PLM-interact: extending protein language models to predict protein-protein interactions

Dan Liu[1], Francesca Young[1], Kieran D. Lamb[1], Adalberto Claudio Quiros[2,4], Alexandrina Pancheva[3], Crispin Miller[2,3], Craig Macdonald[4], David L Robertson[1, #], Ke Yuan[2,3,4,#]

[1]MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom;

[2]School of Cancer Sciences, University of Glasgow, Glasgow, United Kingdom;

[3]Cancer Research UK Scotland Institute, Glasgow, United Kingdom;

[4]School of Computing Science, University of Glasgow, Glasgow, United Kingdom.

#Shared senior/corresponding authors: david.l.robertson@glasgow.ac.uk, ke.yuan@glasgow.ac.uk

Abstract

Computational prediction of protein structure from amino acid sequences alone has been achieved with unprecedented accuracy, yet the prediction of protein-protein interactions (PPIs) remains an outstanding challenge. Here we assess the ability of protein language models (PLMs), routinely applied to protein folding, to be retrained for PPI prediction. Existing PPI prediction models that exploit PLMs use a pre-trained PLM feature set, ignoring that the proteins are physically interacting. Our novel method, PLM-interact, goes beyond a single protein, jointly encoding protein pairs to learn their relationships, analogous to the next-sentence prediction task from natural language processing. This approach provides a significant improvement in performance: Trained on human-human PPIs, PLM-interact predicts mouse, fly, worm, *E. coli* and yeast PPIs, achieving between 16-28% improvements in AUPR compared with state-of-the-art PPI models. Additionally, it can detect changes that disrupt or cause PPIs and be applied to virus-host PPI prediction. Our work demonstrates that large language models can be extended to learn the intricate relationships among biomolecules from sequences alone.

Introduction

Proteins are the main structural components of cells and mediate biological processes by interacting with other proteins[1]. Disruption of these protein-protein interactions (PPIs), e.g., mediated by mutations can underlie human disease[2]. In virology PPIs are particularly important as viruses depend entirely on the host cell for replication, achieved through

specific interactions with host proteins. Understanding PPI mechanisms offers the potential for developing novel therapy strategies for both human disease and pathogen infections[3]. Unfortunately, experimentally identifying PPIs is both costly and time-consuming, such that interaction datasets remain sparse with only a few species having comprehensive coverage[4,5].

Computational algorithms offer an efficient alternative to the prediction of PPIs at scale. Existing approaches mainly leverage protein properties such as protein structures sequence composition and evolutionary information[6,7,8,9]. Leveraging these features from pairs of proteins, classifiers have been trained using classical machine learning[10] and deep learning approaches[11]. Recently, protein language models (PLMs) trained on large public protein sequence databases have been used for encoding sequence composition, evolutionary and structural features[12–14], becoming the method of choice for representing proteins in state-of-the-art PPI predictors. A typical PPI prediction architecture uses a pre-trained PLM to represent each protein in a pair separately, then a classification head is trained for a binary task that discriminates interacting pairs from non-interacting pairs[13,15] (**Figure 1a**). Despite this use of PLMs, identifying positive PPIs remains challenging.

The issue is PLMs are primarily trained using single protein sequences, i.e., while they learn to identify contact points within a single protein[16], they are not aware of interaction partners. In a conventional PLM-based PPI predictor architecture, a classification head is used to extrapolate the signals of inter-protein interactions by grouping common patterns of intra-protein contacts in interacting and non-interacting pairs, respectively (**Figure 1a**). This strategy relies on the classification head being generalisable. Unfortunately, with a feedforward neural network being the dominant option, these classifiers often don't generalise well.

To address the lack of inter-protein context in training, we propose a novel PPI prediction model, PLM-interact, that directly models PPIs by extending and fine-tuning a pre-trained PLM, ESM-2[17]. PLM-interact (trained on human PPI data) achieves a significant improvement compared to other predictors when applied to mouse, fly, worm, yeast and *E. coli* datasets. We demonstrate that PLM-interact is capable at identifying mutations that cause and disrupt interactions. We also show PLM-interact's performance in a virus-human PPI prediction task.

PLM-interact

To directly model PPIs, two extensions to ESM-2 are introduced (**Figure 1b**): (1) longer permissible sequence lengths in paired masked-language training to accommodate amino acid residues from both proteins; (2) implementation of the "next sentence" prediction task[18] to fine-tune ESM-2 where the model is trained with a binary label indicating whether the protein pair is interacting or not (see Methods for more details). Our training task is, thus, a mixture of the next sentence prediction and the mask language modelling tasks. This architecture allows amino acids in a protein to be linked by amino acids from a different protein through transformer's attention mechanism.

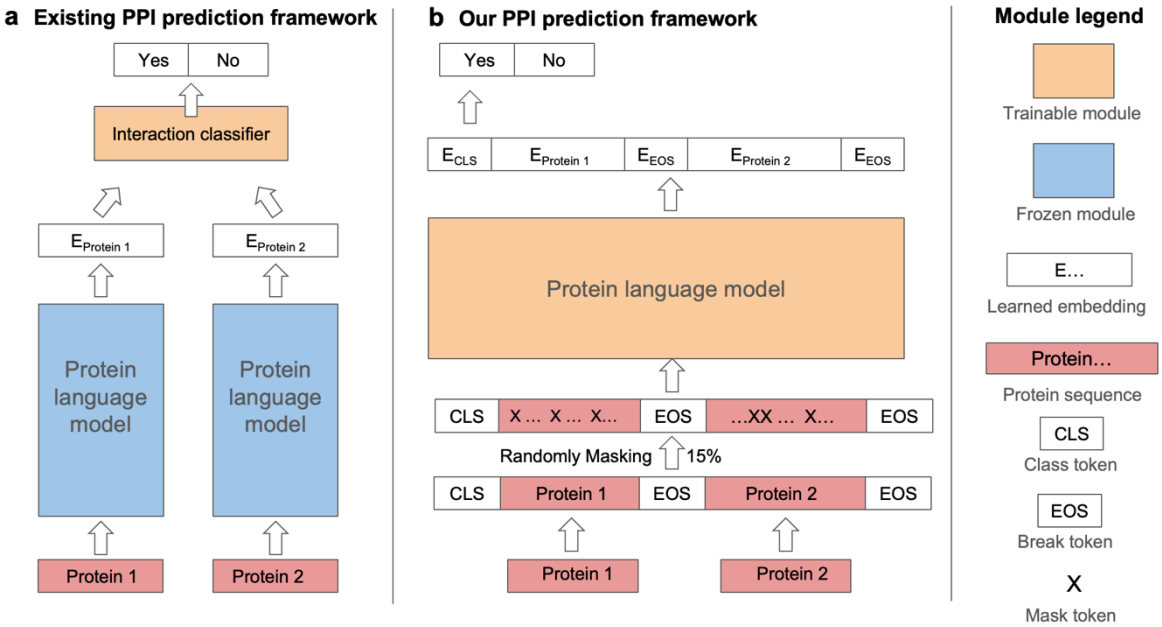


Figure 1. A comparison of PLM-interact to the existing framework. a. Traditional PPI prediction models use pre-trained protein language models to generate embeddings of single proteins and train a classifier using those single protein embeddings as features. **b.** PLM-interact uses a protein language model with longer context to handle a pair of proteins directly. Both the mask language modelling task and a binary classification task predicting interaction status are used to train the model. (see **Supplementary Figure 1**).

The training of PLM-interact begins with the pre-trained ESM-2. We fine-tune it for PPIs, by showing it pairs of known interacting and non-interacting proteins. In contrast to similar training strategies in machine learning[18], we find the next sentence prediction and mask language modelling objectives need to be balanced. We therefore conducted comprehensive benchmarking for different weighting options, before selecting a 1:10 ratio between

classification loss and mask loss, combined with initialisation using the ESM-2 with 650M parameters, as this achieved the best performance. This configuration is the PLM-interact model setting in the following benchmarking tasks (for details, see Methods and **Supplementary Figure 1).**

PLM-interact improves prediction performance

To examine the performance of PLM-interact, we benchmark the model against five PPI prediction approaches: TT3D[13], Topsy-Turvy[19], D-SCRIPT[15], PIPR[6] and DeepPPI[20]. We use a multi-species dataset created by Sledzieski et al.[15]. Each model is trained on human protein interaction data and tested on five other species. The human training dataset contains 421,796 PPIs, and the human, mouse, worm, fly, and yeast test datasets contain 55K, except the *E. coli* test dataset, which contains 22K. The positive PPIs in these datasets are experimentally-derived physical interactions.

PLM-interact achieves the highest AUPR (area under the precision-recall curve)[23] with the next best performer, TT3D[13], which exhibits very similar performance to Topsy-Turvy[19] (**Figure 2b**). Tested on mouse, fly, and worm test species datasets, PLM-interact has an improvement of 13%, 20%, and 20% AUPR compared to TT3D[13], respectively. The predictions for yeast and *E. coli* PPIs are more challenging because they are more evolutionarily divergent from human dataset used for training (see **Figure 2b**): Our model achieved an AUPR of 0.707 on yeast, a 28% improvement over TT3D's AUPR of 0.553, and a 26% improvement on *E. coli* with an AUPR of 0.722.

The improvement in PLM-interact is due to its ability to correctly identify positive PPIs: Comparing the prediction interaction probability of PLM-interact with the second-best predictor TT3D, PLM-interact consistently assigned higher probabilities of interaction to true positive PPIs. TT3D, in contrast, produces a bimodal distribution for interaction probabilities in all held-out species (more details see **Supplementary Figure 3**).

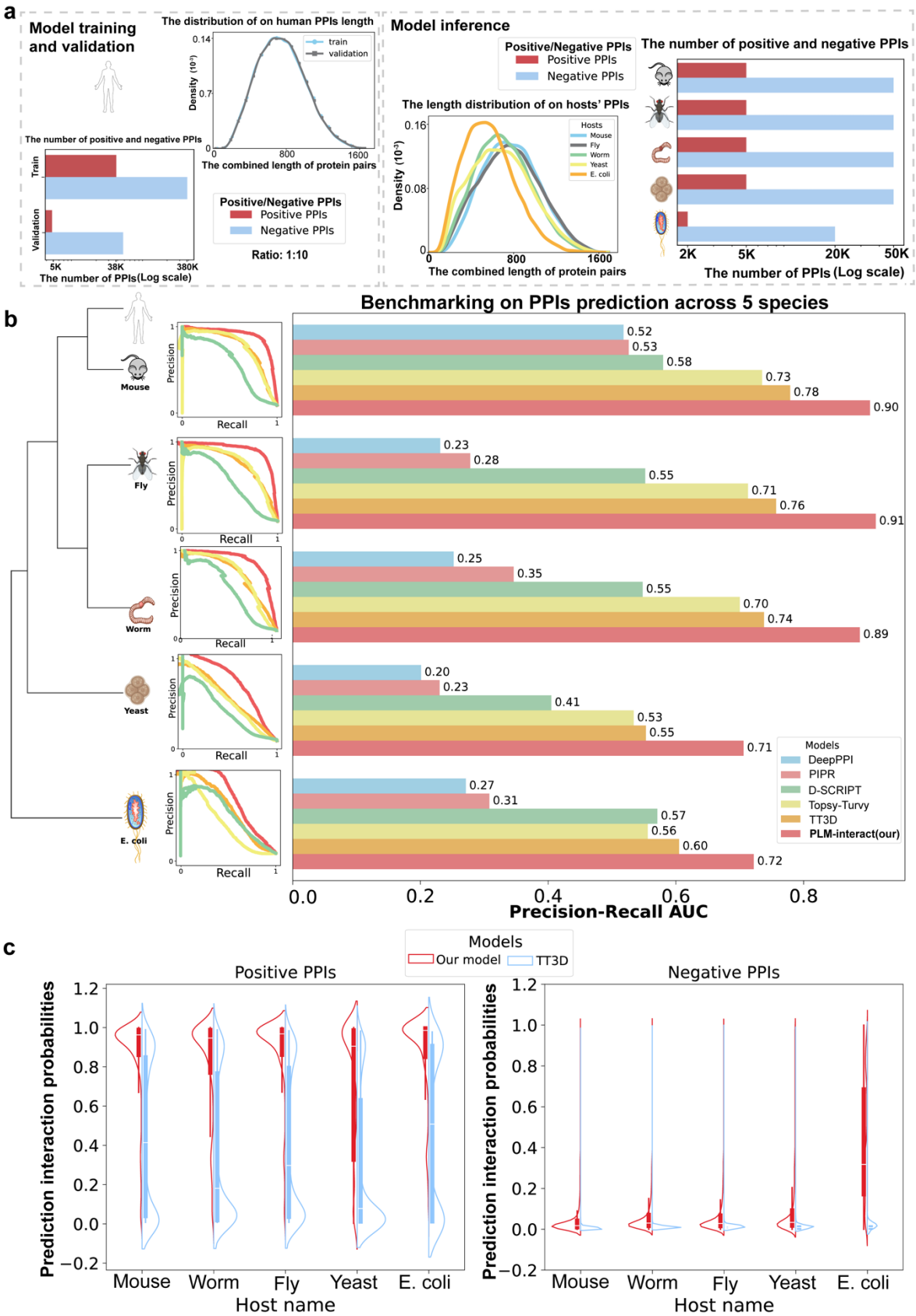


Figure 2 PLM-interact achieves the best performance on within-species PPI prediction. The benchmarking results of PLM-interact with the state-of-the-art PPI models. **a.** The size and

distribution of combined length of protein pairs of training, validation, and test data. **b.** The taxonomy tree of the training and test hosts, aligned with precision-recall curves of each test host. Our model achieved the highest AUPR. **c**. Violin plots of predicted interaction probabilities of PLM-interact and TT3D on positive and negative PPIs, respectively. (see **Supplementary Figure 3** for more information).

Next, we showcase five positive PPI instances, one for each test host species. where our model produces a correct prediction, while TT3D produces an incorrect prediction (**Figure 3**). These PPIs are necessary for essential biology processes in the host species including ubiquinone biosynthesis, RNA polymerisation, ATP catalysis, transcriptional activation and protein transportation. We use Chai-1[21] to predict and visualise these protein-protein structures (**Figure 3**). Despite four out of five predicted structures having low confident scores (ipTM<0.6), PLM-interact gives correct predictions with high confidence.

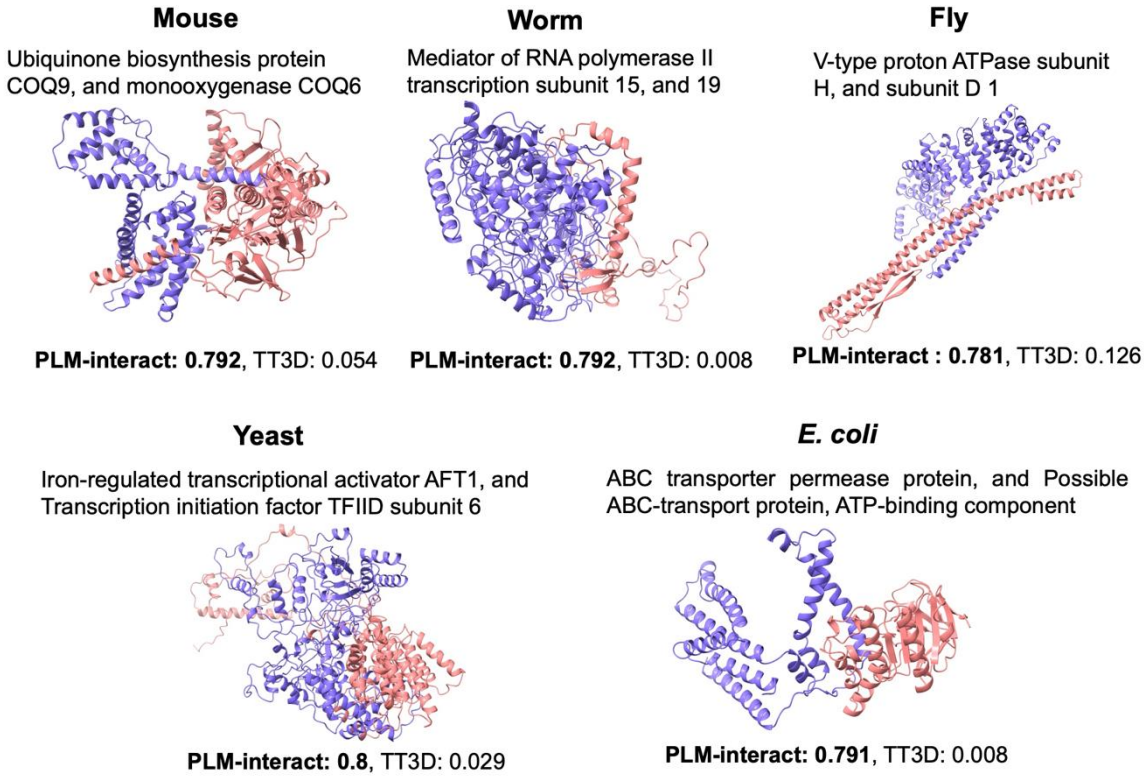


Figure 3. We show a PPI example for each host species that was predicted correctly by PLM-interact while getting false predictions by TT3D. Protein-protein structures are predicted by Chai-1[22] and visualised by ChimeraX[23]. Both models' prediction interaction probability range between 0 and 1. When the predicted interaction probability is > 0.5, it is predicted as a positive PPI. Otherwise, it is predicted as a negative PPI. **Mouse**: The PPI is Q8K1Z0 (Ubiquinone biosynthesis protein COQ9, mitochondrial) and Q8R1S0 (Ubiquinone biosynthesis monooxygenase COQ6,

mitochondrial); **Worm**: The PPI is Q21955 (Mediator of RNA polymerase II transcription subunit 15) and Q9N4F2 (Mediator of RNA polymerase II transcription subunit 19); **Fly**: The PPI is Q9V3J1 (V-type proton ATPase subunit H) and Q9V7D2 (V-type proton ATPase subunit D 1); **Yeast**: The PPI is P22149 (Iron-regulated transcriptional activator AFT1) and P53040 (Transcription initiation factor TFIID subunit 6); ***E. coli***: The PPI is A0A454A7G5 (ABC transporter permease protein) and A0A454A7H5 (Possible ABC-transport protein, ATP-binding component).

PLM-interact can identify impact of mutations on interactions

Here, we demonstrate examples of PLM-interact correctly predicting the consequences of mutations on PPIs. Wild-type protein sequences are retrieved from UniProt[24]. Amino acid substitutions associated with changes in PPIs are obtained from InAct[25]. We obtained predicted interaction probabilities for both wild-type and mutants from PLM-interact trained on human data (see Methods). For mutation-causing interactions (see **Figure 4a**), PLM-interact's predicted interaction probabilities of the wild-type PPI are below 0.5, whereas the interaction probabilities of the mutant PPI exceed 0.5. For mutation-disrupting PPIs (**Figure 4b**), the predicted interaction probabilities of the wild-type PPI exceed 0.5, whereas the interaction probabilities of the mutant PPI are below 0.5. Examples are presented in **Figure 4**, where complex structures are predicted by Chai-1.

We show two examples of mutations that cause PPIs in **Figure 4a,** the protein Ataxin-1 and SOD1 (Superoxide dismutase [Cu-Zn]), which in humans are encoded by the ATXN1 and SOD1 genes separately. Ataxin-1 interacts with many other proteins, and the expansion of a glutamine(Q)-encoding repeat can affect the function of PPIs and cause the genetic disease spinocerebellar ataxia type1 (SCA1) and other polyglutamine diseases[26]. Predictions from PLM-interact show that prior to mutation, this PPI has a predicted interaction probability of 0.411, correctly indicating non-interaction with E2s (Ubiquitin-conjugating enzyme E2 E3). Following mutation, PLM-interact increases this score to 0.771, correctly predicting that the mutation induces interaction. In the second example (**Figure 4a**), the SOD1 gene encodes superoxide dismutase enzymes that break down human superoxide radicals. SOD1 is linked to the nervous system disease amyotrophic lateral sclerosis (ALS)[27], with the A4V mutation being the most common variant in North America[28]. PLM-interact predicts 0.465 and 0.851 before and after the mutation, correctly capturing the change in interaction with CDK4 (Cyclin-dependent kinase 4).

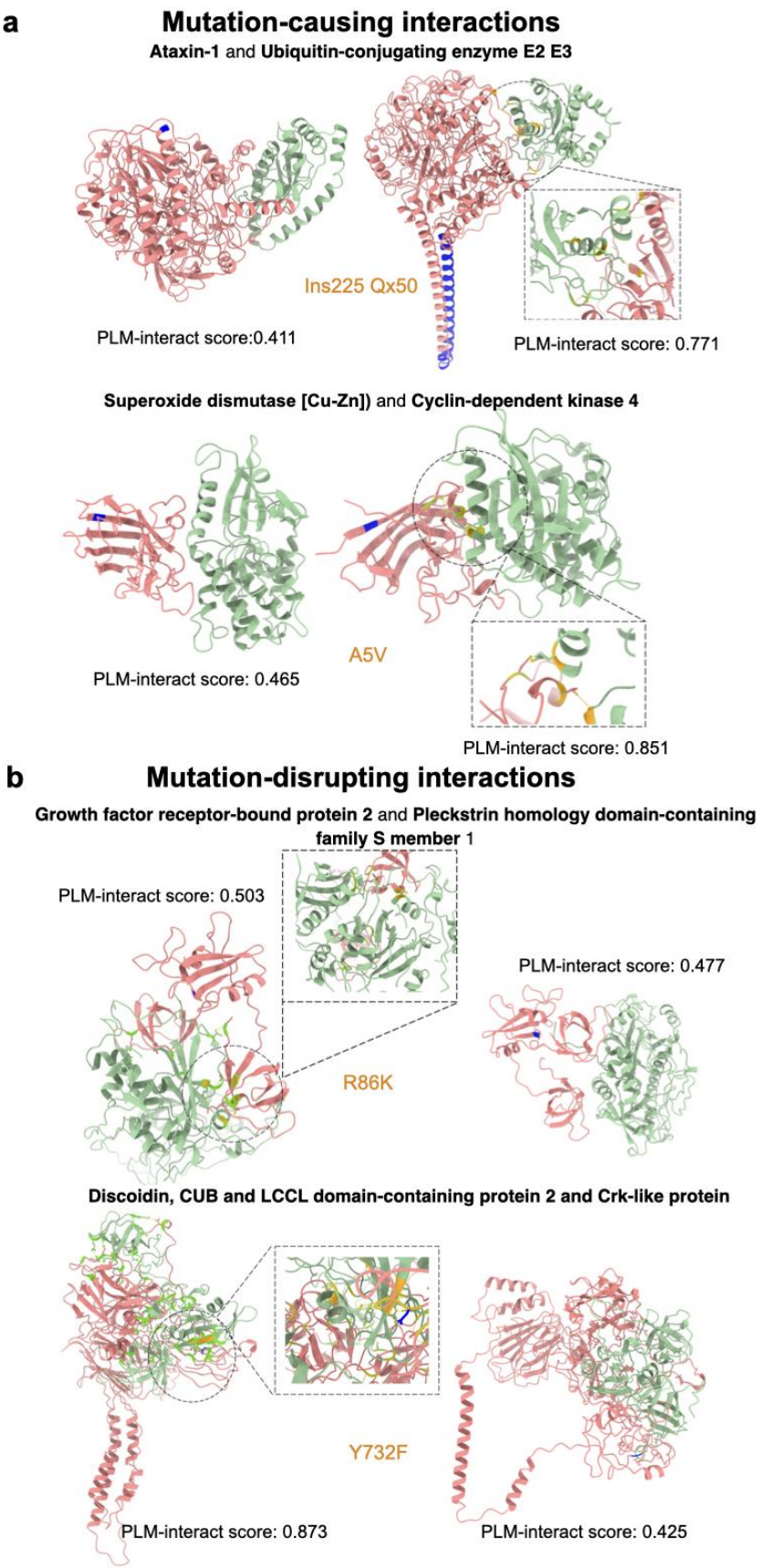


Figure 4. PLM-interact can detect changes in human PPIs caused by mutations. **a** shows two mutation-causing interaction examples, while **b** shows two mutation-disrupting PPI examples. These

PPI structures are predicted using Chai-1[22] and visualised by ChimeraX[23]; here, the mutated amino acids are highlighted in blue, and the H-Bonds involved intramodal are shown in orange and are magnified to emphasise PPIs. Prediction interaction probabilities exceeding 0.5 indicate the proteins interact, while below 0.5 indicate non-interact. **a.** The 225th amino acid Glutamine (Q) of P54253 (Ataxin-1) is mutated to 50 Q, causing interaction with Q969T4 (Ubiquitin-conjugating enzyme E2 E3)[29]; the 5th amino acid Alanine (A) of P00441 (Superoxide dismutase [Cu-Zn]) is mutated to Valine (V), causing interaction with P11802 (Cyclin-dependent kinase 4)[30]. **b.** The 86th amino acid Arginine(R) of P62993 (Growth factor receptor-bound protein 2) is mutated to Lysine (K), disrupting its interaction with Q5SXH7-1 (Pleckstrin homology domain-containing family S member 1)[31]; the 732nd amino acid Tyrosine (Y) of Q96PD2 (Discoidin, CUB and LCCL domain-containing protein 2) is mutated to Phenylalanine (F), disrupting its interaction with P46109 (Crk-like protein)[32].

Next, we show two examples of mutations that disrupt PPIs (**Figure 4b**). First, GRB2 (growth factor receptor bound protein 2) is associated with signal transduction. GRB2 mutations are associated with multiple cancers, including breast cancer[33] and leukaemia[34]. The wild type protein interacts with PLEKHS1 (Pleckstrin homology domain-containing family S member 1). PLM-interact predicts that the missense mutation R86K reduces the interaction probability from 0.503 to 0.477. Second, CLCP1 is a transmembrane protein that regulates cell growth, and this protein is identified as a cancer marker, exhibiting up-regulated expression in lung cancer[35]. Again, PLM-interact predicts that the missense mutation Y732F reduces the interaction probability from 0.873 to 0.425.

Improved virus-human PPI prediction

To study virus-host PPI prediction, we train PLM-interact on a virus-human PPIs dataset from Tsukiyama et al.[11]. The dataset is derived from the Host-Pathogen Interaction Database (HPIDB) 3.0[36], comprising a total of 22,383 PPIs, which include 5,882 human and 996 virus proteins. We compare our model with three recent virus-human PPI models, including the PLM-based approaches: STEP[14], LSTM-PHV[11] and InterSPPI[37]. STEP is similar to existing PPI models benchmarked previously; it leverages protein sequence embedding extracted by the pre-trained PLM ProtBERT[38]. The results show that PLM-interact outperforms the other models, demonstrating improvements in the AUPR, F1 and MCC scores of 5.7%, 10.9% and 11.9%, respectively, compared with STEP **(Figure 5a**). To further analyse our model's performance, we select three pairs of virus-human PPIs from our test data, all with corresponding experimental virus-human complex structures available in the HVIDB[39]. We

then use ChimeraX[23] to visualise these structures and present PLM-interact's predicted interaction probability for each example (see **Figure 5c, d, e**).

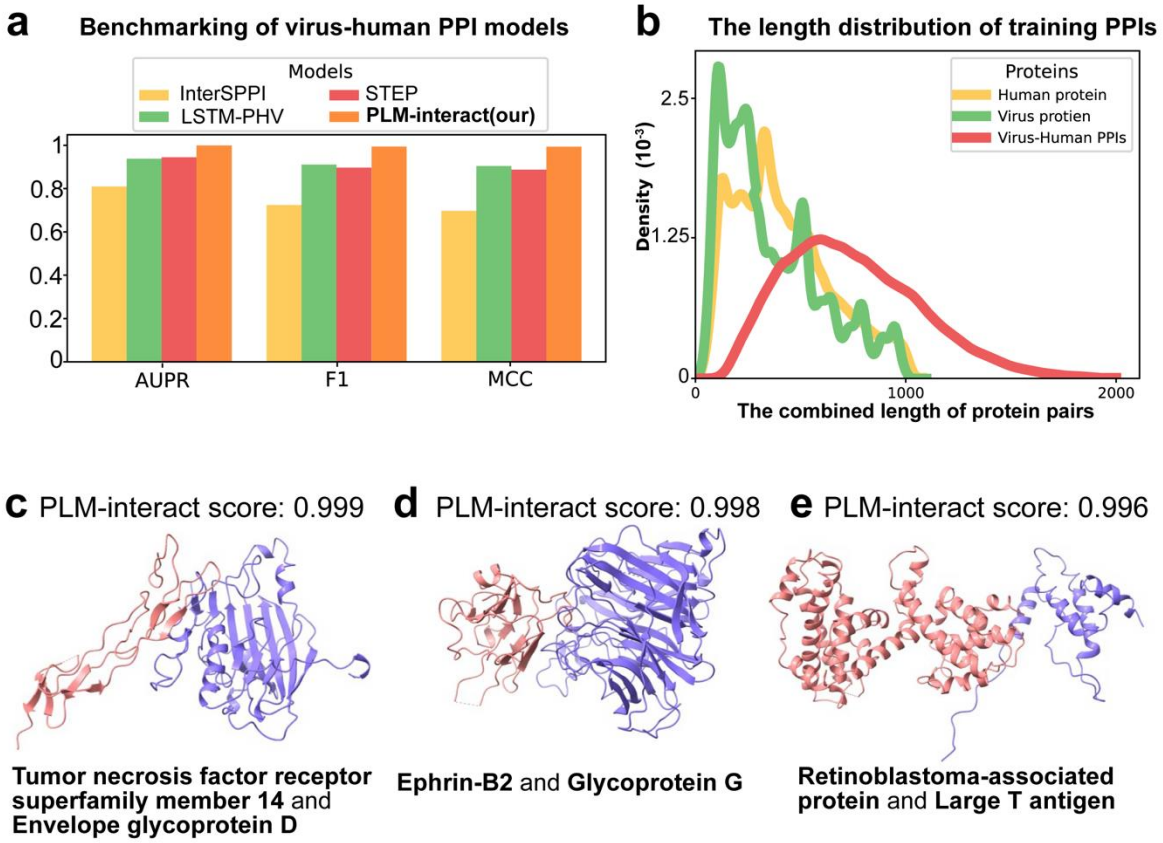


Figure 5. a. The bar plot presents a comparison of the AUPR, F1, and MCC metrics of PLM-interact against recent virus-human PPI models. **b**. The distribution of the combined length of virus proteins, human proteins and virus-human protein-protein pairs. **c-e**. The virus-human PPIs predicted correctly by our model, and the 3D complex structures of virus-human PPIs are experimentally verified structures collected from the human-virus PPI database (**HVIDB**[39]). **c.** The PPI structure between Tumour necrosis factor receptor superfamily member 14 (Human protein: Q92956) and Envelope glycoprotein D (human herpes simplex virus 1: P57083). **d**. The PPI structure between Ephrin-B2 (human protein: P52799) and Glycoprotein G (Nipah virus protein: Q9IH62). **e.** The PPI structure between Retinoblastoma-associated protein (human protein: P06400) and Large T antigen (Simian virus 40: P03070). Note: The metrics results of the other three models in **a** are taken from STEP[14] paper. Human proteins are shown in pink, and virus proteins are shown in purple.

Discussion

In this study, we have developed a novel PPI prediction method, PLM-interact that extends single protein-focused PLMs to their interacting protein partner. We report significant improvements in multiple held-out species and highlight successful examples predicting mutational effects on interactions in both the causing and disrupting categories. We further demonstrate PLM-interact's performance in a virus-human PPI prediction task, showing significant improvement over state-of-the-art approaches.

We show that underlying the benefit of PLM-interact is the improved capability of correctly predicting positive PPIs in held-out species. The positive PPIs are important as they are of high biological interest, but they are challenging to predict due to a lack of high-quality data. Our improved performance relative to TT3D is particularly impressive given the use of sequence alone. TT3D[13] includes explicit structural information, the pre-residual structural alphabet in Foldseek[40], to improve its prediction over Topsy-Turvy. Further improvements could be achieved in a similar fashion.

Furthermore, our results show the potential of predicting mutational effects on PPIs from sequence alone. This could lead to a new generation of interaction aware in-silico variant effect predictors where state-of-the-art methods rely on PLMs of the single proteins[41–43]. However, current training data remains limited. The number of high-quality structures of mutant proteins and their interaction partners are low. Algorithmically, models with long and multimodal context[44–46] that include multiple proteins, structures and nucleotides could be specialised for interaction tasks.

Finally, effective virus-host PPI predictors could provide the much-needed molecular detail to conventional virus-host species prediction tools, which often ignore host molecules that are interacting with viral molecules[47–50]. In those approaches, the host species only act as labels. Recent progress within SARS-CoV-2 PPI studies mapped out a complex interaction landscape between the virus and human proteome[51,52]. Other viruses are likely to have similarly complex interactions with human and animal hosts. Leveraging these interactions could lead to tools that are better at predicting zoonotic events and the emergence of novel viruses. While PLM-interact has demonstrated significant improvements, we want to note

that high-quality virus-host PPI data remains scarce. Further progress would require breakthroughs in data and algorithms.

Methods

In this study, the main objective is to train the protein language model ESM-2 with masking and classification processes based on a sentence-BERT model[18]. Similar to sentence transformers library, it follows the *cross-encoders architecture* for sentence pair scoring and classification tasks. PLM-interact enables PPI prediction based on the entire protein pair training on transformer *cross-encoders*. It encodes protein pairs as an entire sequence to the ESM-2; thereby allowing the model to capture attention across the amino acids between two proteins.

To evaluate our model's performance, we train our model using human PPI benchmarking datasets from Sledzieski et al.[15] and virus-human PPI benchmarking data produced by Tsukiyama et al.[11]. Using the DiRAC Extreme Scaling GPU based system Tursa, the model training on human PPIs from Sledzieski et al.[15] used 16 A100-80 GPUs for 31 hours, while the model trained on virus-human PPIs from Tsukiyama et al.[11] used 8 A100-80 GPUs for 30.5 hours.

Datasets

The benchmarking human PPI dataset, sourced from Sledzieski et al.[15], comprises human training and validation data, and test data from five other host species (Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, and *E. coli Escherichia coli*), all retrieved from STRING V11[53]. We train and validate our model on human PPIs and then conduct inference on PPIs from other host species. All training, validation, and test datasets maintain a 1:10 ratio of positive to negative PPIs, as positive PPIs are significantly fewer than negative PPIs in the actual host PPI networks. The length of protein sequences ranges from 50 to 800, and PPIs are clustered at 40% identity using cd-hit[54] to remove the redundant PPIs. The human training PPI dataset includes 38,345 positive PPIs, whereas the validation set contains 5,000 positive PPIs. Each of the five species includes 5,000 positive interactions, except for $E.\ coli$, which only has 2,000 positive interactions due to the fewer positive PPIs in the STRING.

The benchmarking dataset of 22,383 virus-human PPIs includes 5,882 human and 996 virus proteins. This dataset was obtained from Tsukiyama et al.[11], sourced from the HPIDB 3.0 database[36]; the ratio of positive to negative PPIs is 1:10, and negative PPIs are chosen based on sequence dissimilarities. The length of protein sequences ranges from 30 to 1,000, and the redundant PPIs are filtered based on a threshold of 95% identity using cd-hi[54].

In addition, we download human PPIs from STRING V12[55] and select positive PPIs by collecting physical links with positive experimental scores, while excluding PPIs with positive homology scores and confidence score below 400. Previous studies have typically limited the maximum length of protein sequences to 800 or 1000 due to GPU memory limitations. We process the length of protein sequences up to 2000, with a combined length threshold for PPIs of nearly 2500. Furthermore, protein sequences are clustered at 40% identity using MMSeq2[56], and only PPIs from the distinct clusters are chosen to eliminate redundant PPIs. Again, the positive-to-negative PPI ratio is 1:10, consistent with the aforementioned two benchmarking datasets.

Model architecture

We mask 15% of the amino acids in each protein pair and then use the ESM-2 mask language modelling to train on these protein pairs and obtain mask loss. A linear classification layer is added to the top of the mask language model to get the classification loss for the PPI binary classification task. **Figure 1** shows the processes of our training model, PLM-interact, which trains ESM-2 as a *cross-encoder*.

ESM-2 has a family of transformer PLMs, and the number of parameters ranges from 8 million up to 15 million. ESM-2 is trained to predict amino acids that are randomly masked out of protein sequence pairs[17]:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in M} \log p\left(x_i | x_{\setminus M}\right)$$

where M is the random mask that includes 15% of amino acid position i in the protein sequence pairs. The model predicts the mask amino acids x based on the context surrounding the protein-protein sequence.

Cross-encoder is a type of neural network architecture[18] successfully used in various NLP tasks, such as semantic textual similarity[57] and information retrieval[18]. The *cross-encoder* architecture simultaneously processes both input texts, allowing for self-attention scores and understanding of semantic content across both inputs. The self-attention[58] mechanism enables the model to maintain long-range information throughout the entire input sequence pair. For the input protein P_1 and P_2, the *cross-encoder* processes them as:

$$P_{pair} = [CLS, P_1, EOS, P_2, EOS]$$

Here, the special token CLS (classification) marks the beginning of the sequence. The CLS token embedding aggregates the representation of the entire sequence pair, which is used for sequence classification. The token EOS (end-of-sequence) marks the separator and the end of the sequences. The input sequence pairs are tokenised and converted into vector representations using the pre-trained ESM-2. The *cross-encoder* architecture of PLM-interact uses a shared encoder for both proteins in the pair; the joint encoding allows the model to learn the contact and contextual significance of the PPIs to one another. Each token in the protein pairs is converted into a dense representation using an embedding layer, and the embedded sequence is passed through multiple transformer encoder layers. The representation of CLS is the input to the feedforward classification head of ESM-2 for predicting interaction probabilities in PPIs.

Overall, ESM-2 has been pre-trained on billions of protein sequences from UniRef[59]. We train ESM-2 based on the *cross-encoder* architecture to model PPIs as a pair-wise comparison, treating protein pairs as analogous to sentence-sentence or question-answer formats in NLP tasks. In this study, PLM-interact is initialised with ESM-2 and trained on protein pairs to learn protein interactions.

Model Training and Evaluation

We train PLM-interact for ten epochs using an efficient batch size of 128 on both benchmarking datasets of human PPIs and virus-human PPIs, as mentioned above. To evaluate the model's performance on validation data during training, evaluation is conducted every 2000 steps. In the evaluation process, 128 PPIs are randomly subsampled from the validation dataset, and the results are averaged over 100 iterations to ensure the metric

reliability. The iterations evaluate a subset of validation datasets, enabling effective monitoring of the model's performance while optimising time and GPU resources. In addition, the dataset exhibits a significant imbalance, with a positive-to-negative ratio of 1:10. Consequently, we evaluate our model's performance using the area under the precision-recall curve (AUPR).

We conduct benchmarking inference on other species of PPIs by training and validating PLM-interact using human PPIs established by Sledzieski et al[15]. The optimal model checkpoint is selected based on its highest APPR performance on the validation set. Next, we evaluate the trained model on PPIs from mouse, fly, worms, yeast, and *E. coli*. In addition, PLM-interact is trained on virus-human PPI benchmarking data and tested on independent PPIs produced by Tsukiyama et al.[11]. We provide model checkpoints that include human PPI models trained on Sledzieski et al.'s benchmarking datasets retrieved from STRING V11[53], the virus-human PPIs model trained on the benchmarking virus-human PPIs created by Tsukiyama et al.[11], sourced from HPIDB 3.0[36], as well as human PPIs created by us collected from STRING V12[53].

PLM-interact ablation experiments

Our model is based on the ESM-2 mask language model and linear binary classification for PPI prediction. To investigate the importance of both the mask language model and binary classification on PPI prediction performance, we conducted a benchmarking analysis using the different ratios between mask loss and classification loss in Sledzieski et al.'s benchmarking data. We train five models initialised by ESM-2-35M and ESM-2-650M separately, where four models have different ratios (0:1,1:1,1:5 and 1:10) between the mask loss and classification loss, and one model is based on sequence binary classification (**Supplementary Figure 1 b, c)**. We used the human validation set for each model to identify the optimal epoch checkpoint achieving the best AUPR. Next, the trained human PPI models are evaluated by testing on five other host PPIs, respectively.

Based on the results of McNemar's Test (**Supplementary Figure 1 d, e**), our model, using a 1:10 ratio of the mask to classification loss based on ESM-2-650M, demonstrates the most significant improvements (McNemar test, p-value < 0.05) compared to other models. The model with a 1:5 ratio of mask to classification loss based on ESM-2-35M achieved the

highest number of significant improvements (McNemar test, p-value < 0.05) compared to other models. According to these results, we select a loss ratio of 1:10 for ESM-2-650M and 1:5 for ESM-2-35M for model training. The ratio setting is implemented in both benchmarking of human PPI training and virus-human PPI training, as well as human PPI training using the STRING V12 database. The number of epochs is set to 10, the efficient batch size is 128, and the warmup steps are set to 2000.

To investigate how our model leverages the benefits of mask loss for the model performance, we use token prediction accuracy and perplexity to evaluate the model's performance across three tasks with varying ratios of mask to classification loss (1:1, 1:5 and 1:10). High token prediction accuracy is better, whereas low perplexity is better. **Supplementary Figure 2** shows the performance of these three models; the mask token accuracy of mouse and fly is higher than yeast and *E. coli* in both the 35M and 650M models, and the 650M model is better than 35M. The perplexity of the 35M model is low at a loss ratio of 1:5, especially yeast exhibiting low perplexity, indicating that the model typically achieves lower perplexity on test species at this loss ratio. The mouse, fly, and worm have a closer evolutionary relationship to the training species humans, resulting in higher token prediction accuracy and lower perplexity than the evolutionary divergent species yeast and *E. coli*. These results are consistent with the benchmarking results of within-species PPI prediction. Overall, the varying ratios of the mask-to-classification loss influence model performance, and the larger (650M) model generally performs better with higher token accuracy and low perplexity.

Baselines

We recompute the prediction interaction probabilities based on checkpoints of TT3D[13], Topsy-Turvy[19] and D-SCRIPT[15] to plot precision-recall (PR) curves in **Figure 2b**. Due to the absence of publicly available checkpoints for DeepPPI[20] and PIPR[6], these methods are excluded from the PR curve comparison. The AUPR values for DeepPPI and Topsy-Turvy are sourced from the Topsy-Turvy paper[19], those for D-SCRIPT and PIPR are from the D-SCRIPT paper[15], and the AUPR value of TT3D is obtained through email communication.

MMseq2

We used MMseq2 to obtain the protein sequence-based alignment results between each pair of proteins; the parameters setting is:
--threads 128 --min-seq-id 0.4 --alignment-mode 3 --cov-mode 1.

Chai-1

Chai-1 is the state-of-the-art model for molecular structure prediction, and it is available on a web interface (https://lab.chaidiscovery.com/). We use Chai-1 to predict protein-protein structure complex and visualise PPI structures by using the molecular visualisation program ChimeraX[23].

Evaluation metrics and statistical analysis

Evaluation metrics

In this study, we use AUPR to evaluate models' performance and plot Precision-Recall (PR) curves[60] to compare our PPI model with the recent PPI models. For the imbalanced training and test PPIs, the prediction-recall AUC curve is more informative than the ROC curve in evaluating the model's performance.

AUPR is calculated based on precision and recall at different classification thresholds and represents the area under the precision-recall curve, summarising the model's overall performance across different decision thresholds.

Precision is the proportion of true positives on all predicted positive samples.

$$Precision = \frac{TP}{(TP + FP)}$$

Recall is the proportion of all true class samples predicted to be true classes.

$$Recall = \frac{TP}{(TP + FN)}$$

True positive (TP): The model correctly predicts the positive samples. True Negative (TN): The model correctly predicts the negative samples. False Positive (FP): The model predicts

negative samples to be positive. False Negative (FN): The model predicts positive samples to be negative.

We used the F1 and Matthews Correlation Coefficient (MCC) to evaluate virus-human PPI models' performance.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN))}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

We used perplexity[61] to analysis our models' performance on masking token prediction task; perplexity shows the complexity of a model, and the lower perplexity indicates that the model is better.

McNemar's test

McNemar's Chi-Square test[62] is a statistical test that determines if there are significant differences between paired nominal data.

$$McNemar's\ test = \frac{(\frac{model1_{true}}{model2_{false}} - \frac{model1_{fasle}}{model2_{true}})^2}{(\frac{model1_{true}}{model2_{false}} + \frac{model1_{fasle}}{model2_{true}})}$$

Here, $model1_{true}$ and $model2_{true}$ separately represent the count of correct prediction obtained by $model1$ or $model2$, while $model1_{fasle}$ and $model2_{false}$ separately represent the count of incorrect prediction obtained by $model1$ and $model2$.

To investigate if our models with varying ratios of mask-to-classification loss perform significant differences, we conducted a McNemar's test for any pairs of models in ablation experiments. This test is based on the number of correct and incorrect between two models. Prediction scores of each model are used to get predicted label, which are compared with true label to get the count of correct and incorrect prediction. A McNemar's test p-value ≤ 0.05 indicates a significant difference between the two models. We consider the model with

more correct predictions superior, and when the p-value is below 0.05, we can demonstrate that this improvement is statistically significant.

Data availability

Sledzieski et al.'s benchmarking PPI data is available at https://d-script.readthedocs.io/en/stable/data.html

Tsukiyama et al.'s virus-human benchmarking PPIs dataset is available at: http://kurata35.bio.kyutech.ac.jp/LSTM-PHV/download_page

STRING V12 PPIs database: https://stringdb-downloads.org/download/protein.physical.links.v12.0.txt.gz

The mutations that cause or disrupt PPIs are from IntAct[25] Database; the link is https://ftp.ebi.ac.uk/pub/databases/intact/current/various/mutations.tsv

UniProt: https://www.uniprot.org/

HVIDB: http://zzdlab.com/hvidb/download.php

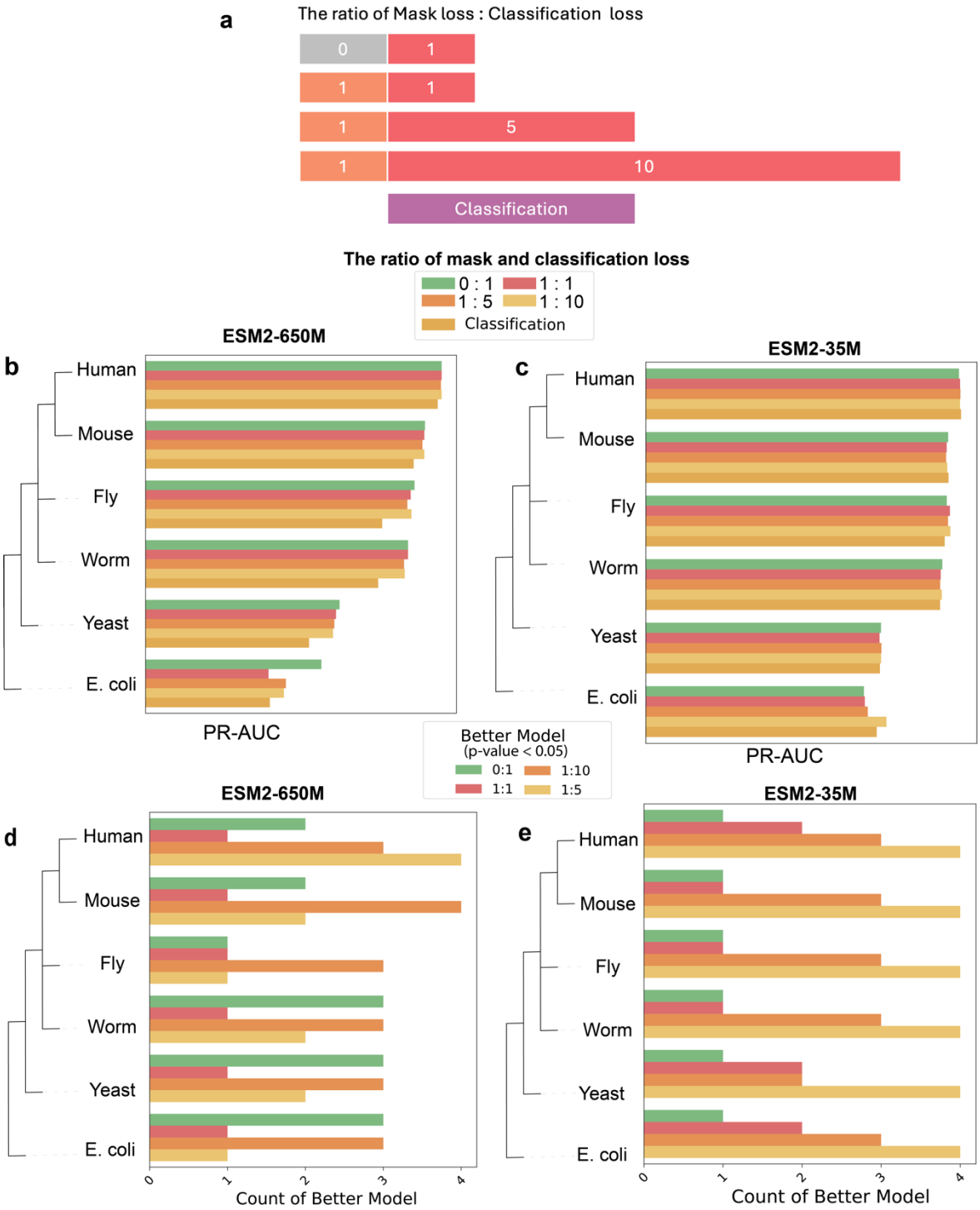
Code availability

The code and trained models are available at https://github.com/liudan111/PLM-interact

Acknowledgements

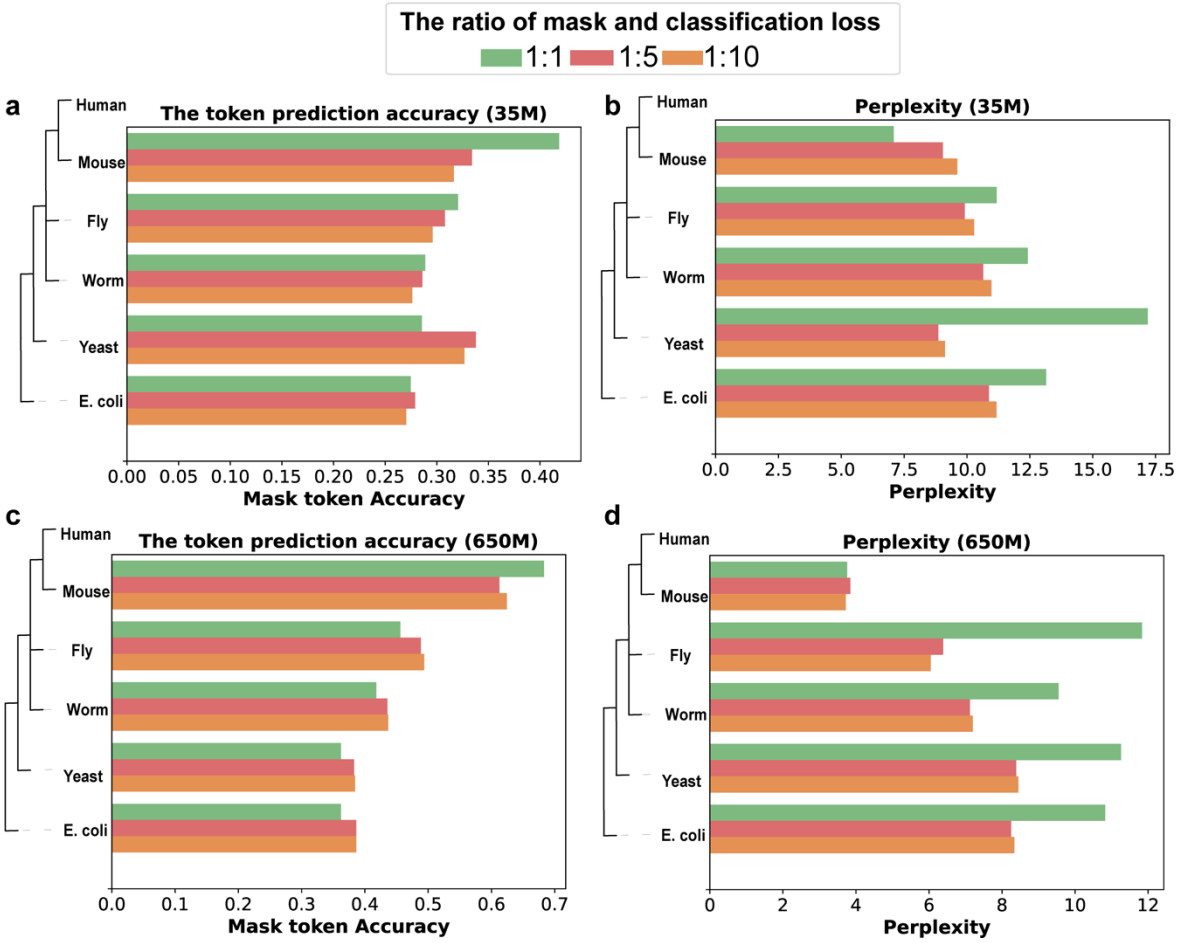
The authors acknowledge funding from the European Union's Horizon 2020 research and innovation 562 program, under the Marie Sklodowska-Curie Actions Innovative Training Networks 563 grant agreement no. 955974 (VIROINF) for DL, a UK Medical Research Council (MRC) Doctoral Training Programme in Precision Medicine studentship (MR/N013166/1) for KDL and MRC grants: MC_UU_00034/5, MC_UU_00034/6 and MR/V01157X/1. This work used the DiRAC Extreme Scaling service (Tursa) at the University of Edinburgh, managed by the Edinburgh Parallel Computing Centre on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The DiRAC service at Edinburgh was funded by BEIS, UKRI and STFC capital funding and STFC operations grants. DiRAC is part of the UKRI Digital Research Infrastructure.

Supplementary Figures

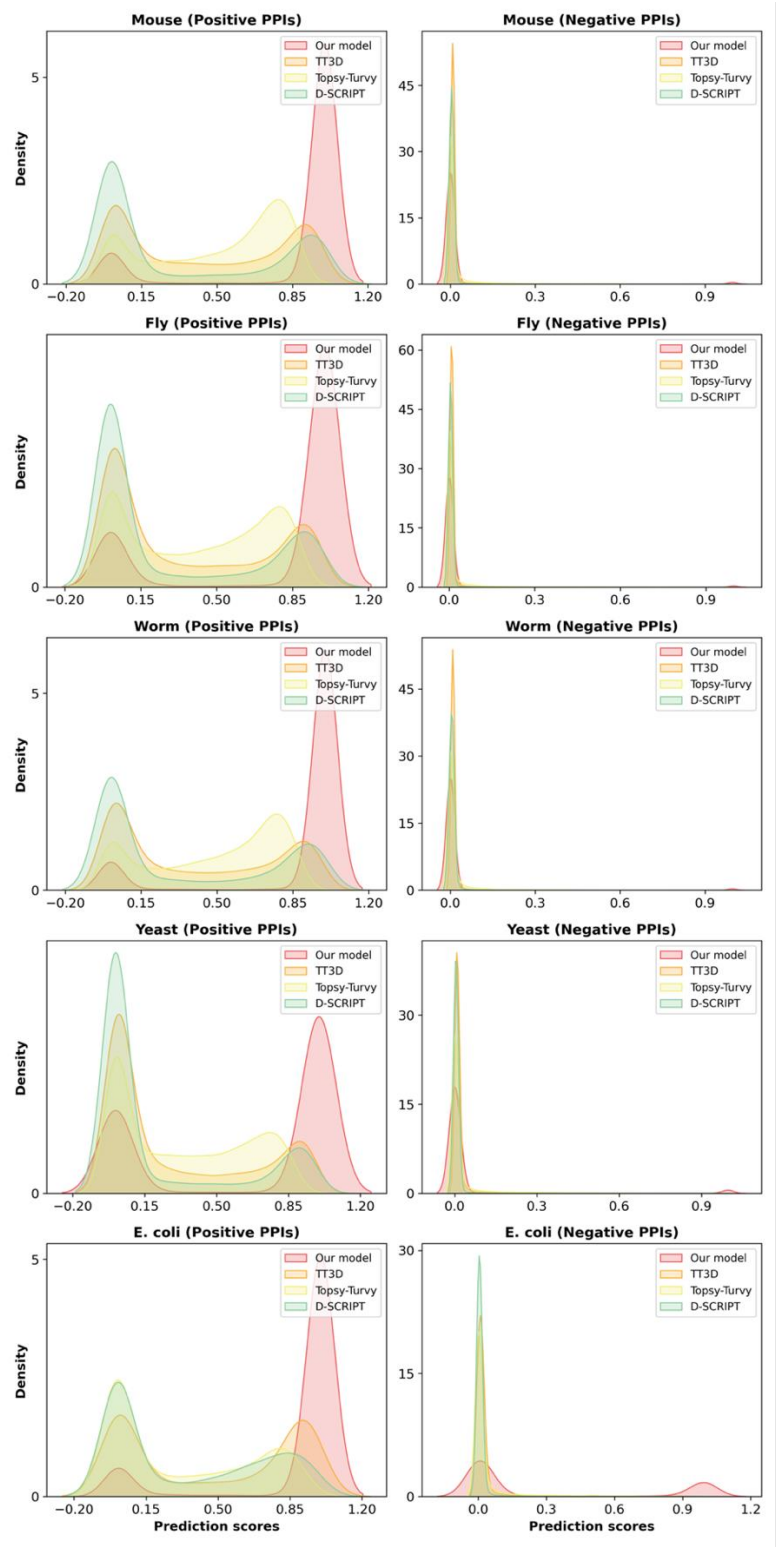


Supplementary Figure 1. The benchmarking of different ratios of mask to classification loss on five species PPI prediction task. **a.** This is the bar plot to show the ratio between mask loss and classification loss. **b** and **c** respectively represent the performance of our model with the different ratios between mask and classification loss on 650M and 35M of ESM-2 models. The left is aligned

with taxonomy tree of the hosts that are used for evaluating our human PPI model. **d** and **e** show the better model with significant improvements than the other, p-value < 0.05 with McNemar's Test.



Supplementary Figure 2. The performance of PLM-interact with the different ratios between mask loss and classification loss. The results show the average accuracy is the highest and perplexity is lowest with a ratio of 1:5 on ESM-2-35M, while the ESM-2-650M performed the best with a ratio of 1:10.



Supplementary Figure 3. The distribution of prediction scores of PPIs of our model, TT3D, TT and D-SCRIPT. Our model identifies the highest number of positive PPIs with prediction scores above 0.5, and the highest number of negative PPIs with scores below 0.5, demonstrated that our model performs best among the recent PPI models, only excluding negative PPIs of *E. coli*.

References

1. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein–protein interactions. *PROTEOMICS* **7**, 2833–2842 (2007).
2. David, A. & Sternberg, M. J. E. The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease. *Journal of Molecular Biology* **427**, 2886–2898 (2015).
3. Vassilev, L. T. *et al.* In Vivo Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2. *Science* **303**, 844–848 (2004).
4. Kotlyar, M., Pastrello, C., Sheahan, N. & Jurisica, I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. (2015) doi:10.1093/nar/gkv1115.
5. Shin, W.-H., Kumazawa, K., Imai, K., Hirokawa, T. & Kihara, D. Current Challenges and Opportunities in Designing Protein–Protein Interaction Targeted Drugs. (2020) doi:10.2147/AABC.S235542.
6. Chen, M. *et al.* Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* **35**, i305–i314 (2019).
7. Hashemifar, S., Neyshabur, B., Khan, A. A. & Xu, J. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* **34**, i802–i810 (2018).
8. Huang, Y., Wuchty, S., Zhou, Y. & Zhang, Z. SGPPI: structure-aware prediction of protein–protein interactions in rigorous conditions with graph convolutional network. *Briefings in Bioinformatics* **24**, bbad020 (2023).
9. Yuan, Q., Chen, J., Zhao, H., Zhou, Y. & Yang, Y. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. *Bioinformatics* **38**, 125–132 (2021).
10. Shen, J. *et al.* Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences* **104**, 4337–4341 (2007).
11. Tsukiyama, S., Hasan, M. M., Fujii, S. & Kurata, H. LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec. *Briefings in Bioinformatics* **22**, bbab228 (2021).
12. Hallee, L. & Gleghorn, J. P. Protein-Protein Interaction Prediction is Achievable with Large Language Models. 2023.06.07.544109 Preprint at https://doi.org/10.1101/2023.06.07.544109 (2023).
13. Sledzieski, S., Devkota, K., Singh, R., Cowen, L. & Berger, B. TT3D: Leveraging precomputed protein 3D sequence models to predict protein–protein interactions. *Bioinformatics* **39**, btad663 (2023).

14. Madan, S., Demina, V., Stapf, M., Ernst, O. & Fröhlich, H. Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings. *Patterns* **3**, 100551 (2022).
15. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems* **12**, 969-982.e6 (2021).
16. Zhang, Z. *et al.* Protein language models learn evolutionary statistics of interacting sequence motifs. 2024.01.30.577970 Preprint at https://doi.org/10.1101/2024.01.30.577970 (2024).
17. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
18. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Preprint at http://arxiv.org/abs/1908.10084 (2019).
19. Singh, R., Devkota, K., Sledzieski, S., Berger, B. & Cowen, L. Topsy-Turvy: integrating a global view into sequence-based PPI prediction. *Bioinformatics* **38**, i264–i272 (2022).
20. Richoux, F., Servantie, C., Borès, C. & Téletchéa, S. Comparing two deep learning sequence-based models for protein-protein interaction prediction. Preprint at https://doi.org/10.48550/arXiv.1901.06268 (2019).
21. Chai Discovery. https://www.chaidiscovery.com/blog/introducing-chai-1.
22. Chai Discovery *et al.* Chai-1: Decoding the molecular interactions of life. Preprint at https://doi.org/10.1101/2024.10.10.615955 (2024).
23. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**, 70–82 (2021).
24. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).
25. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Research* **40**, D841–D846 (2012).
26. Lim, J. *et al.* Opposing effects of polyglutamine expansion on native protein complexes contribute to SCA1. (2008) doi:10.1038/nature06731.
27. Rosen, D. R. *et al.* Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**, 59–62 (1993).

28. Wj, B. *et al.* SOD1A4V-mediated ALS: absence of a closely linked modifier gene and origination in Asia. *PubMed* (2008).

29. C, H. *et al.* Interactome Mapping Provides a Network of Neurodegenerative Disease Proteins and Uncovers Widespread Protein Aggregation in Affected Brains. *PubMed* (2020).

30. Kabuta, T. *et al.* Ubiquitin C-terminal Hydrolase L1 (UCH-L1) Acts as a Novel Potentiator of Cyclin-dependent Kinases to Enhance Cell Proliferation Independently of Its Hydrolase Activity*. *Journal of Biological Chemistry* **288**, 12615–12626 (2013).

31. Grossmann, A. *et al.* Phospho-tyrosine dependent protein–protein interaction network. *Molecular Systems Biology* **11**, 794 (2015).

32. Aten, T. M. *et al.* Tyrosine phosphorylation of the orphan receptor ESDN/DCBLD2 serves as a scaffold for the signaling adaptor CrkL. *FEBS Lett* **587**, 2313–2318 (2013).

33. Daly, R. J., Binder, M. D. & Sutherland, R. L. Overexpression of the Grb2 gene in human breast cancer cell lines. *Oncogene* **9**, 2723–2727 (1994).

34. Ohanian, M. *et al.* Liposomal Grb2 antisense oligodeoxynucleotide (BP1001) in patients with refractory or relapsed haematological malignancies: a single-centre, open-label, dose-escalation, phase 1/1b trial. (2018) doi:10.1016/S2352-3026(18)30021-8.

35. Koshikawa, K. *et al.* Significant up-regulation of a novel gene, CLCP1, in a highly metastatic lung cancer subline as well as in lung cancers in vivo. *Oncogene* **21**, 2822–2828 (2002).

36. Ammari, M. G., Gresham, C. R., McCarthy, F. M. & Nanduri, B. HPIDB 2.0: a curated database for host–pathogen interactions. *Database* **2016**, baw103 (2016).

37. Yang, X., Yang, S., Li, Q., Wuchty, S. & Zhang, Z. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Computational and Structural Biotechnology Journal* **18**, 153–161 (2020).

38. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. 2020.07.12.199554 Preprint at https://doi.org/10.1101/2020.07.12.199554 (2021).

39. Yang, X. *et al.* HVIDB: a comprehensive database for human–virus protein–protein interactions. *Briefings in Bioinformatics* **22**, 832–844 (2021).

40. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* (2023) doi:10.1038/s41587-023-01773-0.

41. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

42. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118**, e2016239118 (2021).

43. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

44. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. Preprint at https://doi.org/10.1101/2024.07.01.600583 (2024).

45. Cornman, A. *et al.* The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling. (2024) doi:10.1101/2024.08.14.607850.

46. Wang, Y. *et al.* LC-PLM: Long-context Protein Language Model. (2024) doi:10.1101/2024.10.29.620988.

47. Babayan, S. A., Orton, R. J. & Streicker, D. G. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science* **362**, 577–580 (2018).

48. Young, F., Rogers, S. & Robertson, D. L. Predicting host taxonomic information from viral genomes: A comparison of feature representations. (2022) doi:10.1371/journal.pcbi.1007894.

49. Liu, D., Young, F., Robertson, D. L. & Yuan, K. Prediction of virus-host associations using protein language models and multiple instance learning. (2023) doi:10.1101/2023.04.07.536023.

50. Roux, S. *et al.* iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. (2023) doi:10.1371/journal.pbio.3002083.

51. Stukalov, A. *et al.* Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature* **594**, 246–252 (2021).

52. J, L. *et al.* Virus-Host Interactome and Proteomic Survey Reveal Potential Virulence Factors Influencing SARS-CoV-2 Pathogenesis. *PubMed* (2021).

53. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607–D613 (2019).

54. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

55. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research* **51**, D638–D646 (2023).

56. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026–1028 (2017).

57. Rep, I. & Čeperić, V. Boosting the Performance of Transformer Architectures for Semantic Textual Similarity. Preprint at https://doi.org/10.48550/arXiv.2306.00708 (2023).

58. Vaswani, A. *et al.* Attention Is All You Need. Preprint at http://arxiv.org/abs/1706.03762 (2023).

59. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

60. Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. in *Proceedings of the 23rd international conference on Machine learning* 233–240 (Association for Computing Machinery, New York, NY, USA, 2006). doi:10.1145/1143844.1143874.

61. Jelinek, F., Mercer, R. L., Bahl, L. R. & Baker, J. K. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* **62**, S63 (2005).

62. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947).