comparisons, which we considered as substantial, were achieved in eight cases, which are labelled in Figure 2. Importantly, these include models across the three domains explored: viral, bacterial and eukaryotic pathogens. Four of these models exhibiting high AUC gains correspond to specific pathogenic species (*P. falciparum*, *C. trachomatis*, Influenza A and *C. difficile*, and the other four were derived for different taxonomic levels (family Filoviridae and its containing order Mononegavirales; genus Lentivirus; and bacterial family Enterobacteriaceae).

3.3 EpitopeTransfer provides high-performance LBCE predictors for some 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 [38], is the one where observed performance differences are most prominent. As shown in Figure 2, 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 absolute performance values for this model, which displays an out-of-sample AUC of 0.97, a PPV of 0.66 and an NPV of 0.99, suggesting it as a robust LBCE predictor for viruses under this family.

Other EpitopeTransfer models where particularly high performance gains were observed include the one derived for *Plasmodium falciparum*, with AUC gains ranging between 0.24 and 0.28 over the current state-of-the-art (EpitopeTransfer out-of-sample AUC = 0.81, PPV = 0.79 and NPV = 0.73); for the genus Lentivirus, which contains the human immunodefficiency virus (HIV), for which AUC gains in the range $0.18 - -0.32$ were observed (EpitopeTransfer out-of-sample AUC = 0.79, PPV = 0.86 and NPV = 1.0); and the Enterobacteriaceae family, which includes medically relevant bacteria such as *E. coli*, *Salmonella* and *Shigella* [39], with AUC gains in the $0.13 - -0.28$ range (EpitopeTransfer out-of-sample AUC = 0.83, PPV = 0.69 and NPV = 0.79).

These models and, to a lesser degree, the others nominally listed in Figure 2, are likely to provide LBCE predictions with considerably better results for their target taxa than the current state-of-the-art. On the other extreme, we observed a few pathogens / taxa for which we would not recommend the EpitopeTransfer models derived in this work. Both the models for SARS-CoV-2 and *M. tuberculosis* performed worse than the external baselines. In both cases, the best model was Bepipred 3.0, against which the two corresponding EpitopeTransfer models presented AUC losses of 0.13 and 0.09, respectively. The remaining few cases of performance degradation of EpitopeTransfer in relation to the current state of the art were either specific to one predictor, or of magnitude lower than 0.05, which we defined as our range of practical equivalence.

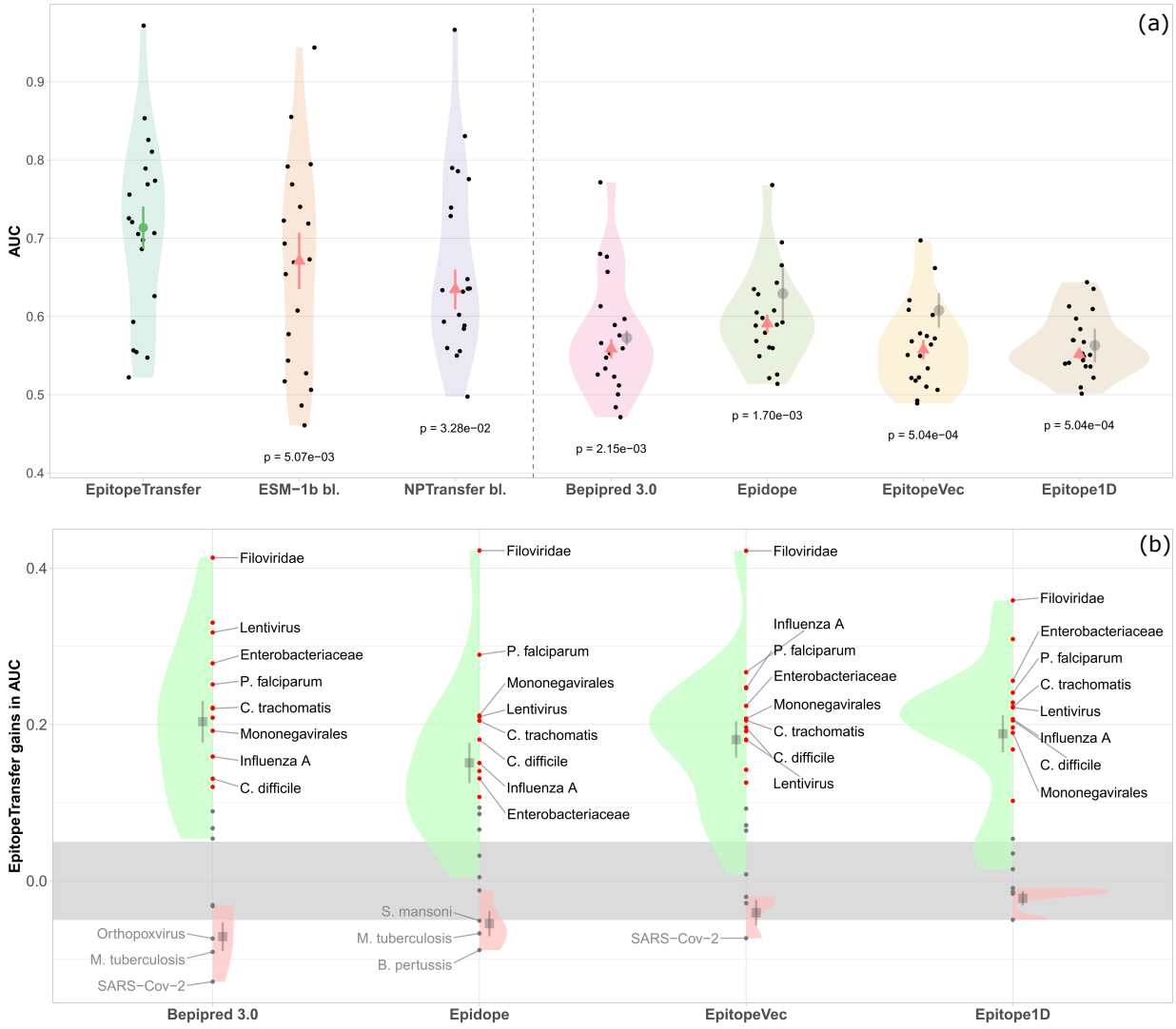

Figure 2: **(a)**: Performance summary for EpitopeTransfer and comparison baselines on the datasets listed in Supplementary Table 1. Black dots indicate performance estimates on individual datasets. Larger, colored marks represent median performances obtained considering the "no leakage" holdout sets, as described in Section *Performance assessment and comparison*), with associated standard errors (note that standard error bars may be smaller than the plot symbol). Upward-facing triangles indicate a statistically significant gain of performance of EpitopeTransfer in relation to a given baseline (Wilcoxon signed-ranks test, adjusted p-value < 0.05). For the external baselines, the median and standard errors calculated using the "EpitopeTransfer holdout" set are also displayed as light gray circles (no inferential results are displayed for these cases; however, they are reported in Supplementary Tables 3 – 5). **(b)**: Distribution of EpitopeTransfer gains in terms of paired differences of AUC, illustrating the gap between observed performance gains (green) and losses (red). We defined a region of practical equivalence comprising differences of ± 0.05 AUC, shown as a gray band. Mean gains and mean losses are displayed as gray squares with associated standard errors. Models for which EpitopeTransfer displayed large performance gains (defined as greater than $+0.1$ AUC) are marked as red dots. Models presenting gains greater than $+0.1$ against every baseline are explicitly labelled, as well as all cases of performance loss outside the region of practical equivalence.

4 Discussion and Conclusions

We have developed EpitopeTransfer, a modeling framework to instantiate taxon-specific predictors of LBCEs using a phylogeny-aware fine tuning of the ESM-1b 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, by contrasting the performance of EpitopeTransfer with similar models developed (i) without the fine tuning step ("ESM-1b baseline") and (ii) 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, at least partially, due to the phylogeny-aware data selection for the refinement of the ESM protein language model.

We also established that EpitopeTransfer models provide expected performance gains between 19 and 25% in AUC over both generalist LBCE predictors [19, 6, 7] and a recent taxon-specific predictor set [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 lineages. By leveraging evolutionary relationships, it is possible to create optimized 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.

The EpitopeTransfer framework presented here is completely modular, making it easily adaptable for 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., by 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 classifiers other than Random Forests, such as gradient boosting or feedforward neural networks (FNN), to build the predictors. Although preliminary testing 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.

The models developed using EpitopeTransfer are adapted from broader to more specific domains (i) using partially labelled protein data from the source domain (i.e., full protein sequences from the higher level taxon containing some labelled peptides) to fine-tune the protein language model, and (ii) then the proportionally few labelled samples from the target domain are used to help train a 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 [40]. In principle, this could be done even in the absence of any labelled data in the target domain, which would make EpitopeTransfer extensible to extremely data scarce but nonetheless critical cases such as fungal pathogens.[1] However, the performance of the framework for such cases still remains to be investigated.

5 Data availability

The source code, data and analysis scripts can be retrieved from `https://github.com/fcampelo/epitopetransfer`. An archival version is also available from FigShare at `https://doi.org/10.6084/m9.figshare.27175953`.

6 Competing interests

The authors declare no competing interests.

7 Author contributions

L.P.L.: Software, Validation, Methodology, Writing – original draft; **T.E.C.**: Supervision, Resources, Writing – Review & Editing; **F.P.L.**: Conceptualization, Writing – Review & Editing; **F.C.**: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Supervision, Writing – original draft.

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

References

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

[2] Senthamil R. Selvan, Jose L. Sanchez-Trincado, Marta Gomez-Perosanz, and Pedro A. Reche. Fundamentals and methods for T- and B-cell epitope prediction. *Journal of Immunology Research*, 2017:2680160, 2017.

[3] Ian W. Hamley. Peptides for vaccine development. *ACS Applied Bio Materials*, 5(3):905–944, February 2022.

[4] Ruoxuan Sun, Mark G. Qian, and Xiaobin Zhang. T and B cell epitope analysis for the immunogenicity evaluation and mitigation of antibody-based therapeutics. *mAbs*, 16(1), March 2024.

[5] Juan Mucci, Santiago J. Carmona, Romina Volcovich, Jaime Altcheh, et al. Next-generation ELISA diagnostic assay for Chagas disease based on the combination of short peptidic epitopes. *PLOS Neglected Tropical Diseases*, 11(10):e0005972, October 2017.

[6] Maximilian Collatz, Florian Mock, Emanuel Barth, Martin Hölzer, Konrad Sachse, and Manja Marz. EpiDope: a deep neural network for linear B-cell epitope prediction. *Bioinformatics*, 37(4):448–455, 09 2020.

[7] Akash Bahai, Ehsaneddin Asgari, Mohammad R K Mofrad, Andreas Kloetgen, and Alice C McHardy. EpitopeVec: linear epitope prediction using deep protein sequence embeddings. *Bioinformatics*, 37(23):4517–4525, 06 2021.

[8] Giancarlo Croce, Sara Bobisse, Dana Léa Moreno, Julien Schmidt, et al. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nature Communications*, 15(1):3211, apr 2024.

[9] Fang Liu, ChengCheng Yuan, Haoqiang Chen, and Fei Yang. Prediction of linear B-cell epitopes based on protein sequence features and bert embeddings. *Scientific Reports*, 14(1):2464, jan 2024.

[10] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, et al. Simulating 500 million years of evolution with a language model, July 2024. bioRxiv preprint. https://doi.org/10.1101/2024.07.01.600583.

[11] Emilio Fenoy, Alejando A Edera, and Georgina Stegmayer. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. *Briefings in Bioinformatics*, 23(4):bbac232, 06 2022.

[12] Yuan Liu, Dianke Li, Xin Zhang, Simin Xia, et al. A protein sequence-based deep transfer learning framework for identifying human proteome-wide deubiquitinase-substrate interactions. *Nature Communications*, 15(1):4519, may 2024.

[13] Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15:7407, August 2024.

[14] Alexander Rives, Joshua Meier, Tom Sercu, Zeming Lin, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118:e2016239118, 04 2021.

[15] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, et al. ProtTrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 07 2021.

[16] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.

[17] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, et al. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[19] Joakim Nøddeskov Clifford, Magnus Haraldson Høie, Sebastian Deleuran, Bjoern Peters, et al. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Science*, 31(12):e4497, 2022.

[20] Magnus Haraldson Høie, Frederik Steensgaard Gade, Julie Maria Johansen, Charlotte Würtzen, et al. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. *Frontiers in Immunology*, 15, February 2024.

[21] Jodie Ashford, João Cunha, Igor Lobo, Francisco Lobo, and Felipe Campelo. Organism-specific training improves performance of linear B-cell epitope prediction. *Bioinformatics (Oxford, England)*, 37:btab536, 07 2021.

[22] Felipe Campelo and Francisco P Lobo. The rise of taxon-specific epitope predictors. *Briefings in Bioinformatics*, 25(2):bbae092, 03 2024.

[23] Bruna da Silva, David B. Ascher, and Douglas E.V. Pires. epitope1D: accurate taxonomy-aware B-cell linear epitope prediction. *Briefings in Bioinformatics*, 24(3), 2023.

[24] Ran Liu, Ye-Fan Hu, Jin Du, et al. Family-specific training improves linear B cell epitope prediction for emerging viruses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(6):3669–3680, November 2023.

[25] Rui Yin, Xianghe Zhu, Min Zeng, et al. A framework for predicting variable-length epitopes of human-adapted viruses using machine learning methods. *Briefings in Bioinformatics*, 23(5), 2022.

[26] Li Cen Lim, Yee Ying Lim, and Yee Siew Choong. Data curation to improve the pattern recognition performance of B-cell epitope prediction by support vector machine. *Pure and Applied Chemistry*, 93(5):571–577, May 2021.

[27] Felipe Campelo, Ana L.G. de Oliveira, Jo ao Reis-Cunha, Vanessa G. Fraga, et al. Phylogeny-aware linear b-cell epitope predictor detects targets associated with immune response to orthopoxviruses. *Briefings in Bioinformatics*, 25(6):bbae527, 2024.

[28] Felipe Campelo, Jodie Ashford, and Francisco P. Lobo. epitopes: Processing, feature extraction and modeling of epitope data from the Immune Epitope Database (IEDB), 2022. GitHub: https://github.com/fcampelo/epitopes.

[29] Randi Vita, Swapnil Mahajan, James A. Overton, Sandeep K. Dhanda, et al. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 2019.

[30] NCBI. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(D1):D7–D19, 2015.

[31] Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, et al. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, November 2022.

[32] Benjamin Buchfink, Klaus Reuter, and Hajk Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18:366–368, 04 2021.

[33] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.

[34] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., USA, first edition, 2005.

[35] David J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC, third edition, 2003.

[36] J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, January 1995.

[37] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1), December 2019.

[38] Jens H. Kuhn, Gaya K. Amarasinghe, Christopher F. Basler, Sina Bavari, et al. ICTV virus taxonomy profile: Filoviridae. *Journal of General Virology*, 100(6):911–912, June 2019.

[39] Seema Alnajar and Radhey S. Gupta. Phylogenomics and comparative genomic studies delineate six main clades within the family Enterobacteriaceae and support the reclassification of several polyphyletic members of the family. *Infection, Genetics and Evolution*, 54:108–127, October 2017.

[40] Gabriela Csurka. *A Comprehensive Survey on Domain Adaptation for Visual Applications*, pages 1–35. Springer International Publishing, Cham, 2017.