PIPENN-EMB: ensemble net and protein embeddings generalise protein interface prediction beyond homology

David P. G. Thomas[1,+], **Carlos M. Garcia Fernandez**[1,+], **Reza Haydarlou**[1,*], **and K. Anton Feenstra**[1,*]

[1]Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, 1081HV, The Netherlands
[*]corresponding authors: k.a.feenstra@vu.nl, r.haydarlou@vu.nl
[+]these authors contributed equally to this work

ABSTRACT

Protein interactions are crucial for understanding biological functions and disease mechanisms, but predicting these remains a complex task in computational biology. Increasingly, Deep Learning models are having success in interface prediction. This study presents PIPENN-EMB which explores the added value of using embeddings from the `ProtT5-XL` protein language model. Our results show substantial improvement over the previously published PIPENN model for protein interaction interface prediction, reaching an MCC of 0.313 vs. 0.249, and AUC-ROC 0.800 vs. 0.755 on the `BIO_DL_TE` test set. We furthermore show that these embeddings cover a broad range of 'hand-crafted' protein features in ablation studies. PIPENN-EMB reaches state-of-the-art performance on the `ZK448` dataset for protein-protein interface prediction. We showcase predictions on 25 resistance-related proteins from *Mycobacterium tuberculosis*. Furthermore, whereas other state-of-the-art sequence-based methods perform worse for proteins that have little recognisable homology in their training data, PIPENN-EMB generalises to remote homologs, yielding stable AUC-ROC across all three test sets with less than 30% sequence identity to the training dataset, and even to proteins with less than 15% sequence identity.
Availability: Webserver, source code and datasets at `www.ibi.vu.nl/programs/pipennemb/`

Introduction

Protein interactions are fundamental to cellular processes[1]. These interactions drive signal transduction, enzymatic activity, structural support, and the regulation of gene expression[2]. Disruptions in this mechanism can lead to various diseases[3], making them a key focus for therapeutic interventions[4–6]. Tackling protein interactions can help understand biological functions and disease mechanisms. Despite having annotations for protein interactions in a pair-based approach, i.e. knowing if two proteins interact, most of these known interactions lack detailed structural information and thereby making them therapeutically non-viable targets[2,7]. Understanding the binding site whereby proteins interact can shed light on how protein interactions come about; residues forming the binding site are called interface residues. Ongoing research is moving forward to understanding which residues belong to the interface by turning the problem into a prediction task[8–13].

Recent advances in machine learning approaches demonstrate strong performance in predicting different types of protein annotations[10,12]. Algorithms addressing protein interface prediction use two main approaches: structure-based and sequence-based. Structure-based methods generally outperform sequence-based ones. However, despite AlphaFold2[14], reliable structural information is still not available for many important types of proteins and protein regions[15,16]. Moreover, the usefulness of predicted structures as input for prediction of functional properties such as interface regions or binding sites may be still quite limited[17–19]. On the other hand, sequence-based approaches, while also dependent on structure-based annotations for training, can exploit the benefits of pre-training and transfer learning from the abundantly available sequence data[20–25]. Transfer learning techniques, such as using embeddings, leverage a self-supervised learning approach to learn powerful representation of proteins at the amino acid level[20,25]. Previous work demonstrated the superior capability of embeddings when compared to 'hand-crafted' features for predicting protein properties[26–28]. Sequence-based models typically employ architectures developed for modeling time series or text data, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) or Transformers[29,30]. Moreover, ensemble networks are an interesting option to maximize performance[12]. These ensembles exhibit greater generalisability compared to the regular architectures, albeit at the expense of reduced interpretability[31].

In this study, we explore two distinct enhancements to the PIPENN model[12], our previously developed sequence-based ensemble predictor for protein interfaces. First, we retrained PIPENN using embeddings from `ProtT5-XL`[20], using the

BioDL_P_TR training set. Subsequently, we updated the ensemble neural network, which we will refer to as PIPENN-EMB, and evaluate its performance on the PIPENN BioDL_P_TE testset, and the external ZK448 benchmark dataset. To evaluate PIPENN-EMB's potential to predict protein interaction interface residues on a new unseen dataset, we applied it to *Mycobacterium tuberculosis* (MTB). MTB infection causes tuberculosis (TB), which continues to be a major cause of death worldwide; in 2022 approximately 1.3 million deaths according to the World Health Organisation[32]. Each year, around 300,000 people contract drug-resistant MTB. Our exploration centers on PIPENN-EMB's potential to predict interface residues in MTB drug resistance-related proteins[33,34]. Such an analysis may contribute to our understanding of how protein interfaces may drive TB drug resistance.

Methods

Data Collection and Preprocessing

We used the PIPENN training data set (BioDL_P_TR) and its fully independent test data set (BioDL_P_TE), which were newly constructed by Stringer et al. (2022)[12], and the ZK448 benchmark data set from Zhang & Kurgan (2018)[10]. Both contain residues annotated for protein-protein interactions. In summary, BioDL_P was generated using annotations from protein structures from PDB, following specific annotation criteria and sequence mapping protocols. Post-processing steps, including sequence clustering at 25% and filtering on sequence length between 30–700, were applied to form non-redundant training and testing sets. For evaluating the performance of our trained models on various independent test sets, we used the *Equal* method[10], by which a cutoff point is selected where number of actual positives is equal to the number of predicted positives.

Model Architectures

The PIPENN architectures are composed of multiple building blocks, each containing various hyperparameters. Different compositions enable learning architectures to provide great flexibility and a broad application area. The following gives a high-level overview of the PIPENN architectures, further detail may be found in Stringer et al. (2022)[12].

PIPENN is based on an ensemble of neural networks, which is referred to as *ensnet*. It is a neural network that uses the protein interface predictions of the following six neural architectures as input to determine whether an amino acid is part of an interface: *ann* is a fully connected architecture; *dnet* is a dilated CNN; *rnet* is a residual CNN; *rnn* consists of two layers of Gated Recurrent Unit (GRU) cells; and *cnet* is a hybrid model that combines the strengths of *rnet* and *rnn* to enhance performance. For training of PIPENN-EMB with embeddings features, we utilised the same hyperparameters as established for the PIPENN models.

Protein Language Model Embeddings

Embeddings are a dense vector representation, capturing complex relationships and characteristics of amino acids that are not readily apparent in their raw form. Protein Language Models (PLMs) extend the concepts of Natural Language Processing (NLP) language models to the realm of protein sequences: amino acids are treated as tokens (cf. words in NLP), and entire protein sequences are regarded as sentences[20,35].

PLMs are initially trained in a self-supervised manner, where the model learns to predict masked amino acids in known protein sequences. This training uses large datasets of protein sequences without any annotations, treating the sequences purely as text data. Here, we used the ProtT5-XL PLM, which was created by Elnaggar et al. (2019)[20] using the UniRef50 database[36]. UniRef50 is approximately eight times larger than the largest datasets previously utilized for PLMs, resulting in a 5-fold increase in the number of tokens[37]. We integrated embeddings from ProtT5-XL in all our datasets and then measured the performance of the new PIPENN-EMB models in comparison with the previous PIPENN.

Feature Ablation

The ablation study conducted at the feature level involved dividing the features into four categories: (1) Structural Information (SI), including secondary structure and surface accessibility; (2) Position Specific Scoring Matrix (PSSM), representing evolutionary conservation information for each sequence, generated using PSI-BLAST; (3) Protein Length (PL), representing the number of amino acids in a protein; (4) Embeddings (EMB), from ProtT5-XL as described above. In order to assess the impact of embeddings on the performance, we retrained the models using different combinations of the four feature categories. The training was performed on the PIPENN BioDL_P_TR data set and tested on the test data set BioDL_P_TE.

Mycobacterium tuberculosis use-case

As an additional external test-set, we selected proteins with known antimicrobial resistance phenotypes in *Mycobacterium tuberculosis* (MTB) from a GWAS dataset acquired from TB-profiler[38]. Interface annotations were retrieved from the PDB-KB API[39]. We excluded proteins longer than 1000 residues, due to length constraints imposed by the prediction methods. Additionally, proteins lacking protein annotation from the PDB-KB API were also omitted from the analysis. The selected 25 proteins constitute the MTB test set.

Comparison to state-of-the-art

We also used the MTB test set to compare the performance of our sequence-based interface prediction model (PIPENN-EMB) with other top-performing sequence-based and parameter-free structure-based prediction methods. Specifically, we evaluated against Seq-InSite[40], CSM-potential2[41], and PeSTo[42].

Homology detection

It is well known that homology is a strong source of predictions for protein properties, however, in the context of method comparison it leads to data leakage between training and test performance[43]. To investigate this beyond the usual 25% sequence identity filtering that is now regularly done[44], we compare observed performance per protein with its similarity to the respective method's training data. For the detection of homology between training data and protein sequences of interest, we ran BLASTP locally (NCBI-BLAST-2.15.0+)[45,46]. Selection of hits was based on the lowest E-value.

We obtained the training data set for CSM-potential2 from their webserver https://biosig.lab.uq.edu.au/csm_potential/data[41]. The PeSTo training data was obtained from their repository https://github.com/LBM-EPFL/PeSTo/blob/main/data/datasets/subunits_train_set.txt[42]. For Seq-InSite the training set trainDset_without500+_Pid_Pseq_label_V10Aug2021.txt was kindly provided by the authors[40]. To ensure that local alignments reflect a meaningful proportion of the total query sequence coverage, we employed:

$$\text{Sequence Identity (as a product of query coverage) (\%)} = \left(\frac{\text{Number of Identical Residues}}{\text{Length of the Query Protein}}\right) \times 100\%$$

This approach enables the calculation of each hit's coverage as a percentage of the total query length, yielding a unified metric. By correlating the predictive power of the models with overall sequence identity and coverage, we can evaluate whether the models are effectively generalising or merely identifying homologous sequences. Pearson correlation coefficient was used to evaluate relationships between sequence similarity and model performance. Correlations and p-value were calculated using the scipy package, considering p-values <0.05 statistically significant.

Results

To corroborate the strength of embeddings as input features compared to hand-crafted features, for protein interface prediction, we constructed several neural network models with and without embedding features. All models were trained on the protein-protein training set BioDL_P_TR. Evaluation and selection of models were done on a 20% subset of the training set. Final selected models were subsequently tested on the ZK448, BioDL_P_TE, and MTB datasets, as detailed below.

Table 1. Impact of embeddings and other feature groups on the performance of the ensemble predictors

Features	F1	MCC	AP	AUC	ACC	SPEC
1H+SI+PSSM+PL	0.339	0.249	0.302	0.755	0.840	0.909
EMB	**0.396**	**0.313**	0.372	**0.798**	**0.854**	**0.917**
EMB+PL	**0.395**	**0.313**	**0.377**	**0.800**	**0.854**	**0.917**
EMB+SI	0.373	0.287	0.329	0.772	0.849	**0.914**
EMB+SI+PSSM	0.384	0.299	0.361	0.793	0.851	**0.915**
EMB+PL+SI+PSSM	0.382	0.290	0.321	0.784	**0.858**	**0.913**

Feature ablation study of *ensnet* model without (top) and with (EMB) embedding features, trained on BioDL_P_TR and tested on BioDL_P_TE. Reported metrics are Accuracy (ACC), Specificity (SPEC), F1 Score, Matthews Correlation Coefficient (MCC), Average Precision (AP), and Area Under the ROC Curve (AUC). Scores up to 0.005 below highest per metric in bold.

Embeddings yield higher performance

We studied the improvement achieved after introducing ProtT5-XL embeddings as input features for the ensemble network (*ensnet*) in the BioDL_P_TE. Substantial improvement can be seen for all metrics when using the embeddings, as shown inTable 1. To further discern the primary contributors to model performance, we examined various combinations of features: Protein Length (PL), Structural Information (SI), Profiles (PSSM) and ProtT5-XL Embeddings (EMB). We evaluated EMB, EMB+PL, EMB+SI, EMB+SI+PSSM and EMB+SI+ PSSM+PL (Table 1). For all combinations, differences were minimal, with MCC ranging from 0.290 for EMB+PSSM to 0.313 for EMB+PL. None of the 'hand-crafted' features, SI and PSSM, demonstrate any improvement. These results align with previous findings on the strengths of embeddings[26]. From a practical standpoint, this implies that the computation of hand-crafted features has become unnecessary.

Table 2. Performance results of ensemble models comprising different architecture combinations

ann	cnet	dnet	rnn	rnet	unet	F1	MCC	AP	AUC	ACC	SPEC
–	–	–	✓	✓	✓	**0.395**	**0.313**	**0.377**	**0.800**	**0.854**	**0.917**
–	✓	✓	✓	✓	✓	0.388	0.294	0.362	0.788	**0.851**	0.890
–	✓	–	✓	✓	–	0.372	0.302	0.360	0.788	**0.855**	0.886
–	–	–	–	✓	✓	0.385	0.294	0.364	0.796	**0.854**	0.901
✓	✓	✓	✓	✓	✓	0.365	0.287	0.352	0.787	0.843	0.895

Selected combinations per row, trained on `BioDL_P_TR` and tested on `BioDL_P_TE`. Metrics as in Table 1, up to 0.005 below highest in bold.

Table 3. Comparing PIPENN-EMB to state-of-the-art sequence-based predictors on the `ZK448` benchmark dataset

Method	F1	MCC	AP	AUC	ACC
Seq-InSite[40]	**0.535**	**0.462**	**0.617**	**0.853**	**0.874**
PIPENN-EMB*	0.505	0.392	0.513	0.805	0.815
EnsemPPIS[29]	0.385	0.291	0.354	0.770	0.821
DELPHI[47]	0.364	0.278	0.326	0.746	0.848
PIPENN[12]	0.339	0.249	0.302	0.755	0.840
SCRIBER[48]	0.333	0.230	0.287	0.715	n.a.
SPRINGS[49]	0.229	0.111	0.201	0.625	n.a.

Metrics cf. Table 1. * this work.

Composition of the ensemble network matters

Our ensemble network is a flexible neural network. Given that it orchestrates the other six PIPENN architectures and utilises their predictions to perform its own predictions, we decided to assess the contribution of each architecture to the ensemble model. Table 2 shows the performance of various combinations of the individual PIPENN architectures. Each combination comprises an ensemble network. Interestingly, we observed that a combination of *rnn*, *rnet*, and *unet* exhibits slightly superior performance compared to other combinations. While this improvement may seem marginal, it holds significance from a practical standpoint, as it allows for simplification of the overall ensemble architecture. Moreover, excluding *ann*, *cnet*, and *dnet* from the ensemble results in a reduction of approximately 6 million parameters.

We designate the suite of *rnn*, *rnet*, and *unet* neural net models along with the *ensnet* ensemble net, incorporating the EMB+PL features, as PIPENN-EMB. We will further explore the performance of the PIPENN-EMB *ensnet* model.

Comparison to state-of-the-art methods

To put the improved PIPENN-EMB performance in context, we compared it against leading sequence-based protein-protein interface prediction methods on the benchmark `ZK448` dataset. Table 3 confirms that PIPENN-EMB significantly surpasses the previous PIPENN model across all metrics except accuracy. Other recent methods such as DELPHI[47] and EnsemPPIs[29] are somewhat inbetween. The very recent Seq-InSite seems to stand out, and according to its authors even exceeds structure-based predictions[40]. Only for the Seq-InSite method we were able to obtain prediction scores on the `BioDL_P_TE`, `ZK448` and `MTB` datasets that allows us to draw ROC and P/R plots for detailed comparison. Figure 1 shows that, consistently for both ROC and P/R, Seq-InSite outperforms PIPENN-EMB.

Table 4. Comparing state-of-the-art sequence and structure based prediction methods on the `MTB` use-case dataset

Method	F1	MCC	AP	AUC	ACC
PIPENN-EMB	0.458	0.319	0.320	0.659	0.779
Seq-InSite	0.572	0.463	0.415	0.731	0.825
PeSTo	0.609	0.508	0.450	0.754	0.840
CSM-potential2	0.400	0.244	0.281	0.623	0.753

Metrics as introduced in Table 1.

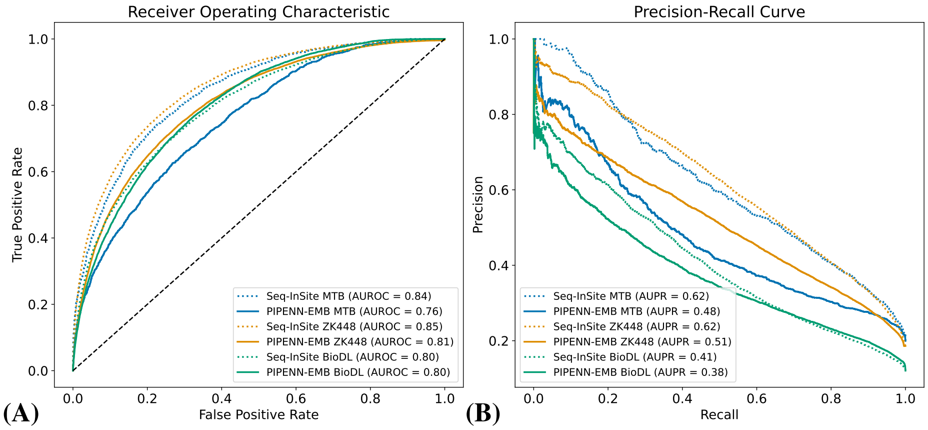

Figure 1. Model performance of state of the art sequence based prediction methods. (A) AUROC and (B) Precision-Recall plots for comparison of models tested on the BioDL_P_TE test set, the MTB use case, and the ZK448 benchmark dataset. Curves correspond to the interface prediction performance for the sequence-based PIPENN-EMB and Seq-InSite. Legends report AUROC and AUPR values as measure of overall prediction capabilities.

Sequence- and structure-based models perform on par on Mycobacterium tuberculosis proteins

Protein-protein interactions in MTB have been suggested as a key factor in driving anti-microbial resistance in TB[50]. To enhance our understanding of these interactions, we evaluate the performance of PIPENN-EMB and selected state-of-the-art methods on this specialised external use-case. This allows for further assessment of the methods' robustness in recognising a particular subset of prokaryotic proteins, thereby providing insights into the models' performances within more specialised scenarios and real life applications.

For comparison, we included sequence-based Seq-InSite[40], and structure-based CSM-Potential2 [41, CSM] and PeSTo[42]. Table 4 shows that sequence-based PIPENN-EMB performing overall quite close with the structure-based predictors, with an AUC score of 0.659, compared to 0.754 for structure-based PeSTo, and 0.623 for CSM, when using AlphaFold2 (AF2) structures. With an MCC of 0.319, PIPENN-EMB performs inbetween both structure-based methods, resp. 0.463 and 0.244. The very recent sequence-based Seq-InSite can be seen to outperform the other methods, including PIPENN-EMB, on both AUC and AP (resp. 0.731 and 0.415). The AP of PIPENN-EMB was 0.320, underscoring its capability to accurately extract critical and generalisable information from sequence characteristics to predict protein interfaces.

The MTB dataset moreover allows us to assess generalisability of the methods tested. Since its introduction in 2018[10], the ZK448 dataset has been extensively used for method development, potentially introducing a bias towards this specific dataset. This raises concerns that supposedly unseen benchmark data, i.e. ZK448, might have influenced design decisions for methods benchmarked on it. Below, we will further explore these differences in generalising capabilities.

Generalising beyond homology

Observed differences in performance may be influenced by homology between training and test datasets, leading to an inflated estimate of the method's capability to generalise to new, unseen data.

To assess the dependency of models' predictions on homology between training and test datasets, we (1) utilise BLASTp to find the closest hits, (2) measure the degree of sequence similarity, and (3) calculate the correlation between the degree of sequence similarity and a model's AUC-ROC. For robustness, we performed this analysis on all three test-sets: BioDL_P_TE, ZK448 and MTB.

The results shown significant correlations between sequence similarity and coverage to the training set and the performance of Seq-InSite in Figure 2A & B, while PIPENN-EMB shows no significant correlation on any of the test-sets in Figure 2. The intersection of the correlation trends between PIPENN-EMB and Seq-InSite suggests PIPENN-EMB's capacity for generalisation beyond homology, contrasting with Seq-InSite's apparent dependency on homologous sequences for performance in Figure 2A & B. Thus, the higher performance of Seq-InSite seen in Table 4 might be inflated through homology. The results emphasise the robustness of PIPENN-EMB against data leakage and its capability to maintain consistent performance irrespective of sequence similarity.

To provide some more insight into each model's ability to generalise predictions of proteins beyond homology with respect to the training set, we categorise proteins from these datasets into close (31-45%) and distant homologs (16-30%) and sequences that are less than 15% similar. Figure 3 shows the distribution of AUC-ROC performance of the two sequence-based methods

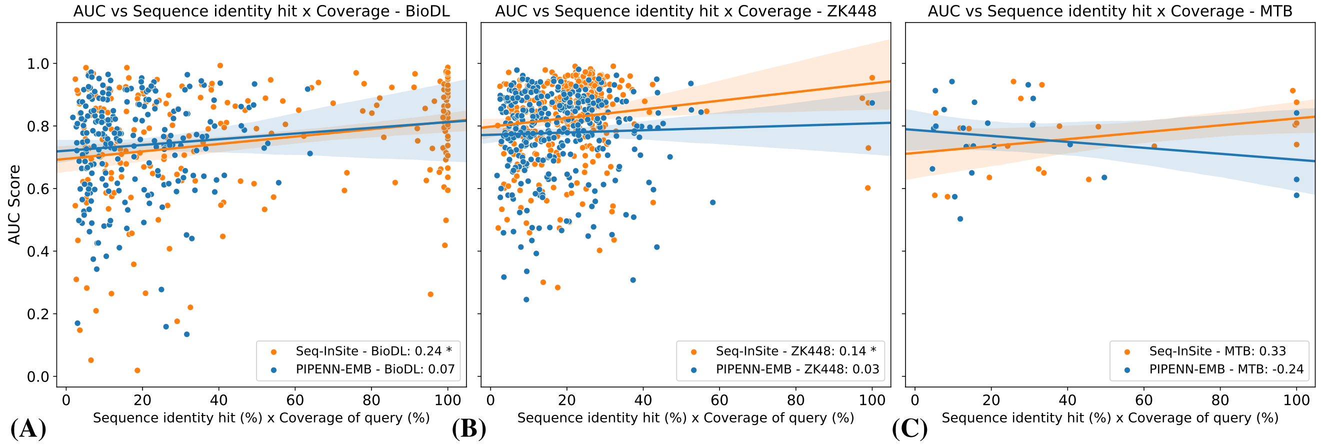

Figure 2. Correlation between model performance and sequence homology for different test datasets. Data points represent individual proteins from the different datasets, showing the method prediction performance in AUC-ROC versus the (the product of) sequence similarity and query coverage for hits in the method's training set. The linear regression lines indicate trends of AUC-ROC with increasing sequence similarity and coverage. Shaded areas show the 95% confidence intervals. Correlation coefficients (R^2) are annotated in the legends for each model, with an asterisk (*) for a p-value < 0.05. Results are shown for three datasets: (A) `BioDL_P_TE`, (B) `ZK448`, and (C) `MTB`.

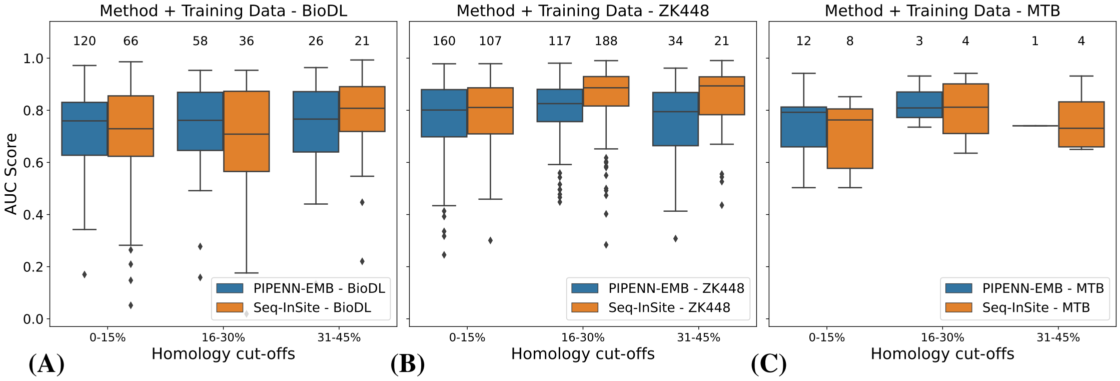

Figure 3. Performance of PIPENN-EMB and Seq-Insite for distant homologs from their respective training data. Boxplots illustrate the differences in the distribution of AUC scores for different homology cut-offs to the respective training data of the prediction methods. Three thresholds based on sequence identity times query coverage were applied: proteins with 0-15% identity times coverage are considered remote homologs, 16-30% are considered distant homologs, and proteins between 31-45% are considered close homologs to the respective training data for the model. The number of proteins used for each distribution is indicated above the boxes. Results are shown for three datasets: (A) `BioDL_P_TE`, (B) `ZK448`, and `MTB`.

for these three categories. For Seq-InSite, there is a noticeable drop in performance towards the more remote catagories, which is most pronounced in the BioDL_P_TE and ZK448 data sets: highest scores are found in the close (31-45%) and distant category (16-30%), with a decline in performance for similarities below 15%. This pattern is indicative of Seq-InSite's reliance on homology for predictive power, which may also be observed in Figure 2. Conversely, PIPENN-EMB displays robust performance across different levels of homology, maintaining consistent average AUC scores that do not significantly differ between categories. This consistency is evident across all three datasets, showcasing PIPENN-EMB's ability to generalize well beyond homology. This suggests that PIPENN-EMB is particularly effective in handling diverse protein sequences, reinforcing its capabilities for predicting protein interfaces of proteins that are distant homologs from the training data.

The Webserver

We also make PIPENN_EMB available as a webserver at www.ibi.vu.nl/programs/pipennemb/. The webserver is simple to use and is aimed at non-experts in both academia and industry. The input is only the protein sequence. The predictions are based on the DNET architecture predictions. We are still working on a version that includes also the ensemble model predictions. Expected runtimes are one to two minutes per protein, including feature generation (actual throughput may be lower due to queue time). For comparison, approximate runtimes for the PIPENN webserver are 5-10 minutes, Seq-InSite 20 minutes, CSM 5-10 minutes, and PeSto 10 seconds per protein.

Discussion

Protein interaction interface prediction remains a challenging task despite the improvements in deep learning models that we have been seeing during the last years[12,29,40,44]. In this work we built on top of PIPENN by introducing embeddings next to a new ensemble combination achieving a comptetent performance in the benchmark ZK448 dataset, and showcase the potential for generalisation on the new MTB dataset.

Embeddings provide a powerful framework to improve protein-based models by utilizing them as features[26,51,52]. Although for protein interface prediction performance seems to be reaching a plateau, these representation models are still being upgraded and show increased capability to encode complex relationships across protein sequence space. Additionally, previous work has shown that deep learning models struggle to reach generalisation in protein interface prediction due to biases in the training set and data leakage to the test set[43]. Strategies to overcome these problems lie in using more robust architectures that aim to achieve a better generalisation[31], such as ensemble architectures, next to using smarter approaches when curating the training set[43]. A practical approach is to include new unseen data when testing new methods.

Training on structures is becoming more popular since AlphaFold2 structures are available[53]. Having access to the 3D information allows models to understand geometrical conformations involved in interactions. However, a certain degree of bias can be introduced due to the annotations that are made based on PDB structures. On top of this, representation models benefit more from sequences because there are more protein sequences available than structures. Models that combine both structural and sequential aspects of proteins are showing promising results, however, they are still in early development[54].

The tuberculosis use-case gives insights into how PIPENN-EMB performs in identifying protein interactions within MTB resistance proteins, which includes new unseen proteins that are not homologous to the training data. Here, PIPENN-EMB outperforms state-of-the-art sequence-based Seq-InSite, particularly for very distant proteins. We verified this trend in the BioDL_P_TE and ZK448 datasets, where we find it even more pronounced. This demonstrates the ability of the PIPENN-EMB model to generalise beyond homology, making it a valuable tool for predicting interactions in proteins less represented in training datasets. Moreover, novel pathogen genomes often contain unknown proteins for which no known homologs are available. Thus, the robustness to generalise interface predictions beyond homology is crucial for advancing our understanding of resistance mechanisms and to help developing new treatments.

In summary, we have contributed with the following:

(1) introducing embeddings to boost PIPENN performance;
(2) introducing a new ensemble architecture (PIPENN-EMB) that achieves state-of-the-art performance;
(3) comparing the performance of PIPENN-EMB with other state-of-the-art sequence-based and structured-based models on the ZK448 and MTB data sets, respectively; and
(4) evaluating performance biases of PIPENN-EMB and another state-of-the-art sequence-based model on the BioDL_P_TE, ZK448, and MTB datasets, introduced by data leakage through homology.

Acknowledgements

We thank Jody Phelan and Nick Furnham for their insightful discussions about the role of protein interactions in the development of MTB drug resistance, and their suggestion to investigate the prediction of the binding residues involved. We also thank SeyedMohsen Hosseini and Lucian Ilie for very kindly providing the training data for their Seq-InSite method, as well as help with technical issues on their webserver.

Author contributions statement

R.H. and K.A.F conceptualised the research goals and aims. R.H. and C.M.G.F. developed the methodology and software. C.M.G.F. and D.P.G.T. conducted the experiments, analysed the raw results, and visualised the data presentation. All authors analysed results, and wrote and edited the manuscript.

Additional information

The MTB dataset is available at https://github.com/jodyphelan/TBProfiler/blob/master/db/tbdb.bed. The training and test sets and code for training and testing are available from the PIPENN github https://github.com/ibivu/pipenn/.

References

1. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc. Nat. Acad. Sci.* **93**, 13–20 (1996).
2. Westermarck, J., Ivaska, J. & Corthals, G. L. Identification of protein interactions involved in cellular signaling. *Mol Cell Proteom* **12**, 1752–1763, DOI: 10.1074/mcp.R113.027771 (2013).
3. Cheng, F. *et al.* Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nat. Genet.* **53**, 342–353, DOI: 10.1038/s41588-020-00774-y (2021).
4. Sperandio, O. Editorial: Toward the design of drugs on protein-protein interactions. *Curr pharmaceut design* **18**, 4585 (2012).
5. Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nat. Rev Drug Discov* **15**, 533–550, DOI: 10.1038/nrd.2016.29 (2016).
6. Mannar, D., Ahmed, S. & Subramaniam, S. AAA ATPase protein–protein interactions as therapeutic targets in cancer. *Curr Opin Cell Biol* **86**, 102291, DOI: 10.1016/J.CEB.2023.102291 (2024).
7. Das, S. & Chakrabarti, S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Sci Rep* **11**, 1–12, DOI: 10.1038/s41598-020-80900-2 (2021).
8. Ezkurdia, I. *et al.* Progress and challenges in predicting protein–protein interaction sites. *Brief Bioinf* **10**, 233–246, DOI: 10.1093/bib/bbp021 (2009).
9. Xue, L. C., Dobbs, D., Bonvin, A. M. & Honavar, V. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Lett.* **589**, 3516–3526, DOI: 10.1016/J.FEBSLET.2015.10.003 (2015).
10. Zhang, J. & Kurgan, L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief Bioinf* **19**, 821–837, DOI: 10.1093/bib/bbx022 (2018). https://academic.oup.com/bib/article-pdf/19/5/821/25861134/bbx022.pdf.
11. Hou, Q., De Geest, P., Vranken, W., Heringa, J. & Feenstra, K. Seeing the trees through the forest: Sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics* **33**, DOI: 10.1093/bioinformatics/005 (2017).
12. Stringer, B. *et al.* PIPENN: protein interface prediction from sequence with an ensemble of neural nets. *Bioinformatics* **38**, 2111–2118, DOI: 10.1093/bioinformatics/btac071 (2022).
13. Lensink, M. F. *et al.* Impact of AlphaFold on structure prediction of protein complexes: The CASP15-CAPRI experiment. *Proteins* **91**, 1658–1683 (2023).
14. Jumper, J. *et al.* Highly accurate protein structure prediction with {AlphaFold}. *Nature* **596**, 583–589, DOI: 10.1038/s41586-021-03819-2 (2021).
15. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596, DOI: 10.1038/S41586-021-03828-1 (2021).
16. Su, H. *et al.* Improved protein structure prediction using a new multi-scale network and homologous templates. *Adv. Sci.* 2102592, DOI: 10.1002/ADVS.202102592 (2021).
17. Xie, Z. & Xu, J. Deep graph learning of inter-protein contacts. *Bioinformatics* DOI: 10.1093/bioinformatics/btab761 (2021).
18. Thornton, J. M., Laskowski, R. A. & Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Medicine 2021 27:10* **27**, 1666–1669, DOI: 10.1038/s41591-021-01533-0 (2021).

19. Jones, D. T. & Thornton, J. M. The impact of AlphaFold2 one year on. *Nat. Methods* **19**, 15–20, DOI: 10.1038/s41592-021-01365-3 (2022).

20. Elnaggar, A., Heinzinger, M., Dallago, C. & Rost, B. End-to-end multitask learning, from protein language to protein features without alignments. *bioRxiv* DOI: 10.1101/864405 (2019).

21. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. *Sci. Reports* **11**, 1160, DOI: 10.1038/s41598-020-80786-0 (2021).

22. Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K. & Rost, B. Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci. Reports* **11**, 1–15, DOI: 10.1038/s41598-021-03431-4 (2021).

23. Capel, H., Feenstra, K. A. & Abeln, S. Multi-task learning to leverage partially annotated data for PPI interface prediction. *Sci Rep* **12**, 10487, DOI: 10.1038/s41598-022-13951-2 (2022).

24. Capel, H. *et al.* ProteinGLUE multi-task benchmark suite for self-supervised protein modeling. *Sci. Reports* **12**, 16047, DOI: 10.1038/s41598-022-19608-4 (2022).

25. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110, DOI: 10.1093/bioinformatics/btac020 (2022).

26. Villegas-Morcillo, A. *et al.* Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *Bioinformatics* **37**, 162–170, DOI: 10.1093/bioinformatics/btaa701 (2021).

27. Elnaggar, A. *et al.* ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112–7127, DOI: 10.1109/TPAMI.2021.3095381 (2022).

28. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Nat. Acad. Sci.* **118**, e2016239118, DOI: 10.1073/PNAS.2016239118 (2021).

29. Mou, M. *et al.* A Transformer-Based Ensemble Framework for the Prediction of Protein–Protein Interaction Sites. *Research* **6**, DOI: 10.34133/research.0240 (2023).

30. Cui, Y., Dong, Q. & Hong, D. Predicting protein-ligand binding residues with deep convolutional neural networks. *BMC Bioinforma.* **20** (2019).

31. Ganaie, M., Hu, M., Malik, A., Tanveer, M. & Suganthan, P. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **115**, 105151, DOI: https://doi.org/10.1016/j.engappai.2022.105151 (2022).

32. World Health Organization. *Global tuberculosis report 2023* (World Health Organization, Geneva, 2023).

33. Tunstall, T. *et al.* Combining structure and genomics to understand antimicrobial resistance. *Comput. Struc Biotech J.* **18**, 3377–3394, DOI: 10.1016/J.CSBJ.2020.10.017 (2020).

34. Tunstall, T., Phelan, J., Eccleston, C., Clark, T. G. & Furnham, N. Structural and Genomic Insights Into Pyrazinamide Resistance in Mycobacterium tuberculosis Underlie Differences Between Ancient and Modern Lineages. *Front Mol Biosci* **8**, 619403, DOI: 10.3389/FMOLB.2021.619403/BIBTEX (2021).

35. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinforma.* **20**, 723, DOI: 10.1186/s12859-019-3220-8 (2019).

36. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288, DOI: 10.1093/bioinformatics/btm098 (2007).

37. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 1–8, DOI: 10.1038/s41467-018-04964-5 (2018).

38. Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome medicine* **11**, 1–7 (2019).

39. PDBe-KB Consortium. PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.* **50**, D534–D542, DOI: 10.1093/nar/gkab988 (2021). https://academic.oup.com/nar/article-pdf/50/D1/D534/42058803/gkab988.pdf.

40. Hosseini, S., Golding, G. B. & Ilie, L. Seq-InSite: sequence supersedes structure for protein interaction site prediction. *Bioinformatics* **40**, DOI: 10.1093/BIOINFORMATICS/BTAD738 (2024).

41. Rodrigues, C. H. M. & Ascher, D. B. Csm-potential2: A comprehensive deep learning platform for the analysis of protein interacting interfaces. *Prot: Struc, Func, Bioinf* DOI: https://doi.org/10.1002/prot.26615 (2023). https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26615.

42. Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F. & Dal Peraro, M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat. Commun.* **14**, 1–11, DOI: 10.1038/s41467-023-37701-8 (2023).

43. Lannelongue, L. & Inouye, M. Pitfalls of machine learning models for protein-protein interaction networks. *Bioinformatics* btae012, DOI: 10.1093/bioinformatics/btae012 (2024). https://academic.oup.com/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btae012/55399607/btae012.pdf.

44. Hou, Q., Waury, K., Gogishvili, D. & Anton Feenstra, K. Ten quick tips for sequence-based prediction of protein properties using machine learning. *PLOS Comput. Biol.* **18**, e1010669, DOI: 10.1371/journal.pcbi.1010669 (2022).

45. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol* **215**, 403–410, DOI: 10.1016/S0022-2836(05)80360-2 (1990).

46. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinforma.* **10**, 1–9, DOI: 10.1186/1471-2105-10-421 (2009).

47. Li, Y., Brian Golding, G. & Ilie, L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinf* **37**, 896–904, DOI: 10.1093/BIOINFORMATICS/BTAA750 (2021).

48. Zhang, J. & Kurgan, L. SCRIBER: Accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* **35**, i343–i353, DOI: 10.1093/bioinformatics/btz324 (2019).

49. Singh, G., Dhole, K., Pai, P. P. & Mondal, S. SPRINGS: Prediction of Protein-Protein Interaction Sites Using Artificial Neural Networks. *J Proteom & Comput. Biol* **1**, 7 (2014).

50. Singh, P. *et al.* Computational modeling and bioinformatic analyses of functional mutations in drug target genes in mycobacterium tuberculosis. *Comput. Struct. Biotechnol. J.* **19**, 2423–2446 (2021).

51. Heinzinger, M. *et al.* Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics Bioinforma.* **4**, DOI: 10.1093/NARGAB/LQAC043 (2022).

52. Outeiral, C. & Deane, C. M. Codon language embeddings provide strong signals for protein engineering. *bioRxiv* 2022.12.15.519894, DOI: 10.1101/2022.12.15.519894 (2022).

53. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589, DOI: 10.1038/s41586-021-03819-2 (2021).

54. Jha, K. & Saha, S. Amalgamation of 3D structure and sequence information for protein–protein interaction prediction. *Sci. Reports* **10**, 1–14, DOI: 10.1038/s41598-020-75467-x (2020).