Mechanism-Aware Protein-Protein Interaction Prediction via Contact-Guided Dual Attention on Protein Language Models

Shuchen Deng, Xuanjun Wan, Zichun Mu, Sheng-You Huang*, Chengfei Yan*

School of Physics, Huazhong University of Science and Technology, China

Correspondence: chengfeiyan@hust.edu.cn, huangsy@hust.edu.cn

Abstract

Accurate prediction of protein-protein interactions (PPIs) is essential for understanding cellular mechanisms and disease. However, existing deep learning models often fail to generalize to novel proteins with low sequence similarity to training data. Here, we present a biophysical principle-inspired deep learning framework to enhance generalization by incorporating residue-level interaction modeling. Our model integrates a dual-attention-based PPI prediction module with an inter-protein residue-residue contact predictor based on protein language model-embedded geometric graphs. Joint training on structure-informed PPIs from the Protein Data Bank (PDB) enables the model to predict both interactions and the key residue pairs that mediate these interactions. The model generalizes well to unseen proteins across species in the high-quality interactome dataset (HINT). Fine-tuning on *H. sapiens* PPI data from HINT further improves the performance. Compared to D-SCRIPT, Topsy-Turvy, TT3D, and TUnA, our model, referred to as PLMDA-PPI, achieves superior accuracy and robustness under strict sequence-dissimilarity constraints. Moreover, when evaluating PLMDA-PPI against computationally intensive methods such as AF2Complex, RF2-Lite, and RF2-PPI, PLMDA-PPI delivers significantly better predictive performance while requiring far fewer computational resources.

Introduction

Protein–protein interactions (PPIs) play essential roles in nearly all biological processes. Identifying PPIs involved in these processes is critical for understanding life at the molecular level, investigating disease mechanisms, and facilitating drug discovery[1–4]. However, experimental identification of PPIs is both time-consuming and costly, and currently, interaction data remain highly incomplete for most species. Even for humans, the most extensively studied organism, experimentally validated PPIs represent only a small fraction of the complete interactome[3,5–7]. Moreover, a significant portion of available PPI data originates from high-throughput screening techniques, which are often associated with high false-positive rates[8,9]. Therefore, developing effective computational methods for PPI prediction holds substantial practical importance.

In recent years, the application of deep learning has led to remarkable advances in this field[10–13]. Nevertheless, existing approaches still face critical limitations that hinder their broader applicability and accuracy. For instance, to train deep learning models for protein-protein interaction (PPI) prediction, a large amount of high-quality interaction data is required. Currently, most training data for these models are derived from literature-mined PPI databases such as STRING[14]. However, these databases contain a high proportion of false positives and are biased toward well-studied proteins, leading to biases in both model training and evaluation[15–17]. For

instance, evaluation results by Bernett et al.[15] showed that when sequence similarity between training and test sets is minimized, almost all previously reported PPI prediction models that claimed high predictive performance lose their predictive capability. Further analysis revealed that the previously reported high performance was largely attributed to information leakage caused by random splitting of PPI datasets into training and test sets, where proteins in both sets often exhibited significant overlap or high sequence similarity. Consequently, these models primarily relied on shortcuts, such as sequence similarity between proteins in the training and test sets or node degrees within the training PPI network, rather than truly learning generalizable features. As a result, they fail to generalize to novel proteins that have low sequence similarity to those seen during training.

From the perspective of biological physics, protein–protein interactions arise from physicochemical interactions between interfacial residues that stabilize the complex in its functional conformation (i.e., the native state), representing the global minimum in the free-energy landscape. A predictive model with strong generalization should therefore capture the underlying physicochemical interactions of protein-protein associations. As residue-level interactions predominantly drive PPIs, we hypothesize the model's ability to accurately identify the residue-level interactions reflects its understanding of the biophysical principles governing protein-protein associations. Based on this hypothesis, we developed an inter-protein

residue-residue contact-guided deep learning framework for PPI prediction. Specifically, given a pair of proteins, our model first infers the inter-protein residue-residue contact probability map using protein language model (PLM)-embedded geometric graphs[18], which is then fused with sequence-based PLM representations through a dual-attention network[19,20] to predict the interaction probability of the two proteins.

We first trained our model using structure-informed PPIs collected from Protein Data Bank (PDB)[21], in which the inter-protein contact prediction module and PPI prediction module were jointly trained using a linear combination of the inter-protein contact loss and the PPI loss. Validated on structure-informed PPIs formed by proteins with low sequence similarities to the training set, we show that the model not only identifies PPIs with high accuracy, but also pinpoints key residue pairs mediating these interactions through its inter-protein contact prediction module. We tested the model on high-quality interactome (HINT)[22] data for *Homo sapiens* (*H. sapiens*), *Mus musculus* (*M. musculus*), *Arabidopsis thaliana* (*A. thaliana*), *Drosophila melanogaster* (*D. melanogaster*), and *Schizosaccharomyces pombe* (*S. pombe*), and the result shows that the model trained on the PDB dataset can generalize well to the HINT datasets. To address the discrepancy in data collection standards between PDB and HINT, we further fine-tuned the model using the *H. sapiens* PPIs from HINT. The evaluation results across multiple species show that the fine-

tuning can significantly improve model performance, and that even when sequence similarity between the training and test sets is minimized, the performance remains largely maintained. To provide a comparison, we evaluated four recently developed PPI prediction models including D-SCRIPT[23], Topsy-Turvy[24], TT3D[25], and TUnA[26]. The result shows that our model referred to as PLMDA-PPI consistently outperforms these models by large margins. We further benchmarked PLMDA-PPI against the computationally intensive PPI prediction methods, including AF2Complex[27,28], RF2-Lite[12], and RF2-PPI[29], which are extensions of the groundbreaking protein structure prediction models AlphaFold2[19] and RoseTTaFold2[30]. The results demonstrate that PLMDA-PPI achieves markedly superior predictive performance while requiring substantially fewer computational resources.

Results

1. Overview of PLMDA-PPI

1.1 Biophysical principle underlying protein-protein associations

Thermodynamically, any two proteins could, in principle, come into contact within a cell in the absence of spatiotemporal constraints. However, functional protein-protein associations (i.e., PPIs) typically adopt specific configurations (i.e., the native state) with much lower free energy than non-native states, forming basins in the free-energy landscape. In contrast, non-functional protein-protein collisions (i.e., non-interacting protein pairs)

typically fail to form stable or specific configurations, leading to a free-energy landscape lacking distinct basins. If long-timescale molecular dynamics simulations were performed for both functional protein-protein associations and non-functional protein-protein collisions, and the contact probabilities between their interfacial residues were analyzed, distinct, high-probability contact signals for specific residue pairs would be observed in the former, whereas such signals would not be observed in the latter (Figure 1a). These prominent contact signals correspond to the key residue pairs that stabilize the interaction, which are spatially adjacent residues in the native complex structure.

1.2 The model architecture of PLMDA-PPI

We hypothesize that a PPI prediction model capable of effectively capturing residue-level interaction patterns between proteins has, to some extent, learned the underlying mechanisms of protein-protein associations, thereby exhibiting stronger generalization capability. Specifically, for interacting protein pairs, residue-level interactions correspond to key residue pairs that stabilize the interaction, whereas for non-interacting pairs, no significant residue-level interaction signals are expected to be present. Building upon this hypothesis, we developed PLMDA-PPI for PPI prediction. The model architecture of PLMDA-PPI is shown in Figures 1b~c. Given a pair of proteins, PLMDA-PPI requires sequences, structures (experimental or predicted structures) and multiple sequence alignments

(MSAs) as the inputs. First, following the protocol of our previous work, the protein language model-embedded geometric graphs are constructed based the structures and MSAs of the proteins, from which the protein embeddings were obtained. Then, the pair representation of the two proteins were built through outer concatenation, followed by dimensional hybrid residual network[31] formed by 12 residual blocks to transform the pair representation to the predicted inter-protein contact map, which represents the probabilities of residue–residue contacts between the two proteins. Finally, the predicted inter-protein contact map is used to integrate the protein sequence embeddings (generated from ESM-1b[32]) to generate the PPI score (i.e. the interaction probability of the two proteins) using a dual-attention network modified from Visual Question Answering (VQA) tasks[20,33] (Figure 1c). A more detailed description of PLMDA-PPI can be found in the *Methods* section.

1.3 The training and evaluation protocols of PLMDA-PPI

In Figure 1d, we show the protocols we used to train and evaluate PLMDA-PPI. Specifically, we first trained the model using a non-redundant, structure-informed PPI dataset collected from the PDB. The inter-protein contact prediction module and the PPI prediction module were trained simultaneously using a linear combination of the inter-protein contact loss and the PPI loss. The structure-informed PPI dataset was divided into training and validation sets, ensuring that no protein in the

training set shares more than 40% sequence identity with any protein in the validation set. This constraint was applied to minimize the risk of overfitting caused by data leakage (a more detailed discussion of the data leakage issue can be found in Section 3.1 of the Methods). We did not generate an additional test set for the structure-informed PPIs due to the limited number of non-redundant PPIs available from the PDB. Instead, we used literature-curated, high-quality binary PPIs from HINT for *H. sapiens*, *M. musculus*, *A. thaliana*, *D. melanogaster*, and *S. pombe* to further evaluate the model. Considering the discrepancy in PPI collection criteria between the PDB and HINT, we further fine-tuned the prediction model using *H. sapiens* PPIs from HINT through transfer learning. The *H. sapiens* binary PPIs from HINT were divided into training, validation, and test sets, ensuring that proteins in the three subsets do not share more than 40% sequence identity. Finally, the *H. sapiens* test set, along with the PPI datasets for *M. musculus*, *A. thaliana*, *D. melanogaster*, and *S. pombe* from HINT, were used to evaluate the performance of the fine-tuned models. It is worth noting that for PPI data derived from the PDB database, we extracted input features using experimentally determined monomer structures from the corresponding complex structures during both training and testing. In contrast, for PPI data from the HINT dataset, where experimental monomer structures are not always available, we employed predicted monomer models from AlphaFoldDB[34] as substitutes.

For all datasets used in training, validation, and testing, negative samples (i.e., non-interacting protein pairs) were generated by randomly pairing protein monomers within each dataset, after filtering out protein pairs that share interacting partners. A 10:1 negative-to-positive sample ratio was adopted for each dataset. A more detailed description of our training and evaluation protocols is provided in the *Method*s section.

2. Performance of PLMDA-PPI with initial training on the structure-informed PPI dataset from PDB

2.1 Validation of the model using structure-informed PPIs

Following initial model training, we evaluated the performance of PLMDA-PPI on the validation set derived from structure-informed PPIs in the PDB. The evaluation involved two tasks: (1) PPI prediction as the primary objective, and (2) inter-protein contact prediction for protein pairs forming physical interactions (i.e. positive samples) as an auxiliary task. For comparison, we trained four reference models: D-SCRIPT, Topsy-Turvy, TT3D, and TUnA, using the same dataset. D-SCRIPT, Topsy-Turvy, and TT3D share a common architecture that includes an intermediate module for predicting inter-protein contact maps, which are subsequently summarized into a PPI score. The key differences lie in their input representations and training procedures: D-SCRIPT is purely sequence-based; Topsy-Turvy is also sequence-based but incorporates PPI network-derived index during training; and TT3D extends Topsy-Turvy by

integrating a 21-letter discretized structural alphabet (3Di), derived using Foldseek[35], to provide richer structural representations. PLMDA-PPI differs from D-SCRIPT, Topsy-Turvy, and TT3D in two important aspects in addition to input features. First, they compute the final PPI score using only a customized max-pooling operation over the predicted contact maps. In contrast, PLMDA-PPI employs dual-attention network to fuse the predicted residue-level contact maps with protein language model embeddings, resulting in more informative PPI representations. Second, during training, the three models only use a sparsity-inducing regularization term derived from the predicted contact maps in the loss , but they do not incorporate ground-truth contact maps as supervision. PLMDA-PPI, on the other hand, introduces a multi-task loss that explicitly trains the contact prediction module using inter-protein contact maps calculated based experimental structures as labels. TUnA is a recently developed model that employs a Transformer-based encoder architecture for PPI prediction. Its key innovation lies in the integration of the Spectral-normalized Neural Gaussian Process (SNGP) method for uncertainty estimation, which enables the model to identify out-of-distribution samples and mitigate overfitting to the training data, thereby improving generalization. Unlike PLMDA-PPI, D-SCRIPT, Topsy-Turvy, and TT3D, TUnA does not incorporate an intermediate module for inter-protein contact map prediction.

In Table 1, we present the performance of PLMDA-PPI and the four reference models on the validation set of structure-informed PPIs. The evaluation covers two aspects: PPI prediction and inter-protein contact prediction. For PPI prediction, we report the Area Under the Precision-Recall Curve (AUPRC) and the Area Under the Receiver Operating Characteristic Curve (AUROC). For contact prediction, we evaluate the mean precision of the top 50 predicted residue-residue contacts. Given the 10:1 negative-to-positive sample ratio in the validation set, a null model would achieve an AUPRC of 0.1 and an AUROC of 0.5. Since AUPRC is generally more informative for imbalanced datasets[36], we additionally visualize the Precision-Recall curves for all models in Figure 2a.

As shown in both Table 1 and Figure 2a, PLMDA-PPI shows encouraging performances for PPI prediction with an AUPRC 0.637 and AUROC 0.911, dramatically outperforming the reference models. For inter-protein contact prediction, PLMDA-PPI shows a mean precision of 0.287, although there is still substantial room for improvement, it clearly outperforms the reference models, which exhibit minimal capability in predicting inter-protein contacts (TUnA does not predict inter-protein contacts). It is worth noting that in PLMDA-PPI, the inter-protein contact prediction serves only as an auxiliary task within an intermediate module of the model, primarily designed to enhance the generalization ability of the main task — PPI prediction. Therefore, it is acceptable that its

performance on inter-protein contact prediction does not match that of specialized models dedicated to this task. In Figure 2b, we compare the distributions of the mean contact score of the top 50 predicted residue pairs for interacting and non-interacting protein pairs. Interacting pairs exhibit significantly stronger contact signals, further validating the residue-residue contact signals are employed to guide the PPI prediction (Figure 1c). Finally, Figures 2c and 2d illustrate an example of the experimental inter-protein contact map and the predicted contact map by PLMDA-PPI, respectively. The predicted map clearly identifies key residue-residue contacts that mediate the interaction, demonstrating the model's effectiveness in capturing physically meaningful contact patterns.

2.2 Evaluation of the model using PPIs of multiple species from HINT

We further evaluated performances of PLMDA-PPI and the four reference models on the high-quality literature-curated binary interactions for *H. sapiens, M. musculus, A. thaliana, D. melanogaster, and S. pombe* from HINT database. It should be noted that most PPIs in HINT database do not have structural data available. Therefore, we only evaluated their performances on PPI prediction. In Table 2, we show the model performances in terms of AUPRC and AUROC on the HINT PPI datasets, and in Figures 3a-e, we show the Precision-Recall curves. As can be seen

from the results, PLMDA-PPI consistently outperforms the reference models by large margins (also see Figure S1).

The HINT and PDB databases follow different criteria for collecting the PPI data. For instance, the PDB focuses on interactions with experimentally resolved complex structures, whereas many interactions in HINT—such as transient interactions—are difficult to characterize structurally and thus lack associated structural information. As a result, most PPIs in the HINT database do not have corresponding entries in the PDB. To further evaluate our model's ability—trained on PDB-based data—to predict interactions lacking structural references, we excluded from the HINT dataset all interactions with any PDB reference. For each species, we then generated a new set of negative samples via random pairing at a 10:1 ratio (negatives to positives), and re-assessed the performances of these models using this revised dataset, which were also shown in Table 2 and Figure 3f. As shown in the results, although the predictive performance of PLMDA-PPI declined across datasets other than *A. thaliana*, it still maintained a reasonable level of accuracy. The other four reference models exhibited similar trends. Nevertheless, PLMDA-PPI consistently outperformed these models on the filtered datasets. These findings indicate that models trained solely on PDB-derived data may not generalize well to PPIs whose structures are difficult to resolve experimentally. Therefore, incorporating PPI data without structural

information into model training is necessary to further improve predictive performance.

Additionally, we applied another filtering step by excluding all PPIs in the HINT datasets that involve any protein with more than 40% sequence identity to any protein in the training set. For each species, we re-generated negative samples via random pairing, maintaining a 10:1 negative-to-positive ratio. The resulting filtered datasets were then used to evaluate model performance on novel proteins that were absent from the training set and shared low sequence similarity with training proteins. The results are also shown in Table 2 and Figure 3f. As observed, PLMDA-PPI does not show any systematic performance degradation on the filtered datasets; instead, modest improvements are observed on *H. sapiens, D. melanogaster, and S. pombe*. Similar patterns are seen across the other four reference models. These results suggest that the models do not rely on sequence similarity for interaction prediction. This is consistent with the fact that the PDB and HINT datasets differ substantially in their protein coverage. The performance gains may result from filtering out particularly challenging prediction cases.

3. Performance of PLMDA-PPI after transfer learning on the HINT *H. sapiens* PPI dataset

3.1 Model evaluation on HINT PPI datasets

Given the discrepancy in data collection criteria between PDB and HINT, we adopted a transfer learning strategy to enhance the performance of our models in predicting PPIs under the HINT standard. Specifically, we fine-tuned the models initially trained on PDB data—including PLMDA-PPI and four reference models—using the *H. sapiens* subset from the HINT dataset. To evaluate their performance after fine-tuning, we tested the models on the *H. sapiens* test set, as well as on datasets from *M. musculus*, *A. thaliana*, *D. melanogaster*, and *S. pombe* in HINT. As shown in Table 3 and Figures 4a–e, PLMDA-PPI consistently outperformed the four reference models across all five datasets (also see Figure S2), and transfer learning led to a substantial improvement in its predictive performance compared to the pre-transfer version (Figure 4f). In addition, we analyzed the residue-level contact probability maps of the positive (interacting protein pairs) and negative (non-interacting protein pairs) samples across datasets from different species (Figure S3). Overall, the positive samples exhibit significantly stronger residue contact signals compared to the negative ones. This finding further supports that residue-level contact signals are leveraged to guide the PPI predictions.

3.2 Re-assessing model performance after excluding proteins with high sequence similarity to the training data

In the above evaluations, only the test set of *H. sapiens* PPIs was strictly filtered to exclude any PPIs involving proteins with more than 40%

sequence identity to those in the training set of transfer learning in the beginning of the dataset preparation. For the PPI data from other species, no such sequence similarity-based filtering was applied. Therefore, the model may still rely on sequence similarity to the training data when making predictions on these datasets. To provide a more objective evaluation of the models' performance in predicting interactions between novel proteins that were not present in the training set and shared low sequence similarity with the training proteins, we applied stringent filtering to the HINT datasets. Specifically, we removed all PPIs involving any protein with a sequence identity greater than 40% to any protein in the training set of *H. sapiens* PPIs. Based on the remaining interaction pairs of each species, we re-generated a negative set by randomly pairing proteins at a 10:1 ratio relative to the number of positive samples. The filtered datasets were then used to re-assess the models' performances, which were also shown in Table 3 and Figures 4g-j.

As shown in the table and figure, unlike the previous filtering strategy based on the PDB dataset, all models—including ours and the four reference models—exhibited a certain degree of performance drop on the filtered datasets. This may be attributed to the evolutionary relationships between some PPIs from non-*H. sapiens* species in HINT and those from *H. sapiens*, which could allow models to rely on sequence similarity for prediction. Nevertheless, despite the decline in performance, our model

still maintains considerable predictive power on the filtered datasets. In contrast, the performance of D-SCRIPT, Topsy-Turvy, and TT3D dropped sharply, approaching the level of a null model (AUPRC =0.1)，which is consistent with the findings of Bernett et al. in their previous study[15]. In comparison, although TUnA exhibits a decline in predictive performance, it still maintains a certain level of prediction capability, indicating that it possesses better generalization ability than the other three reference models. However, its predictive performance remains significantly lower than that of our model. These results further demonstrate that our model exhibits significantly superior generalization ability to proteins unseen during training compared to the reference models.

In addition, we also evaluated the model before transfer learning on the same dataset. The results show that the model after transfer learning still significantly outperforms the one without transfer learning in most cases. The only exception is on the *S. pombe* dataset, where the AUPRC slightly decreased (from 0.342 to 0.338); however, the AUROC improved from 0.794 to 0.826 (Supplementary Table S1). These findings suggest that the performance gains from transfer learning are not merely due to the presence of proteins in the test set that share high sequence similarity with those in the training set.

3.3 Re-assessing model performance on the structure-informed PPI dataset

To assess the generalization ability of the models, we evaluated whether the models fine-tuned on the HINT dataset could still maintain their predictive performances on the original PDB-based structure-informed PPI dataset. Accordingly, we further tested the transfer-learned models on the previously used validation set of structure-informed PPIs. In Table 4, we present the performance of the model after transfer learning on this dataset, covering both the primary task of PPI prediction and the auxiliary task of inter-protein contact prediction. As shown in the table, our model significantly outperforms the others on both tasks. Notably, our model even exhibits a slight improvement in PPI prediction performance, whereas the other models suffer a substantial drop in accuracy (Figures 5a-f). For the auxiliary task of contact prediction, our model experiences a moderate performance decline after transfer learning (with the mean precision of the top 50 predicted contacts dropping from 0.287 to 0.230), yet it still maintains considerable predictive power. In contrast, the reference models exhibit virtually no ability to predict residue-level contacts. These results further demonstrate the superior generalization capability of our model compared to the others.

4. Benchmarking PLMDA-PPI against AF2Complex, RF2-Lite, and RF2-PPI

We further conducted a systematic comparison between PLMDA-PPI and several recently developed PPI prediction models derived from the

landmark protein structure prediction frameworks AlphaFold2 and RoseTTAFold2, including AF2Complex, RF2-Lite, and RF2-PPI. Among them, AF2Complex is built upon AlphaFold-Multimer[37] and introduces a specialized interface score (iScore) to evaluate and filter potential PPIs. However, due to the extremely high computational cost of structure prediction with AlphaFold-Multimer, AF2Complex is impractical for large-scale PPI screening tasks. In contrast, RF2-Lite and RF2-PPI are simplified versions of the RoseTTAFold2 model that were optimized to accelerate inference and specifically trained for bacterial and human PPIs, respectively. Nevertheless, compared with PLMDA-PPI, these models still incur substantially higher computational costs during deep learning inference and require generating paired MSAs containing inter-protein co-evolutionary information for each candidate protein pair. This process further increases the overall computational burden, whereas PLMDA-PPI only requires one-time feature generation for individual protein monomers. Consequently, PPI screening based on these models demands far greater computational resources than PLMDA-PPI.

Due to computational resource constraints, in the benchmarking experiments on PPI datasets from five representative species, we randomly selected 50 positive PPIs for each species and generated ten times as many non-interacting protein pairs (i.e., 500) by random pairing as negative samples for testing. Notably, in their original studies, RF2-Lite and RF2-

PPI constructed species-specific proteome/genome databases (for bacteria and humans, respectively) to produce paired MSAs, which are not universally applicable to other organisms. Therefore, in our evaluation, we employed ColabFold[38], a more general-purpose tool, to generate paired MSAs. Moreover, since the training of AF2Complex, RF2-Lite, and RF2-PPI requires substantially larger datasets and computational resources than lightweight models such as PLMDA-PPI, we did not retrain RF2-PPI using the same training set as PLMDA-PPI. Instead, we directly evaluated its publicly available model parameters for fair and practical comparison.

In Figures 6a–e and Table 5, we present the evaluation results of AF2Complex, RF2-Lite, RF2-PPI, and PLMDA-PPI on the subsets of PPIs sampled from HINT datasets. For AF2Complex, we employed two sets of AlphaFold-Multimer model weights recommended in the documentation: model_1_multimer_v3 (using monomer templates) and model_3_multimer_v3 (without monomer templates)—denoted as AF2Complex(1) and AF2Complex(3), respectively. As shown in the figures and table, PLMDA-PPI achieves the best predictive performance in four of the five species. The only exception is *S. pombe*, where PLMDA-PPI outperforms RF2-Lite but performs worse than AF2Complex and RF2-PPI. Figure 6f summarizes the inference time required for each model to predict protein–protein interactions (PPIs), with all tests conducted on a single NVIDIA A6000 GPU. The results show that PLMDA-PPI has an

average inference time of approximately 0.24 seconds per prediction, which is 30–50 times faster than RF2-Lite (average 12.8 seconds) and RF2-PPI (average 7.8 seconds), and 1,000–2,000 times faster than AF2Complex(1) (535.4 seconds) and AF2Complex(3) (285.8 seconds). Moreover, unlike the other three methods that require generating paired multiple sequence alignments (MSAs) for every protein pair, PLMDA-PPI only needs to extract monomer features for each protein once. Therefore, its computational advantage would become even more pronounced in large-scale PPI screening tasks.

Collectively, these results demonstrate that PLMDA-PPI offers substantial advantages in both prediction accuracy and computational efficiency compared to AF2Complex, RF2-Lite, and RF2-PPI. Furthermore, the findings suggest that the inter-protein co-evolutionary information contained in the paired MSAs are not indispensable for accurate PPI prediction.

5. Ablation studies

We further conducted ablation studies to systematically assess the contributions of the key components and training strategies of PLMDA-PPI, covering both the architectural design and training approach. Specifically, we retrained the model after removing major modules using the structure-informed PPI training set derived from PDB and evaluated performance on the corresponding validation set. When we removed the

MSA and structural representations embedded in the geometric graph by the protein language models—retaining only the geometric graph—and replaced the sequence-based protein language model embeddings in the interaction prediction module with geometric node features, the model's performance dropped substantially (AUPRC decreased from 0.637 to 0.229). Restoring the protein representations in the interaction module to language model embeddings led to a significant performance recovery (AUPRC = 0.593), although still inferior to the full model. These results highlight the crucial role of language model representations in our framework. Additionally, replacing the dual-attention network in our model with the customized max pooling module used in the baseline model also led to a notable performance decline (AUPRC = 0.411), underscoring the importance of the dual-attention network module (Fig. 7a).

Moreover, when training the model without the multitask loss for residue-residue contact prediction—i.e., optimizing only for the primary interaction prediction task—the performance on the PDB validation set showed a negligible improvement (AUPRC = 0.643). However, the same model exhibited consistently worse performance across the HINT multi-species datasets. This result indicates that enabling the model to predict inter-protein contacts contributes significantly to its generalization ability (Fig. 7b).

Finally, we trained the model directly on the training set of *H. sapiens* PPIs from HINT without any pretraining on PDB data. Across all species, this directly trained model performed significantly worse than the fine-tuned from the PDB-pretrained model, further demonstrating the importance of pretraining on structure-informed datasets for achieving strong cross-species generalization (Fig. 7c and Supplementary Table S2).

Conclusion

In this study, we sought to enhance the generalization ability of deep learning models for PPI prediction, addressing a key limitation of existing models—their inability to accurately predict novel, unseen proteins that exhibit low sequence similarity to those in the training set. The central hypothesis guiding our work is that a robust, generalizable model should not only predict interactions but also effectively capture residue-level interaction patterns between proteins, which is inspired by the biophysical principle of protein-protein associations. Guided by this hypothesis, we designed a model that integrates a dual-attention PPI prediction module into our prior framework for inter-protein contact prediction based on protein language model–embedded geometric graphs. By co-training both modules with structural PPI information, the model achieves simultaneous prediction of interactions and identification of the residue pairs mediating them.

Our experimental results confirm the validity of this hypothesis. The model, trained on PDB data, successfully predicts PPIs across multiple species from the HINT database. Fine-tuning on PPI data of *H. sapiens* from the HINT database further enhances the model's performance. When compared with other recently developed models such as D-SCRIPT, TOPSY-TURVY, TT3D, and TUnA, our approach not only achieves superior accuracy but also demonstrates remarkable generalization to novel proteins When compared with other recently developed models such as D-SCRIPT, TOPSY-TURVY, TT3D, and TUnA, our approach not only achieves superior accuracy but also demonstrates remarkable generalization to novel proteins with low sequence similarity to those in the training data. Further comparisons with computationally intensive methods, including AF2Complex, RF2-Lite, and RF2-PPI, demonstrate that PLMDA-PPI exhibits substantial advantages in both predictive accuracy and computational efficiency. Compared with these models, PLMDA-PPI relies solely on monomer features without the need to regenerate features for protein pairs, making it more suitable for large-scale interaction screening. In addition, the superior predictive performance of PLMDA-PPI indicates that the inter-protein co-evolutionary information contained in MSAs is not indispensable for accurate PPI prediction.

The impact of this work extends beyond protein–protein interaction (PPI) prediction. By formalizing the integration of structural and non-

structural data in the context of biomolecular interactions, our model offers a novel and adaptable framework for predicting other interaction types, such as protein–small molecule and protein–nucleic acid interactions. These interactions, like PPIs, often involve heterogeneous data—some with structural information, others without. As hypothesized, combining both data types enables deeper insights into the biophysical principles underlying these interactions, ultimately enhancing the generalization capability of predictive models for unseen cases.

While our model substantially outperforms existing approaches in terms of both accuracy and generalizability, there remain opportunities for further enhancement. A key challenge lies in narrowing the performance gap between our inter-protein contact prediction module and advanced end-to-end structure prediction tools including AlphaFold-Multimer[37] and AlphaFold3[39]. Future work could explore integrating AlphaFold's structural modeling capabilities into our dual-attention-based PPI framework to further boost predictive power. However, due to AlphaFold's high computational cost, improving its efficiency will be critical for enabling large-scale interaction screening.

In summary, our study demonstrates that incorporating residue-level interaction mechanisms, guided by our central hypothesis, markedly improves both the accuracy and generalization of PPI prediction. Beyond

advancing PPI prediction, this work lays the foundation for extending such frameworks to a broader range of complex biomolecular interactions.

Methods

1. Dataset preparation

1.1 PPI database selection

The PDB database contains a vast collection of experimentally resolved protein multimer structures, offering direct access to physically interacting protein pairs and their complex structural details. Therefore, we used the PDB database as a source of PPI data with known structures. However, due to the high cost and time-consuming nature of experimental structure determination, the coverage of non-redundant PPIs for a single species in the PDB database remains limited. This necessitates the integration of other high-quality PPI databases to expand the dataset.

Currently, multiple databases store protein interaction information categorized by species. For example, HuRi[3] and BioPlex[4] have systematically screened human PPIs using yeast two-hybrid (Y2H) and affinity purification-mass spectrometry (AP-MS), respectively. However, these high-throughput screening methods often have a high false positive rate, and AP-MS can only detect co-complexed proteins without determining whether direct physical interactions occur. Additionally, computationally curated databases such as STRING[14] and HINT[22] extract PPI data from literature using various mining algorithms and filtering

criteria. Due to differences in these approaches, the number and types of PPIs included in each database vary significantly for the same species. In particular, the *"Physical Interaction"* category of PPIs in the STRING database were widely used for training and testing PPI prediction models, but it does not distinguish between direct interactions and proteins that merely co-occur within a complex[14]. In contrast, the HINT database explicitly differentiates between direct and indirect interactions and provides high-quality PPI datasets that have been confirmed by at least two independent small-scale experimental studies (the high-quality literature-curated datasets). Analysis by Donald Petrey et al.[9] suggests that the high-quality literature-curated binary interaction datasets from the HINT database is the best available source for direct physical protein interactions. Therefore, we also leveraged these datasets for model development and testing in this study.

1.2 Structure-informed PPI collection from PDB

To construct a dataset of structure-informed PPIs, we filtered PDB entries deposited before October 24, 2022, using the following criteria: 1) Experimental Method: X-RAY DIFFRACTION, 2) Data Collection Resolution: ⩽4Å, 3) Entry Polymer Composition: heteromeric protein, 4) Chain count: ≥2. This resulted in 43,983 PDB entries. We then extracted protein-protein pairs from their biological assemblies, yielding a total of 438,554 pairs. To remove redundancy, we clustered all the 877,108 protein

sequences (from 438,554 pairs) using CD-HIT[40] with parameters "*-c 0.4 -n 0.2 -g 1*", grouping sequences at a 40% identity threshold, with which each protein was assigned a unique cluster ID (e.g. cluster 0, cluster 1, …), and each protein-protein pair was labeled with a cluster pair (e.g. cluster 0-cluster 1). The protein-protein pairs belonging to the same cluster pair (note: cluster n-cluster m and cluster m-cluster n were considered as the same pair) were considered as redundant. We computed inter-protein contact maps (defining residue contacts as heavy-atom distances $\leq$8Å) for all the protein-protein pairs, and only kept protein-protein pairs with more than 50 inter-protein contacts as PPIs. Among redundant PPIs (i.e., those in the same cluster pair), we retained only the one with the highest number of contacts. We further filtered these PPIs with the following criteria:1) the chain length of each protein between 30 and 521, 2) $\leqslant$ 20% missing residues in the structure of each chain, 3) removal of self-interacting proteins (i.e., proteins from the same cluster) . After these steps, we obtained 6,189 structure-informed, non-redundant PPIs, comprising 11,119 unique protein monomers from 5952 sequence clusters.

1.3 PPI collection for multiple species from HINT

We obtained high-quality, literature-curated binary PPI datasets for *H. sapiens, M. musculus*, *A. thaliana*, *D. melanogaster*, and *S. pombe* from the HINT database. We applied the following filtering steps: 1) removal of self-interacting proteins, 2) excluding PPIs involving proteins with chain

lengths shorter than 30 or longer than 1024, 3) discarding PPIs formed by protein isoforms, to further refine the datasets. After filtering, the datasets contained 23,311 PPIs for *H. sapiens,* 2,369 PPIs for *M. musculus,* 2,789 PPIs *for A. thaliana,* 1,336 PPIs *for D. melanogaster*, *and* 791 PPIs *for S. pombe*.

2. Model construction of PLMDA-PPI

2.1 Input features

2.1.1 Protein geometric graph

Given the 3D structure of a protein, its geometric graph is constructed following the same protocol as PLMGraph-Inter[18], our previous model for inter-protein contact prediction. Each residue is represented as a node, and edges are formed between residues with a $C\alpha$ distance ≤ 18Å. The graph incorporates geometric features for each node and edge in the graph, including inter-atom distances (scalars) and orientations (vectors) derived from the 3D structure. To ensure SE(3) invariance, the geometric vectors are extracted using local coordinate systems. Further details on the geometric graph representation module can be found in our previous work.

2.1.2 Embeddings of sequence, MSA, and structure

ESM-1b[32] is employed to generate embeddings for protein sequences, and ESM-IF[41] is employed to generate embeddings for protein structures. Jackhmmer[42] is used to search the Uniref90[43] database to construct the multiple sequence alignment (MSA) (using default parameters) for a given

protein sequence. The MSA embeddings are formed by two components: the embeddings generated by ESM-MSA-1b[44] and the position-specific scoring matrices (PSSM) derived from the statistics of multiple sequence alignment.

2.2 Protein encoder

The protocol used PLMGraph-Inter is also used to encode proteins in this study. Specifically, the scalar features of each node in the protein graph are formed by concatenating MSA embeddings, structure embeddings, and scalar geometric features of individual residues. The protein graph's scalar and vector features are then processed through three layers of message passing-based graph neural network formed by geometric vector perceptrons (GVP)[45], which consist of a scalar track and vector track, allowing rotationally invariant transformations of the scalar features and rotationally equivariant transformations of the vector features. Finally, the scalar and vector features of each node are stitched to form a 1D representation of the protein. A more detailed description of the graph transformation module can also be found in our previous work[18]. A key difference between PLMDA-PPI and PLMGraph-Inter is that PLMDA-PPI does not incorporate protein sequence embedding (from ESM-1b) into the protein graph. Instead, the sequence embeddings are moved to the interaction prediction module, as we found that including them in both the protein graphs and the interaction module significantly increased the

likelihood of false positive predictions.

2.3 Inter-protein contact prediction module

The inter-protein contact prediction module takes the protein 1D representation output by the protein encoder for each protein and models the interaction between residues of two proteins. For the two protein 1D representations $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{L_1} \times \boldsymbol{d}}$ and $\boldsymbol{B} \in \mathbb{R}^{\boldsymbol{L_2} \times \boldsymbol{d}}$ ($\boldsymbol{L_1}$ and $\boldsymbol{L_2}$ denote the protein lengths, $\boldsymbol{d}$ equaling to 448 denotes the representation dimension), We perform the following broadcast and concatenation operations (Equation (1)~(3)) to generate 2D representation $\boldsymbol{C} \in \mathbb{R}^{\boldsymbol{2d} \times \boldsymbol{L_1} \times \boldsymbol{L_2}}$ for the protein pair:

$$\boldsymbol{A} = Broadcast(Unsqueeze(\mathbf{A})) \in \mathbb{R}^{\boldsymbol{L_1} \times \boldsymbol{L_2} \times \mathbf{448}} \quad (1)$$

$$\boldsymbol{B} = Broadcast(Unsqueeze(\mathbf{B})) \in \mathbb{R}^{\boldsymbol{L_1} \times \boldsymbol{L_2} \times \mathbf{448}} \quad (2)$$

$$\boldsymbol{C} = Transpose(Concat(\mathbf{A}, \mathbf{B})) \in \mathbb{R}^{\mathbf{896} \times \boldsymbol{L_1} \times \boldsymbol{L_2}} \quad (3)$$

The 2D representation of protein pair are then transformed by dimensional hybrid residual block to generate the predicted inter-protein contact map. Specifically, we first use a convolution layer with kernel size of 1*1 to reduce the number of channels of C from 896 to 96, which are then transformed successively by 12 dimensional hybrid residual blocks[31,46], and another convolution layer with kernel size of 1*1 for channel reduction (from 96 to 1). Finally, the predicted inter-protein contact map is obtained through a sigmoid transformation of the feature map.

2.4 Interaction prediction module

PPIs are mediated by physicochemical interactions between interfacial residues. Compared to non-interacting protein pairs, the predicted inter-protein contact maps of interacting proteins should exhibit more significant contact signals. Therefore, the inter-protein contact map generated by the previous module is used to integrate the single sequence-base protein language model embeddings (using ESM-1b) to the protein interaction prediction module. Here, we use only the single sequence-based protein language model embeddings and exclude the representations used to generate the inter-protein contact maps, as our in-house tests indicated that directly incorporating these representations significantly increases the false positive rate. One possible reason is that PPIs exhibit a high degree of sequence specificity. Introducing explicit structural information and multiple sequence alignment-based representations may reduce the model's sensitivity to sequence features. For example, the model may misidentify protein pairs with similar monomeric folding patterns to known interacting proteins as interacting, even when they do not actually interact.

We modify the Dual Attention Network for multimodal reasoning (r-DAN) proposed by Nam et al.[33] to model PPI prediction as a Visual Question Answering (VQA) task, in which the protein sequence embeddings are used to replace the visual and textual features in VQA, and the predicted inter-protein contact map outputted from previous module is

leveraged as a weight matrix to calculate the information transferred between residues of the two proteins. It should be noted that the Dual Attention Networks were also previously used by Li et al. to build the compound-protein binding affinity prediction module in MONN[20]. The architecture details of our interaction prediction module are described as follows:

First, a linear transformation layer is applied to reduce the dimension of the sequence embeddings generated by ESM-1b from 1280 to 448 ($\boldsymbol{P_1} \in \mathbb{R}^{L_1 \times 448}$, $\boldsymbol{P_2} \in \mathbb{R}^{L_2 \times 448}$). Then two feed-forward network (FNN) blocks were successively applied to generate the protein context vectors ($\boldsymbol{p_1^{(k)}} \in \mathbb{R}^{1 \times 448}$, $\boldsymbol{p_1^{(k)}} \in \mathbb{R}^{1 \times 448}$) and update the memory vector ($\boldsymbol{m_p^{(k)}} \in \mathbb{R}^{1 \times 448}$) based on the protein embeddings ($\boldsymbol{P_1}, \boldsymbol{P_2}$) the memory vector ($\boldsymbol{m_p^{(k-1)}}$) in the previous step, and the predicted inter-protein contact map ($\boldsymbol{Cmap}$) through Equation (4)~(5):

$$\boldsymbol{p_1^{(k)}}, \boldsymbol{p_2^{(k)}} = P_\text{attention}(\boldsymbol{P_1}, \boldsymbol{P_2}, \boldsymbol{m_p^{(k-1)}}, \boldsymbol{Cmap}) \quad (4)$$

$$\boldsymbol{m_p^{(k)}} = Layernorm\left(GRUCell\left(\boldsymbol{m_p^{(k-1)}}, \boldsymbol{p_1^{(k)}} \odot \boldsymbol{p_2^{(k)}}\right)\right) \quad (5)$$

where *GRUCell* is the computational unit that implements the GRU mechanism for processing sequential data in recurrent neural networks, $Layernorm$ represents Layer Normalization, and *P_attention* represents Protein Attention, designed to generate protein context vectors by focusing on specific regions in the protein embeddings. Details of this block was

described in the following section.

The Protein Attention block: First, two intermediate hidden states ($\boldsymbol{h}_{\boldsymbol{p_1}}^{(k)} \in \mathbb{R}^{L_2 \times 448}$, $\boldsymbol{h}_{\boldsymbol{p_2}}^{(k)} \in \mathbb{R}^{L_1 \times 448}$) correspond to information transferred from each protein to its partner are calculated using the predicted inter-protein contact map ($\boldsymbol{Cmap} \in \mathbb{R}^{L_1 \times L_2}$) as the weight matrix through Equation (6)~(7):

$$\boldsymbol{h}_{\boldsymbol{p_1}}^{(k)} = Layernorm\left(\boldsymbol{Cmap}^{\boldsymbol{T}} \cdot \left(\boldsymbol{P_1}\boldsymbol{W}_{\boldsymbol{p_1}}^{(k)}\right)\right) \quad (6)$$

$$\boldsymbol{h}_{\boldsymbol{p_2}}^{(k)} = Layernorm\left(\boldsymbol{Cmap} \cdot \left(\boldsymbol{P_2}\boldsymbol{W}_{\boldsymbol{p_2}}^{(k)}\right)\right) \quad (7)$$

where $\boldsymbol{W}_{\boldsymbol{p_1}}^{(k)}$, $\boldsymbol{W}_{\boldsymbol{p_2}}^{(k)}$ are the network parameters.

Then two protein attention weights ($\boldsymbol{\alpha}_{\boldsymbol{p_1}}^{(k)} \in \mathbb{R}^{L_1 \times 448}$, $\boldsymbol{\alpha}_{\boldsymbol{p_2}}^{(k)} \in \mathbb{R}^{L_2 \times 448}$) are calculated based on $\boldsymbol{m}_{\boldsymbol{p}}^{(k-1)}$, $\boldsymbol{P_1}$, $\boldsymbol{P_2}$, $\boldsymbol{h}_{\boldsymbol{p_1}}^{(k)}$, $\boldsymbol{h}_{\boldsymbol{p_2}}^{(k)}$ through Equation (8)~(9):

$$\boldsymbol{\alpha}_{\boldsymbol{p_1}}^{(k)} = Sigmoid\left(\left(Layernorm\left(\left(\boldsymbol{P_1}\boldsymbol{W}_{\boldsymbol{p_1,tmp}}^{(k)}\right) \odot \left(\boldsymbol{m}_{\boldsymbol{p}}^{(k-1)}\boldsymbol{W}_{\boldsymbol{p,m}}^{(k)}\right)\right) \boldsymbol{W}_{\boldsymbol{tmp_{p_1}}}^{(k)} \odot \boldsymbol{h}_{\boldsymbol{p_2}}^{k}\right)\right) \quad (8)$$

$$\boldsymbol{\alpha}_{\boldsymbol{p_2}}^{(k)} = Sigmoid\left(\left(Layernorm\left(\left(\boldsymbol{P_2}\boldsymbol{W}_{\boldsymbol{p_2,tmp}}^{(k)}\right) \odot \left(\boldsymbol{m}_{\boldsymbol{p}}^{(k-1)}\boldsymbol{W}_{\boldsymbol{p,m}}^{(k)}\right)\right) \boldsymbol{W}_{\boldsymbol{tmp_{p_2}}}^{(k)} \odot \boldsymbol{h}_{\boldsymbol{p_1}}^{k}\right)\right) \quad (9)$$

where $\boldsymbol{W}_{\boldsymbol{tmp_{p_1}}}^{(k)}$, $\boldsymbol{W}_{\boldsymbol{tmp_{p_2}}}^{(k)}$, $\boldsymbol{W}_{\boldsymbol{p_1,tmp}}^{(k)}$, $\boldsymbol{W}_{\boldsymbol{p_2,tmp}}^{(k)}$, $\boldsymbol{W}_{\boldsymbol{p,m}}^{(k)}$ are the network parameters, and $Sigmoid$ represents the Sigmoid Activation.

Finally, the protein context vectors are obtained through first

elementwise multiplication of the protein embeddings and the attention weights, followed by summation and average over protein sequences and layer normalization through Equation (10)~(11):

$$\boldsymbol{p}_{\mathbf{1}}^{(k)} = Layernorm\left(\overline{\boldsymbol{P}_{\mathbf{1}} \odot \boldsymbol{\alpha}_{\boldsymbol{p}_{\mathbf{1}}}^{(k)}}\right) \quad (10)$$

$$\boldsymbol{p}_{\mathbf{2}}^{(k)} = Layernorm\left(\overline{\boldsymbol{P}_{\mathbf{2}} \odot \boldsymbol{\alpha}_{\boldsymbol{p}_{\mathbf{1}}}^{(k)}}\right) \quad (11)$$

Memory vector initialization and interaction probability calculation:

The initial memory vector $\boldsymbol{m}_{\boldsymbol{p}}^{(0)}$ is calculated through Equation (12):

$$\boldsymbol{m}_{\boldsymbol{p}}^{(0)} = Layernorm(\overline{\boldsymbol{P}_{\mathbf{1}}} \odot \overline{\boldsymbol{P}_{\mathbf{2}}}) \quad (12)$$

The final predicted interaction probability between two proteins is calculated through Equation (13):

$$\textbf{PPI Score} = Sigmoid\left(Layernorm\left(\boldsymbol{p}_{\mathbf{1}}^{(2)} \cdot \boldsymbol{p}_{\mathbf{2}}^{(2)}\right)\right) \quad (13)$$

3. Training of PLMDA-PPI

3.1 The data leakage issue in PPI prediction

Previous studies often randomly split PPI datasets into training set and test set. Before splitting the dataset, PPIs are clustered to reduce redundancy. Specifically, if both proteins in a PPI exhibit sequence similarities exceeding a predefined threshold (typically 40%) to their respective counterparts in another PPI, the two interactions are considered redundant. In such cases, only one representative PPI is retained to ensure data diversity and minimize redundancy. This approach can mitigate data leakage to some extent but cannot completely eliminate it. For example, although protein pairs in the test set do not share sequence similarity with

those in the training set, individual proteins from a test pair may still appear in the training data. Park et al. proposed three classes of test sets：C1 (both proteins in test protein pairs occur in training)，C2 (only one protein in test protein pairs occur in training), and C3 (no protein overlap between training and test), and found the PPI prediction performances dropped significantly between C1 and C2 as well as C2 and C3[16]. Bernett et al.'s analysis further reveals that the overlap between training and test sets causes models to rely on sequence similarity and node degrees in the network as shortcuts for prediction, leading to severe overfitting[15]. When they minimize sequence similarity between the training and test sets to prevent data leakage, the performance of many PPI prediction models—previously reported as highly effective—drops to nearly random levels. Therefore, careful handling of data leakage is essential when preparing datasets for model training and evaluation.

3.2 Initial training using structure-informed PPIs from PDB

3.2.1 Datasets for training and validation

We randomly divided the 5,952 sequence clusters, formed by protein monomers in the structure-informed PPI dataset, into two groups: Group 1 with 5,500 clusters and Group 2 with 452 clusters. Positive samples in the training set consisted of 5,352 PPIs formed between proteins within Group I. The validation set included 211 PPIs formed within Group II . Additionally, 626 PPIs involving proteins from both groups were discarded

to avoid data leakage. To train the PPI prediction models, we also generated negative samples (non-interacting protein pairs). Specifically, protein monomers within the training and validation sets were randomly paired to create candidate negative samples. We then filtered out any pairs that shared common interacting partners, as these may represent true interactions. For both training and validation sets, we adopted a 10:1 negative-to-positive ratio, following previous studies to reflect the sparsity of true PPIs among protein pairs. Notably, the training and validation sets do not share proteins with sequence similarity higher than 40%, thereby minimizing the risk of overfitting due to data leakage. For the structure-informed PPI datasets, we extracted input features using experimentally determined monomer structures from the corresponding complex structures.

3.2.2 Training protocol

We employed a loss function formed by a linear combination of the inter-protein contact loss and the PPI loss calculated through Equation (14) to train the model:

$$\boldsymbol{Loss_{total}} = \boldsymbol{\mu} \cdot \boldsymbol{Loss_{contact}} + \boldsymbol{v} \cdot \boldsymbol{Loss_{interaction}} \quad (14)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{v}$ are the combination weights, and $\boldsymbol{\mu}$ is chosen as 100 and $\boldsymbol{v}$ is chosen as 1.

$\boldsymbol{Loss_{contact}}$ is the inter-protein contact loss calculated through

Equation (15):

$$\boldsymbol{Loss_{contact}} = -\frac{\mathbf{1}}{\boldsymbol{L_1 * L_2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{L_1}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{L_2}}\left[\hat{\boldsymbol{x}}_{\boldsymbol{ij}}\,\boldsymbol{log}\,\boldsymbol{x_{ij}} + \left(\mathbf{1} - \hat{\boldsymbol{x}}_{\boldsymbol{ij}}\right)\boldsymbol{log}\left(\mathbf{1} - \boldsymbol{x_{ij}}\right)\right]$$

(15)

where $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$ are the protein lengths, $\boldsymbol{x_{ij}}$ represents the contact probability between residue $\boldsymbol{i}$ and $\boldsymbol{j}$ in the predicted contact map, and $\hat{\boldsymbol{x}}_{\boldsymbol{ij}}$ is the contact label (set as 1 if contact, and 0 if not) in the true contact map. For positive samples (i.e. interacting proteins pairs), the true contact maps are computed based on the heavy-atom-distance between residues in the experimental structures using 8Å as the threshold , and for negative samples (i.e. non-interacting proteins pairs), all residue pairs are considered as non-contact (i.e. $\hat{\boldsymbol{x}}_{\boldsymbol{ij}}$ is always set as 0).

$\boldsymbol{Loss_{interaction}}$ is the interaction loss calculated through Equation (16):

$$\boldsymbol{Loss_{interaction}} = -[\boldsymbol{y}\,\boldsymbol{log}\,\hat{\boldsymbol{y}} + (\mathbf{1} - \boldsymbol{y})\,\boldsymbol{log}(\mathbf{1} - \hat{\boldsymbol{y}})] \quad (16)$$

where $\boldsymbol{y}$ represents the interaction label (set as 1 for positive samples, and 0 for negative samples), and $\hat{\boldsymbol{y}}$ represents the predicted interaction probability.

The model training was conducted using the PyTorch (v2.1.0) on a single NVIDIA A100 GPU. Specifically, the initial learning rate was set as 1×10^{-5}, the AdamW optimizer was used to update the model weights, and the batch size was set to 1 . After each epoch, we calculated the total

loss on the validation set, and if the total loss on the validation set did not decrease over 3 epochs, the learning rate was reduced by a factor of 0.1 in the next epoch. After two learning rate reductions, we selected the model weights with the lowest loss on the validation set as the final model weights.

3.3 Transfer learning using *H. sapiens* PPIs from HINT database

3.3.1 Datasets for training and validation

To further enhance the model, the *H. sapiens* PPI dataset from the HINT database was used for fine-tuning, initialized with the weights obtained from training on the structure-informed PPI dataset. To minimize the possibility of data leakage in dataset splitting, following the same procedure as previously used in the structure-informed PPI dataset, we first clustered protein monomers in the 23,311 PPIs with CD-HIT using CD-HIT with parameters "-c 0.4 -n 0.2 -g 1", grouping sequences at a 40% identity threshold, which yielded 5873 clusters. The 5873 clusters were then randomly categorized into three groups: group I containing 3800 clusters, group II containing 700 clusters, and group III containing 1373 clusters. We retained only the PPIs formed by proteins within each group and discarded the rest. In addition, to reduce redundancy, for PPIs belonging to the same cluster pair, we randomly selected one representative interaction. As a result, we obtained 4,585 PPIs formed by proteins within group I, 345 PPIs within group II, and 1,502 PPIs within group III. These PPIs were used as the positive samples for the training, validation, and test

sets during model fine-tuning. To generate negative samples for each set, we also followed the previous procedure. Specifically, the negative samples were generated through random pairing protein monomers within each group after filtering out protein pairs sharing interacting partners, and a 10:1 negative-to-positive ratio was adopted for each set. For the HINT PPI datasets, where experimental monomer structures are not always available, we employed predicted monomer models from AlphaFoldDB as substitutes to generate the input features.

3.3.2 Training protocol

The training set prepared above was used to further update the model weights initialized with the weights obtained from training on the structure-informed PPI dataset. Since PPIs from HINT database generally do not have their complex structure information available, we only employed the interaction labels to update the model weights following the $\boldsymbol{Loss_{interaction}}$ defined in Equation (16) as the loss function. To preserve the model's ability to infer inter-protein contact information, the inter-protein contact maps of the structure-informed PPIs were incorporated as constraints to the inter-protein contact prediction module in the training using $\boldsymbol{\mu \cdot Loss_{contact}}$ as the loss function, where $\boldsymbol{Loss_{contact}}$ was the defined in Equation (15), and $\boldsymbol{\mu}$ was still set as 100. It is worth noting that during the fine-tuning stage, the structure-informed PPIs were not used to train the interaction prediction module. Still the training was conducted

using the PyTorch (v2.1.0) on a single NVIDIA A100 GPU. The initial learning rate was set as 1×10^{-5}, the AdamW optimizer was used to update the model weights, and the batch size was set to 1 . After each epoch, we calculated the total loss using $\boldsymbol{Loss_{interaction}}$ as the loss function on the validation set, and if the total loss on the validation set did not decrease over 3 epochs, the learning rate was reduced by a factor of 0.1 in the next epoch. After two learning rate reductions, we selected the model weights with the lowest loss on the validation set as the final model weights. It should be noted that the inter-protein contact maps were incorporated solely during training to guide the inter-protein contact prediction module, without being involved in model validation or selection.

4. Training of D-SCRIPT, Topsy-Turvy, TT3D, and TUnA

The training scripts for D-SCRIPT, Topsy-Turvy, and TT3D were downloaded from https://dscript.csail.mit.edu/. Both the initial training and the fine-tuning of these models were implemented under the default settings provided in the official tutorial (https://d-script.readthedocs.io/en/stable/usage.html#training) using the same datasets as the training of PLMDA-PPI. The training scripts for TUnA were downloaded from https://github.com/Wang-lab-UCSD/TUnA. The initial training and fine-tuning of the model were also implemented under the default settings using the same datasets as the training of PLMDA-PPI.

5. Implementation of AF2Complex, RF2-Lite, and RF2-PPI

AF2Complex was obtained from https://github.com/FreshAirTonight/af2complex and executed using the settings provided in example1.sh. RF2-Lite and RF2-PPI were obtained from https://github.com/SNU-CSSB/RF2-Lite and https://github.com/CongLabCode/RoseTTAFold2-PPI, respectively, and implemented under default settings using paired MSAs generated by ColabFold (https://github.com/sokrypton/ColabFold).

Data Availability

The training and test data for this work can be accessed at https://github.com/ChengfeiYan/PLMDA-PPI/tree/main/data.

Code Availability

The code for training and implementing PLMDA-PPI can be accessed at https://github.com/ChengfeiYan/PLMDA-PPI.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (grant number: 32571438 and 32101001) and the Major project of Guangzhou National Laboratory (grant number: GZNL2023A03007). Computational resources were provided by the HPC platform of Huazhong University of Science and Technology. The authors also utilized ChatGPT to assist in linguistic refinement and to enhance the clarity and readability of this manuscript.

Author Contributions

C.Y. and S.-Y.H. conceived the project. C.Y. designed the experiments and supervised the study. S.D. and X.W. performed the experiments. S.D., M.Z., and C.Y. analyzed the data. D.S., X.W., and C.Y. drafted the manuscript. C.Y. and S.-Y.H. revised the manuscript. All authors discussed the results and reviewed the manuscript.

Ethics declarations

The authors declare no competing interests.

References

1. Stelzl, U. *et al.* A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **122,** 975-68 (2005).
2. Mintseris, J. & Weng, Z. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10930-5 (2005).
3. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402-8 (2020).
4. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-40 (2015).
5. Drew, K., Wallingford, J. B. & Marcotte, E. M. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol. Syst. Biol.* **17**, e10016 (2021).
6. Burke, D. F. *et al.* Towards a structurally resolved human protein

interaction network. *Nat. Struct. Mol. Biol.* **30**, 216–5 (2023).

7. Vidal, M. How much of the human protein interactome remains to be mapped? *Science Signaling* **9**, eg7 (2016).
8. Snider, J. *et al.* Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.* **11**, 848 (2015).
9. Petrey, D., Zhao, H., Trudeau, S. J., Murray, D. & Honig, B. PrePPI: A Structure Informed Proteome-wide Database of Protein–Protein Interactions. *J. Mol. Biol.* **435**, 168052 (2023).
10. Zhang, J., Durham, J. & Qian Cong. Revolutionizing protein–protein interaction prediction with deep learning. *Curr. Opin. Struct. Biol.* **85**, 102775 (2024).
11. Humphreys, I. *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
12. Humphreys, I. R. *et al.* Protein interactions in human pathogens revealed through deep learning. *Nat. Microbiol.* **9**, 2642–2652 (2024).
13. Hu, L., Wang, X., Huang, Y. A., Hu, P. & You, Z. H. A survey on computational models for predicting protein-protein interactions. *Brief. Bioinform.* **22,** bbab036 (2021).
14. Szklarczyk, D. *et al.* The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605-12

(2021).

15. Bernett, J., Blumenthal, D. B. & List, M. Cracking the black box of deep sequence-based protein-protein interaction prediction. *Brief. Bioinform.* **25**, bbae076 (2024).

16. Park, Y. & Marcotte, E. M. Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods* **9,** 1134-6 (2012).

17. Lannelongue, L. & Inouye, M. Pitfalls of machine learning models for protein–protein interaction networks. *Bioinformatics* **40**, btae012 (2024).

18. Si, Y. & Yan, C. Protein language model-embedded geometric graphs power inter-protein contact prediction. *eLife* **12,** RP92184 (2023).

19. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-9 (2021).

20. Li, S. *et al.* MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities. *Cell Syst.* **10**, 308-22(2020).

21. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

22. Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 1-12 (2012).

23. 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 Syst.* **12**, 969-82 (2021).

24. Singh, R., Devkota, K., Sledzieski, S., Berger, B. & Cowen, L. Topsy-Turvy: Integrating a global view into sequence-based PPI prediction. *Bioinformatics* **38**, i264-72 (2022).

25. 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).

26. Ko, Y. S., Parkinson, J., Liu, C. & Wang, W. TUnA: an uncertainty-aware transformer model for sequence-based protein–protein interaction prediction. *Brief. Bioinform.* **25**, bbae359 (2024).

27. Gao, M. & Skolnick, J. Improved deep learning prediction of antigen–antibody interactions. *Proc. Natl. Acad. Sci.* **121**, e2410529121 (2024).

28. Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022).

29. Zhang, J. *et al.* Predicting protein-protein interactions in the human proteome. *Science* **390**, eadt1630 (2025).

30. Baek, M. *et al.* Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv* 2023.05.24.542179 (2023)

31. Si, Y. & Yan, C. Improved protein contact prediction using dimensional hybrid residual networks and singularity enhanced loss function. *Brief. Bioinform.* **22**, bbad039 (2021).

32. 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).

33. Nam, H., Ha, J. W. & Kim, J. Dual attention networks for multimodal reasoning and matching. in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* vols 2017-January (2017).

34. Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439-44 (2022).

35. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243-6 (2024).

36. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432 (2015).

37. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034 (2021).

38. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

39. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

40. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-9 (2006).

41. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. *Proceedings of the 39th International Conference on Machine Learning, PMLR 162:8946-8970* (2022).

42. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 1-8 (2010).

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

44. Rao, R. *et al.* MSA Transformer. *Proceedings of the 38th International Conference on Machine Learning, PMLR* **139**:8844-8856 (2021).

45. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. & Dror, R. O. Learning from Protein Structure with Geometric Vector Perceptrons. in *ICLR 2021 - 9th International Conference on*

Learning Representations (2021).

46. Si, Y. & Yan, C. Improved inter-protein contact prediction using dimensional hybrid residual networks and protein language models. *Brief. Bioinform.* **24**, bbad039 (2023).

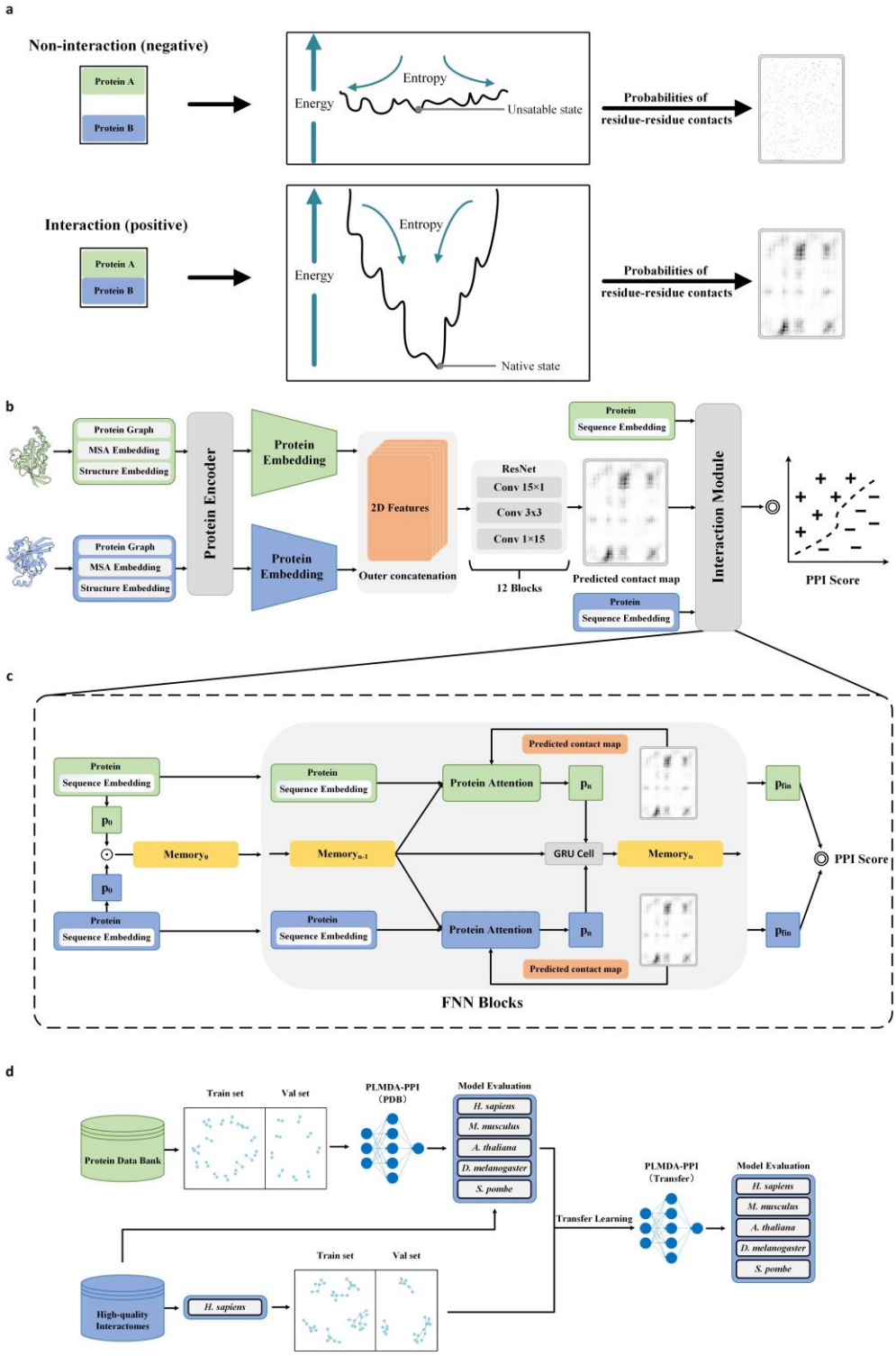


Figure 1. The theoretical basis, model architecture, and training protocol of PLMDA-PPI. (a) Schematic illustration of the biophysical principles underlying functional protein-protein associations (i.e. interaction) and non-functional collisions (i.e. non-interaction). (b) The overall model architecture of PLMDA-PPI. (c) The interaction prediction module of PLMDA-PPI. (d) The workflow for training and evaluation of PLMDA-PPI.

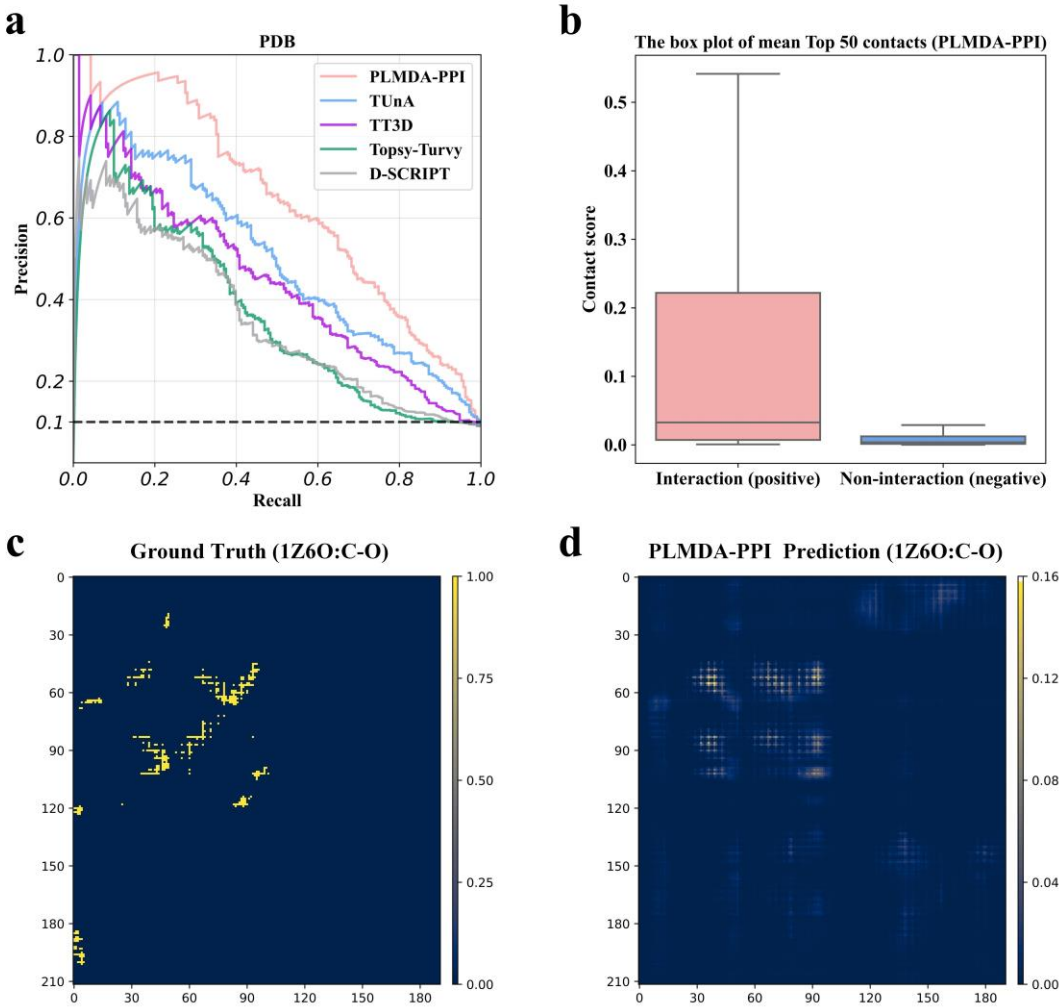


Figure 2. Model validation after initial training on the structure-informed PPI dataset from PDB. (a) The Precision-Recall curves of PLMDA-PPI, D-SCRIPT, Topsy-Turvy, TT3D, and TUnA on the validation set of structure-informed PPIs. (b) Comparison of the distribution of the mean contact score of the top 50 predicted residue pairs for interacting and non-interacting protein pairs. (c)~(d) An example (PDB 1Z6O:C-O) of the experimental inter-protein contact map (c) and the predicted contact map by PLMDA-PPI (d).

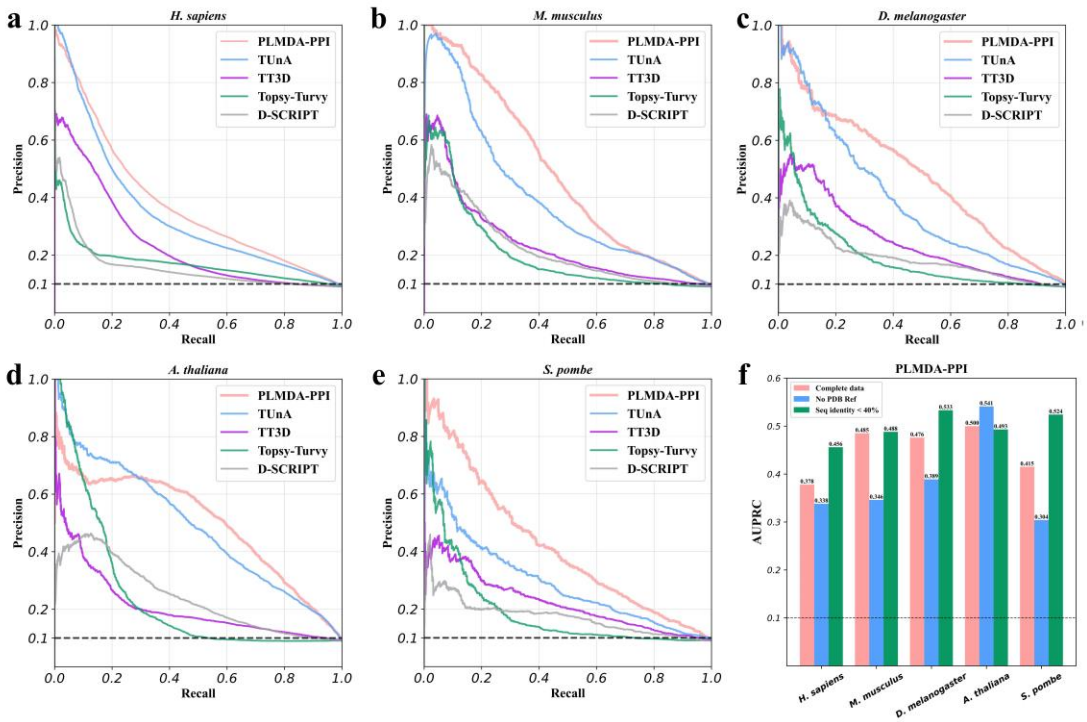


Figure 3. Performances of PLMDA-PPI, D-SCRIPT, Topsy-Turvy, TT3D, and TUnA on the HINT datasets of multiple species after initial training on the structure-informed PPI dataset from PDB. (a)~(e) The Precision-Recall curves on *H. sapiens* (a), *M. musculus* (b), *D. melanogaster* (c), *A. thaliana* (d), and *S. pombe* (e). (f) Performance variations of PLMDA-PPI (in terms of AUPRC) on the HINT datasets after using different criteria to filter the datasets. Blue bars denote the complete HINT datasets, red bars denote the datasets after removing PPIs with PDB references and green bars denote the datasets after excluding all PPIs involving proteins with >40% sequence identity to any protein in the training set.

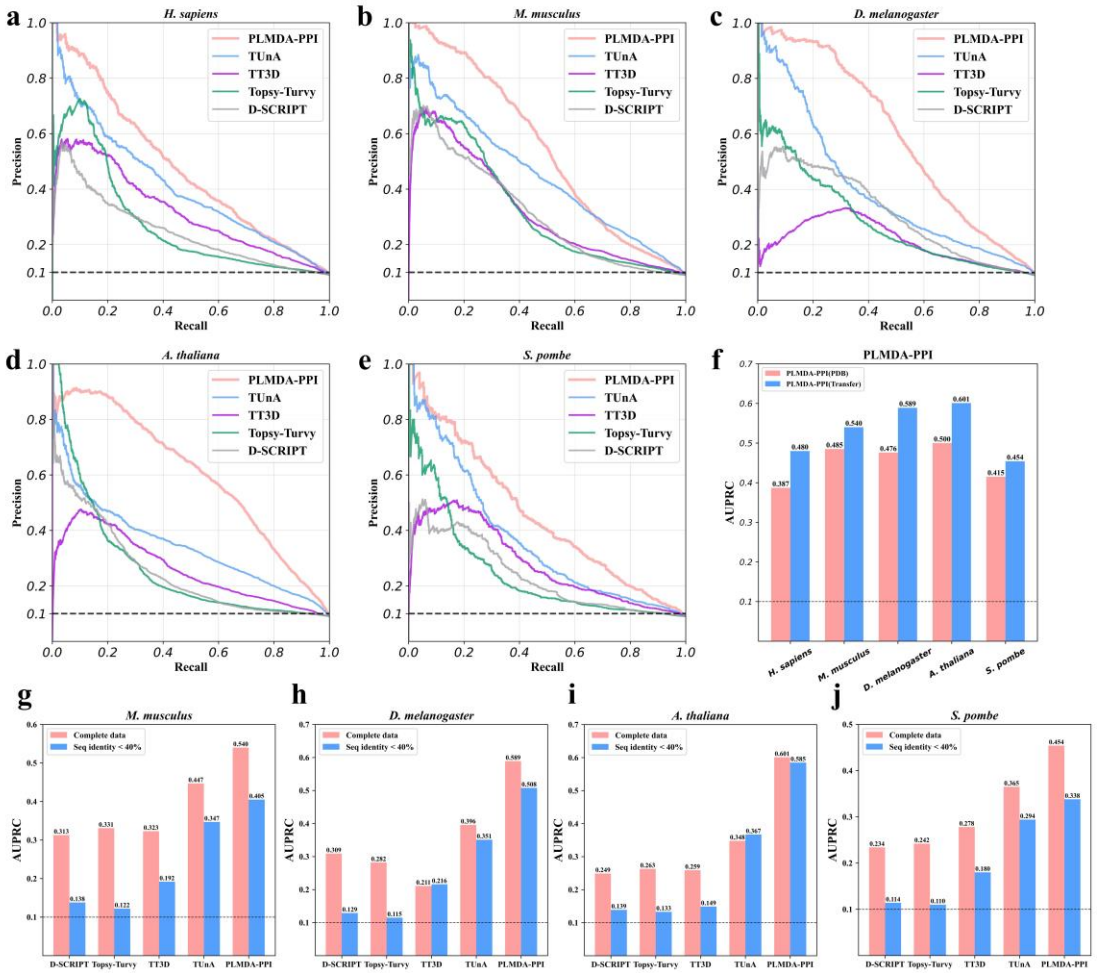


Figure 4. Performances of PLMDA-PPI, D-SCRIPT, Topsy-Turvy, TT3D, and TUnA on the HINT PPI datasets of multiple species after transfer learning on the HINT *H. sapiens* PPI dataset. (a)~(e) The Precision-Recall curves on *H. sapiens* (test set) (a), *M. musculus* (b), *D. melanogaster* (c), *A. thaliana* (d), and *S. pombe* (e). (f) Performance comparison of PLMDA-PPI (in terms of AUPRC) on the HINT datasets before (blue bars) and after transfer learning (red bars). (g)~(j) Performance variations (in terms of AUPRC) of PLMDA-PPI, D-SCRIPT, Topsy-Turvy, TT3D, and TUnA on the HINT PPI datasets of *M. musculus* (g), *D. melanogaster* (h), *A. thaliana* (i), and *S. pombe* (j) after excluding all PPIs involving proteins with >40% sequence identity to any protein in the *H. sapiens* training set.

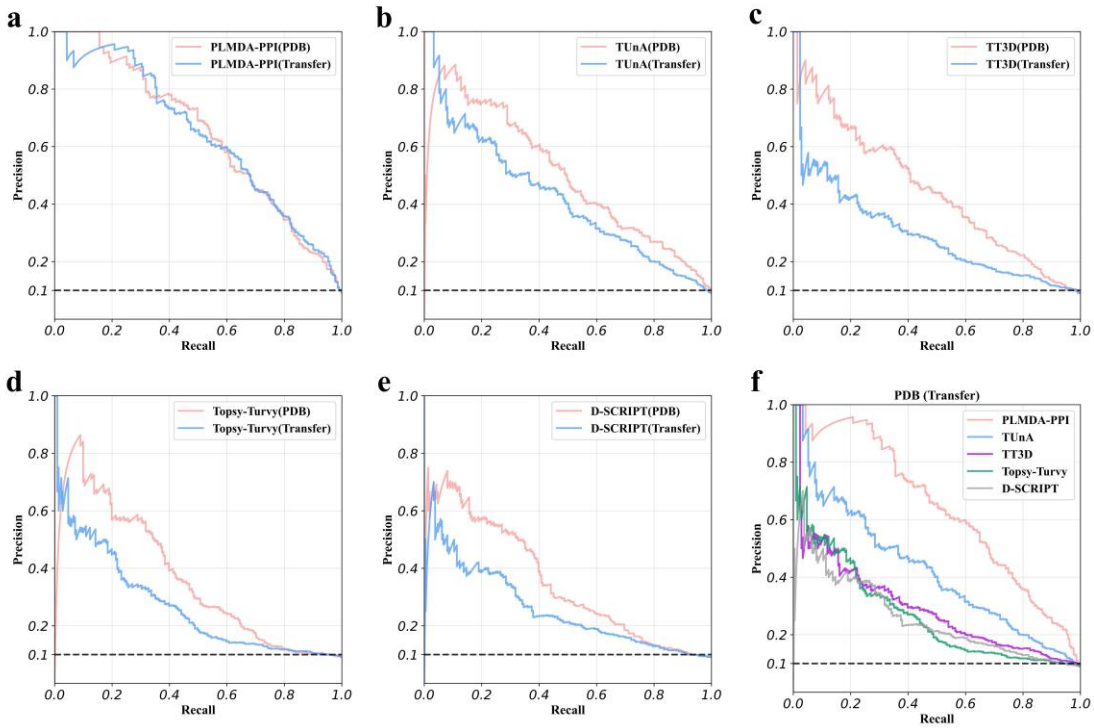


Figure 5. Re-evaluation of the models after transfer learning–based fine-tuning using the HINT *H. sapiens* PPI dataset on the validation set of structure-informed PPIs derived from PDB. Comparisons of the performances (using Precision-Recall curves) before and after transfer learning for PLMDA-PPI (a), TUnA (b), TT3D (c), Topsy-Turvy (d), and D-SCRIPT (e). (f) Comparisons of the performances (using Precision-Recall curves) of the five fine-tuned models on the validation set of structure-informed PPIs derived from PDB.

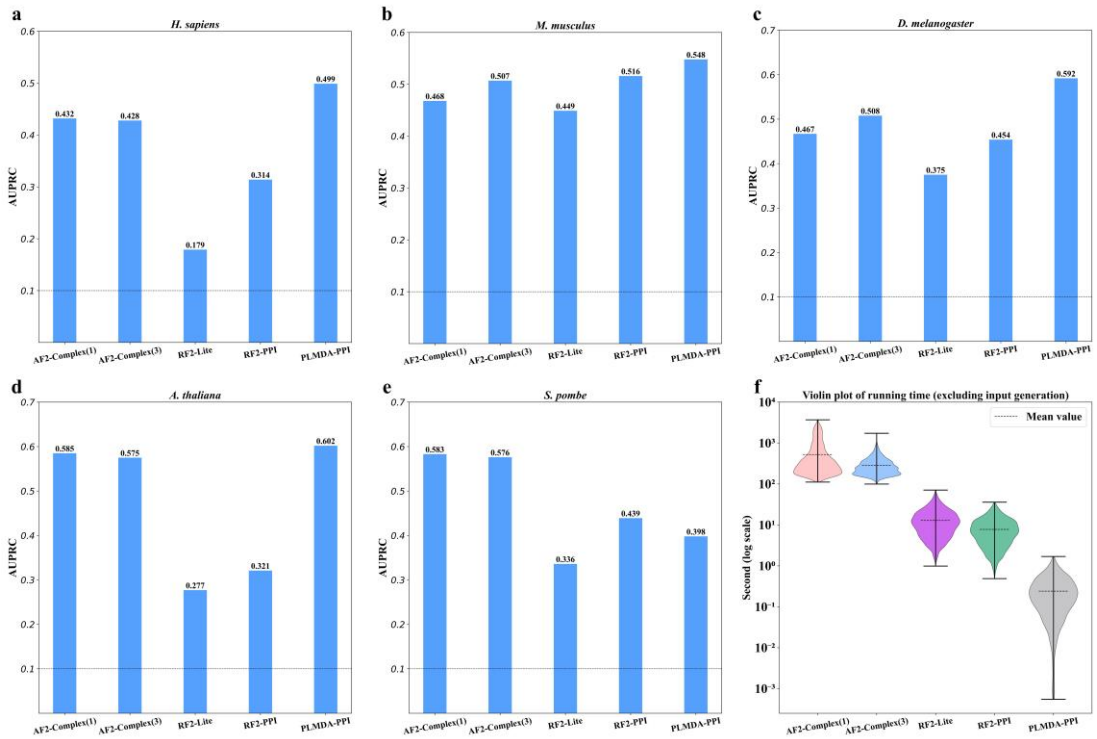


Figure 6. Performances of PLMDA-PPI, AF2-Complex, RF2-Lite, and RF2-PPI on subsets of PPIs sampled from HINT datasets for *H. sapiens* (a), *M. musculus* (b), *D. melanogaster* (c), *A. thaliana* (d), and *S. pombe* (e). For each species, 50 PPIs were randomly selected as positive samples, and ten times as many (i.e., 500) non-interacting protein pairs were generated through random pairing as negative samples. (f) Violin plots of inference-only running times for AF2-Complex, RF2-PPI, and RF2-Lite (time for input feature generation excluded). For AF2Complex, two sets of AlphaFold-Multimer model weight: model_1_multimer_v3 (using monomer templates) and model_3_multimer_v3 (without monomer templates) were employed—denoted as AF2Complex(1) and AF2Complex(3), respectively.

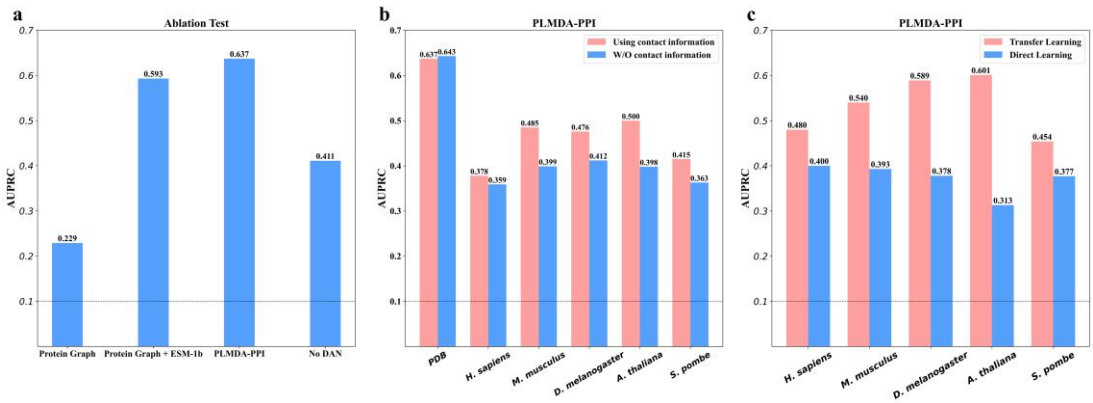


Figure 7. Ablation studies for PLMDA-PPI. (a) Performance variations (in terms of AUPRC) on the validation set of structure-informed PPIs when key components are removed from the model architecture. (b) Performance comparison (in terms of AUPRC) between models trained with and without inter-protein contact information. (c) Performance comparison (in terms of AUPRC) between models with and without the incorporation of structure-informed PPIs from the PDB for the initial training.

Table 1. Validation of the models trained from structure-informed PPIs from PDB.

Dataset	Model	AUPRC	AUROC	Mean Precision (Top 50 Contact Prediction)
PDB dataset	D-SCRIPT	0.324	0.684	0.026
	Topsy-Turvy	0.352	0.756	0.029
	TT3D	0.448	0.826	0.036
	TUnA	0.496	0.873	-
	PLMDA-PPI	**0.637**	**0.911**	**0.287**

We show performances of PLMDA-PPI, D-SCRIPT, Topsy-Turvy, TT3D, and TUnA on the validation set of structure-informed PPIs from PDB after initial learning. AUPRC: area under precision-recall curve; AUROC: area under receiver operating characteristic curve; Mean Precision (Top 50 Contact Prediction): the mean precisions of the top 50 predicted contacts. TUnA does not predict inter-protein contacts. Top-performing metrics are highlighted in bold.

Table 2. Evaluation of the models trained from structure-informed PPIs using PPIs of multiple species from HlNT dataset.

Species	Model	AUPRC	AUROC	AUPRC (No PDB ref)	AUROC (No PDB ref)	AUPRC (Seq id<40%)	AUROC (Seq id<40%)
H. sapiens	D-SCRIPT	0.158	0.605	0.137	0.596	0.151	0.572
	Topsy-Turvy	0.172	0.67	0.149	0.653	0.189	0.661
	TT3D	0.237	0.646	0.223	0.632	0.213	0.635
	TUnA	0.343	0.776	0.312	0.764	0.387	0.786
	PLMDA-PPI	**0.378**	**0.801**	**0.338**	**0.792**	**0.456**	**0.816**
M. musculus	D-SCRIPT	0.219	0.664	0.204	0.647	0.183	0.663
	Topsy-Turvy	0.206	0.62	0.123	0.588	0.189	0.616
	TT3D	0.237	0.69	0.158	0.684	0.222	0.679
	TUnA	0.395	0.799	0.280	0.742	0.437	0.833
	PLMDA-PPI	**0.485**	**0.804**	**0.346**	**0.777**	**0.488**	**0.844**
D. melanogaster	D-SCRIPT	0.188	0.682	0.209	0.665	0.196	0.721
	Topsy-Turvy	0.196	0.629	0.183	0.642	0.241	0.674
	TT3D	0.249	0.709	0.185	0.676	0.327	0.745
	TUnA	0.390	0.796	0.315	0.765	0.487	0.853
	PLMDA-PPI	**0.476**	**0.851**	**0.389**	**0.812**	**0.533**	**0.889**
A. thaliana	D-SCRIPT	0.239	0.697	0.230	0.681	0.205	0.661
	Topsy-Turvy	0.237	0.57	0.209	0.534	0.199	0.567
	TT3D	0.207	0.678	0.176	0.667	0.202	0.707
	TUnA	0.486	0.867	0.472	0.881	0.473	0.868
	PLMDA-PPI	**0.500**	**0.871**	**0.541**	**0.912**	**0.493**	**0.877**
S. pombe	D-SCRIPT	0.173	0.664	0.229	0.680	0.183	0.667
	Topsy-Turvy	0.19	0.592	0.171	0.604	0.161	0.603
	TT3D	0.223	0.712	0.222	0.708	0.255	0.744
	TUnA	0.292	0.760	0.253	0.710	0.344	0.788
	PLMDA-PPI	**0.415**	**0.816**	**0.304**	**0.765**	**0.524**	**0.865**

We show performances of PLMDA-PPI, D-SCRIPT, Topsy-Turvy, TT3D, and TUnA on PPIs of multiple species from HINT with initial training on structure-informed PPIs from PDB. Top-performing metrics are highlighted in bold.

No PDB ref: PPIs in the HINT dataset without PDB reference tags; Seq id<40%: PPIs in the HINT dataset formed by proteins with less than 40% sequence identity to any protein the training dataset.

Table 3. Evaluation of the models after transfer learning on the *H. sapiens* PPI dataset.

Species	Model	AUPRC	AUROC	AUPRC (Seq id < 40%)	AUROC (Seq id < 40%)
H. sapiens (test set)	D-SCRIPT	-	-	0.245	0.713
	Topsy-Turvy	-	-	0.273	0.696
	TT3D	-	-	0.318	0.784
	TUnA	-	-	0.409	0.792
	PLMDA-PPI	-	-	**0.480**	**0.840**
M. musculus	D-SCRIPT	0.313	0.719	0.138	0.661
	Topsy-Turvy	0.331	0.732	0.122	0.583
	TT3D	0.323	0.755	0.192	0.656
	TUnA	0.447	0.844	0.347	0.795
	PLMDA-PPI	**0.540**	**0.840**	**0.405**	**0.801**
D. melanogaster	D-SCRIPT	0.309	0.67	0.129	0.643
	Topsy-Turvy	0.282	0.671	0.115	0.559
	TT3D	0.211	0.743	0.216	0.665
	TUnA	0.396	0.808	0.351	0.809
	PLMDA-PPI	**0.589**	**0.871**	**0.508**	**0.855**
A. thaliana	D-SCRIPT	0.249	0.754	0.139	0.637
	Topsy-Turvy	0.263	0.721	0.133	0.578
	TT3D	0.259	0.724	0.149	0.593
	TUnA	0.348	0.818	0.367	0.824
	PLMDA-PPI	**0.601**	**0.898**	**0.585**	**0.899**
S. pombe	D-SCRIPT	0.234	0.682	0.114	0.570
	Topsy-Turvy	0.242	0.664	0.110	0.537
	TT3D	0.278	0.744	0.180	0.625
	TUnA	0.365	0.770	0.294	0.759
	PLMDA-PPI	**0.454**	**0.830**	**0.338**	**0.826**

We show performances of PLMDA-PPI, D-SCRIPT, Topsy-Turvy, TT3D, and TUnA on PPIs of multiple species from HINT after transfer learning on the HINT *H. sapiens* PPI dataset. Top-performing metrics are highlighted in bold.

Seq id<40%: PPIs in the HINT dataset formed by proteins with less than 40% sequence identity to any protein the training dataset.

Table 4. Reassessment of the models after transfer learning on the validation set of structure-informed PPIs.

Dataset	Model	AUPRC	AUROC	Mean Precision (Top50 Contact Prediction)
PDB dataset	D-SCRIPT	0.258	0.719	0.026
	Topsy-Turvy	0.271	0.697	0.029
	TT3D	0.296	0.757	0.036
	TUnA	0.421	0.827	-
	PLMDA-PPI	**0.642**	**0.907**	**0.230**

We re-evaluated the performance of PLMDA-PPI, D-SCRIPT, Topsy-Turvy, TT3D, and TUnA after transfer learning on the HINT *H. sapiens* PPI dataset, using the validation set of structure-informed PPIs derived from PDB. Top-performing metrics are highlighted in bold.

Table 5. Performances of AF2-Complex, RF2-Lite, RF2-PPI, and PLMDA-PPI on PPIs sampled from HINT datasets

Species	Model	AUPRC	AUROC
H. sapiens	AF2-Complex(1)	0.432	0.696
	AF2-Complex(3)	0.428	0.702
	RF2-Lite	0.179	0.594
	RF2-PPI	0.314	0.645
	PLMDA-PPI	**0.499**	**0.839**
M. musculus	AF2-Complex(1)	0.468	0.687
	AF2-Complex(3)	0.507	0.734
	RF2-Lite	0.449	0.717
	RF2-PPI	0.516	0.727
	PLMDA-PPI	**0.548**	**0.877**
D. melanogaster	AF2-Complex(1)	0.467	0.748
	AF2-Complex(3)	0.508	0.753
	RF2-Lite	0.375	0.716
	RF2-PPI	0.454	0.675
	PLMDA-PPI	**0.592**	**0.861**
A. thaliana	AF2-Complex(1)	0.585	0.729
	AF2-Complex(3)	0.575	0.766
	RF2-Lite	0.277	0.689
	RF2-PPI	0.321	0.631
	PLMDA-PPI	**0.602**	**0.906**
S. pombe	AF2-Complex(1)	**0.583**	0.762
	AF2-Complex(3)	0.576	0.757
	RF2-Lite	0.336	0.711
	RF2-PPI	0.439	0.665
	PLMDA-PPI	0.398	**0.818**

We show performances of AF2Complex, RF2-Lite, RF2-PPI, and PLMDA-PPI on the subsets of PPIs sampled from HINT datasets. For AF2Complex, we employed two sets of AlphaFold-Multimer model weights recommended in the documentation: model_1_multimer_v3 (using monomer templates) and model_3_multimer_v3 (without monomer templates)—denoted as AF2Complex(1) and AF2Complex(3), respectively. Top-performing metrics are highlighted in bold.

Supplementary

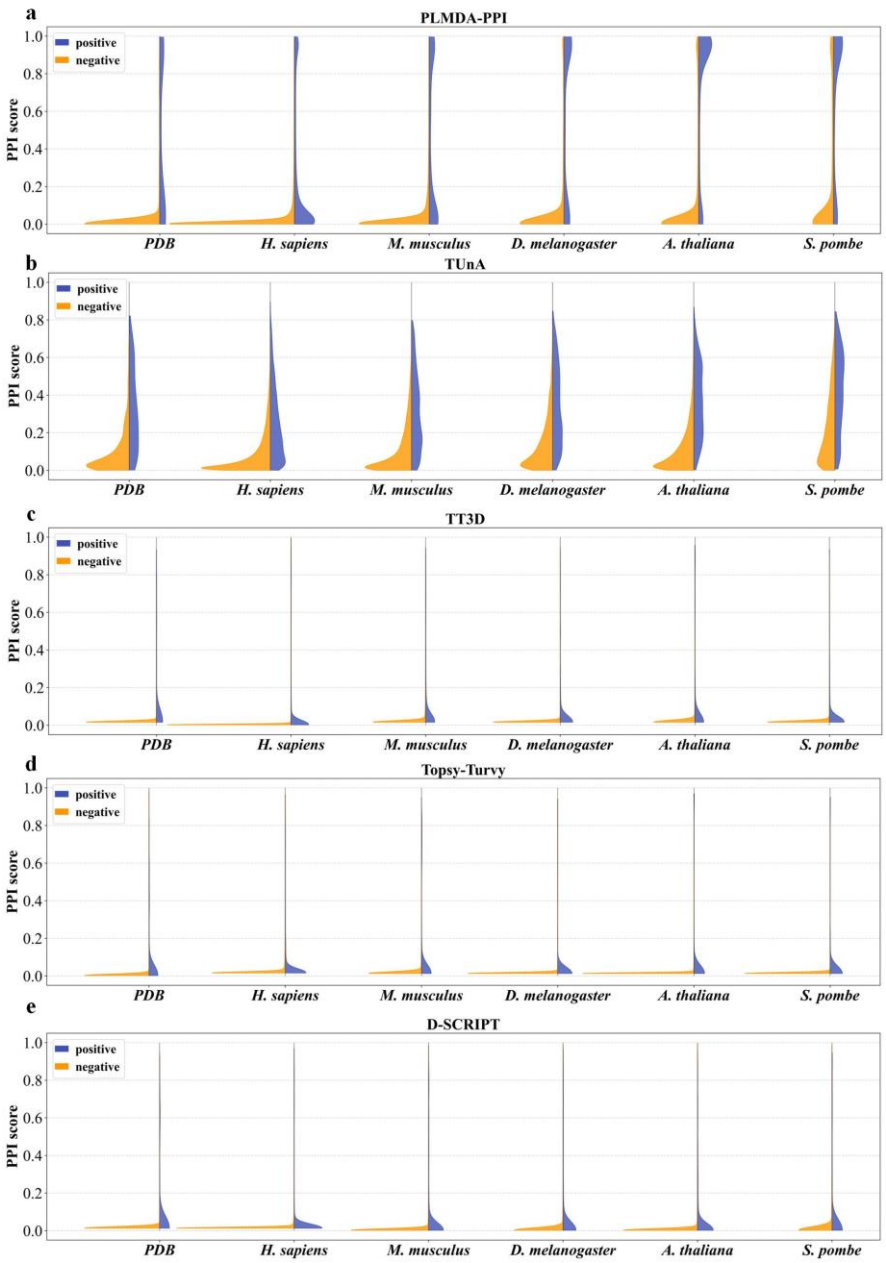


Figure S1. Distributions of PPI scores predicted by PLMDA-PPI (a), TUnA (b), TT3D (c), Topsy-Turvy (d), and D-SCRIPT (e) after initial training on structure-informed PPIs, for positive and negative samples in the validation set derived from structure-informed PPIs in PDB and the HINT test datasets across multiple species.

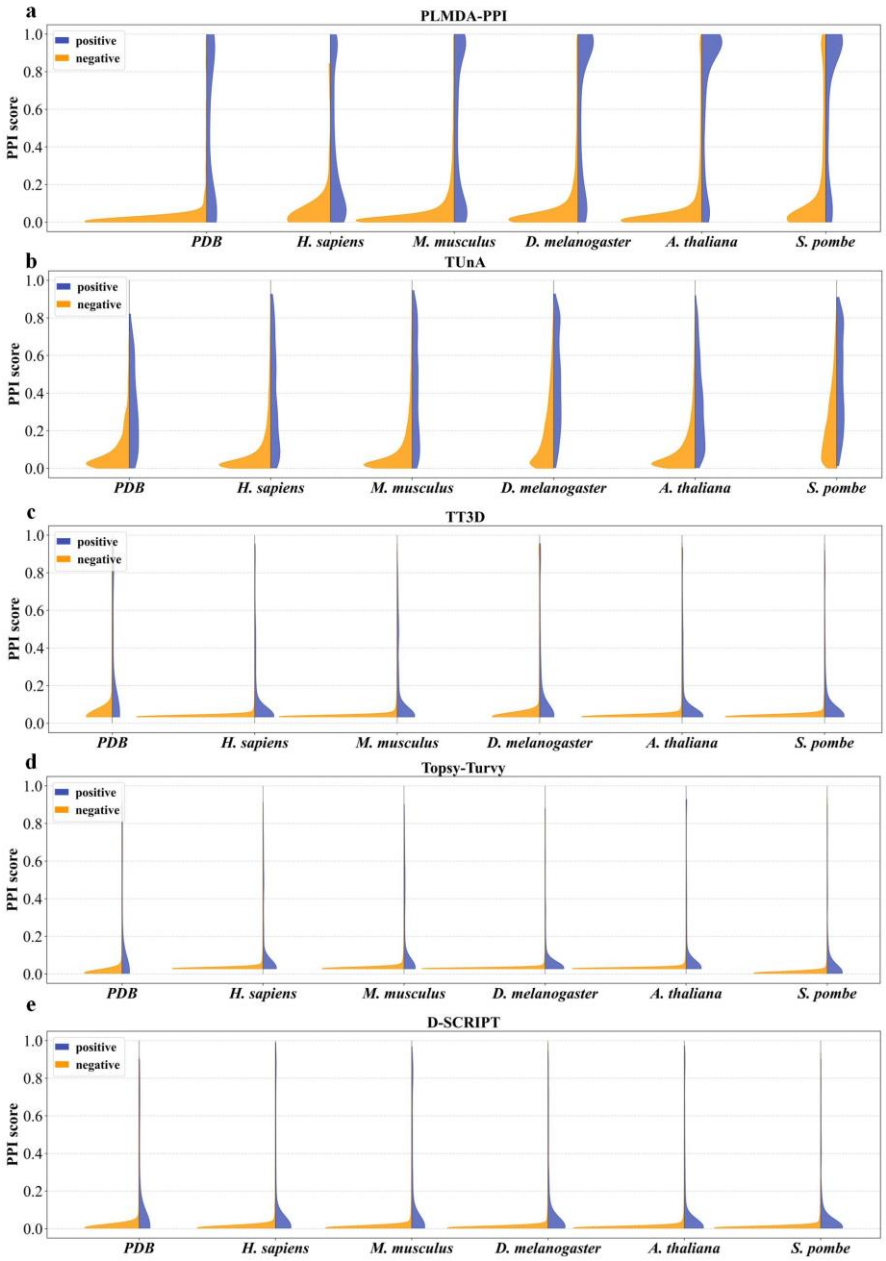


Figure S2. Distributions of PPI scores predicted by PLMDA-PPI (a), TUnA (b), TT3D (c), Topsy-Turvy (d), and D-SCRIPT (e) after transfer learning, for positive and negative samples in the validation set derived from structure-informed PPIs in PDB and the HINT test datasets across multiple species.

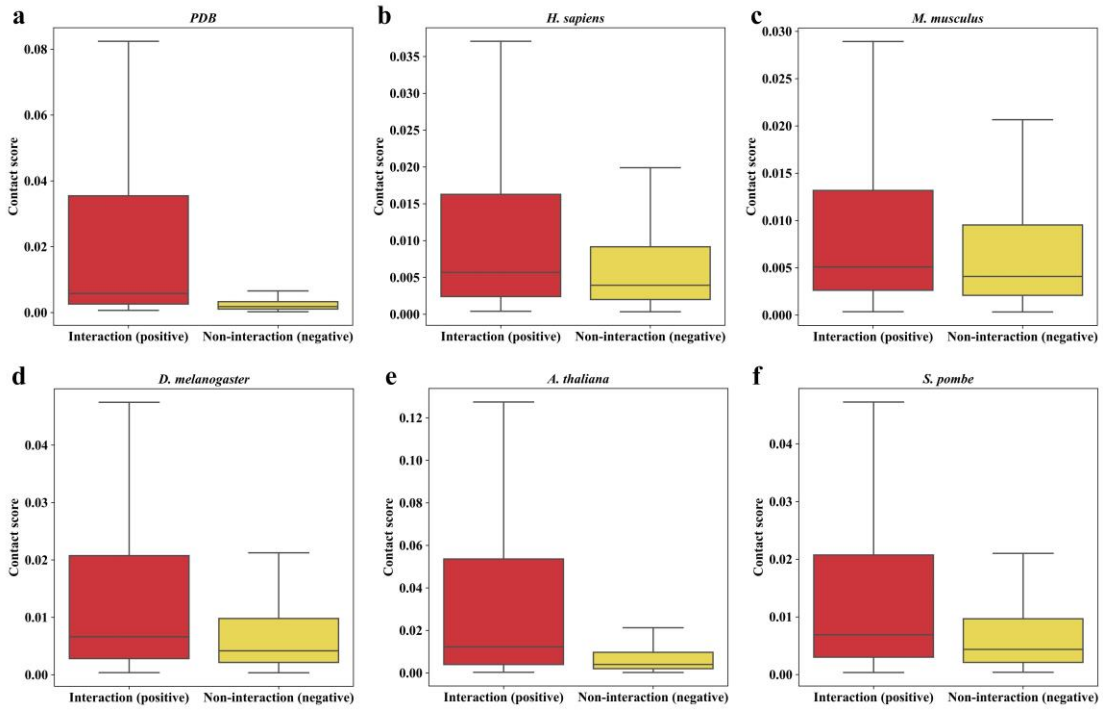


Figure S3. Comparison of the distributions of the mean contact score of the top 50 predicted residue pairs by PLMDA-PPI after transfer learning, for positive and negative samples in the validation set derived from structure-informed PPIs in PDB and the HINT test datasets across multiple species.

Table S1. Performance comparison of PLMDA-PPI before and after transfer learning on HINT datasets across multiple species, following the exclusion of PPIs involving any protein sharing more than 40% sequence identity with proteins in the *H. sapiens* training set.

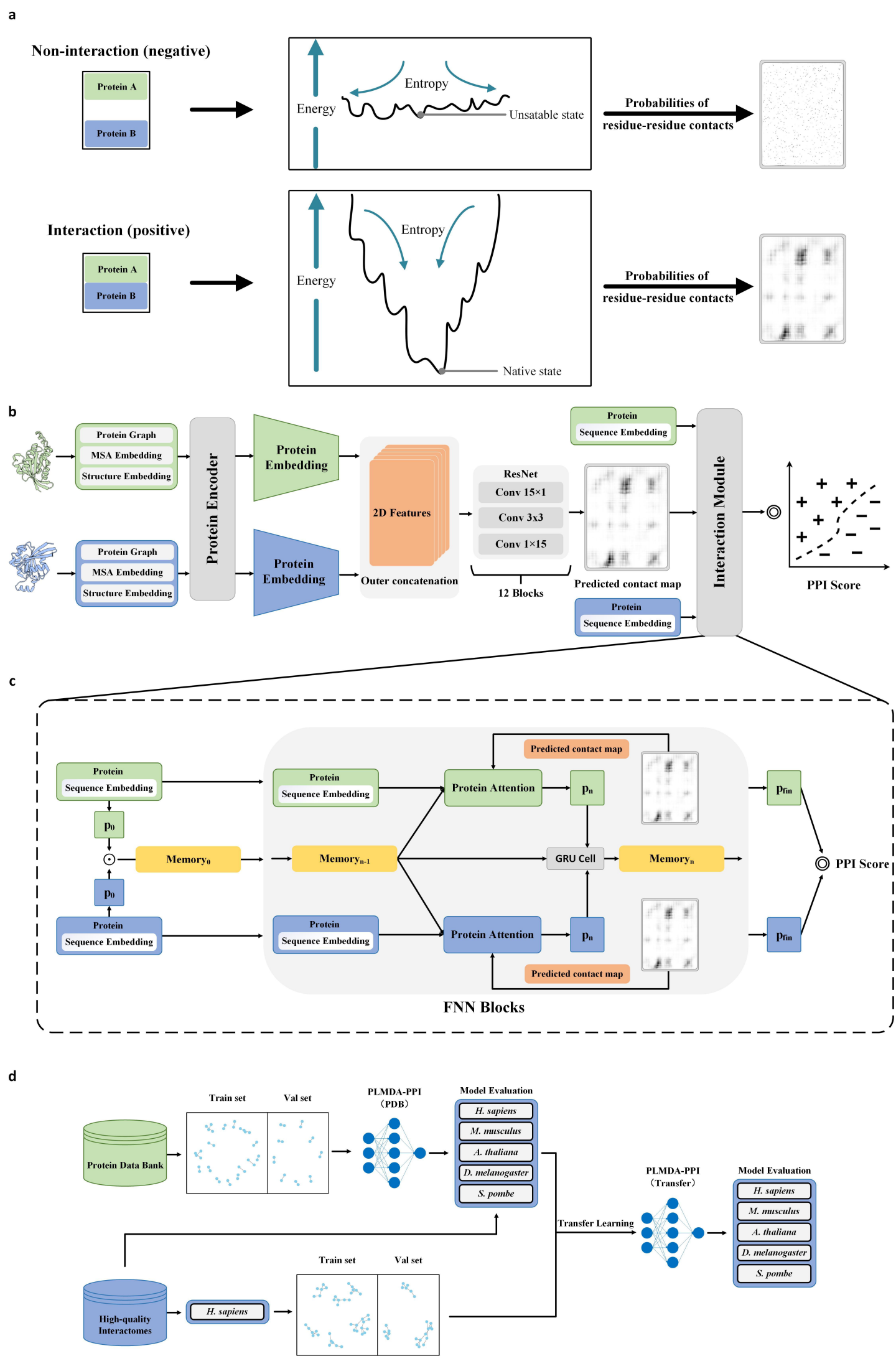
Species	AUPRC (PDB)	AUROC (PDB)	AUPRC (Transfer)	AUROC (Transfer)
H. sapiens	0.387	0.795	**0.480**	**0.840**
M. musculus	0.389	0.789	**0.405**	**0.801**
D. melanogaster	0.429	0.838	**0.508**	**0.855**
A. thaliana	0.490	0.883	**0.585**	**0.899**
S. pombe	**0.342**	0.794	0.338	**0.826**

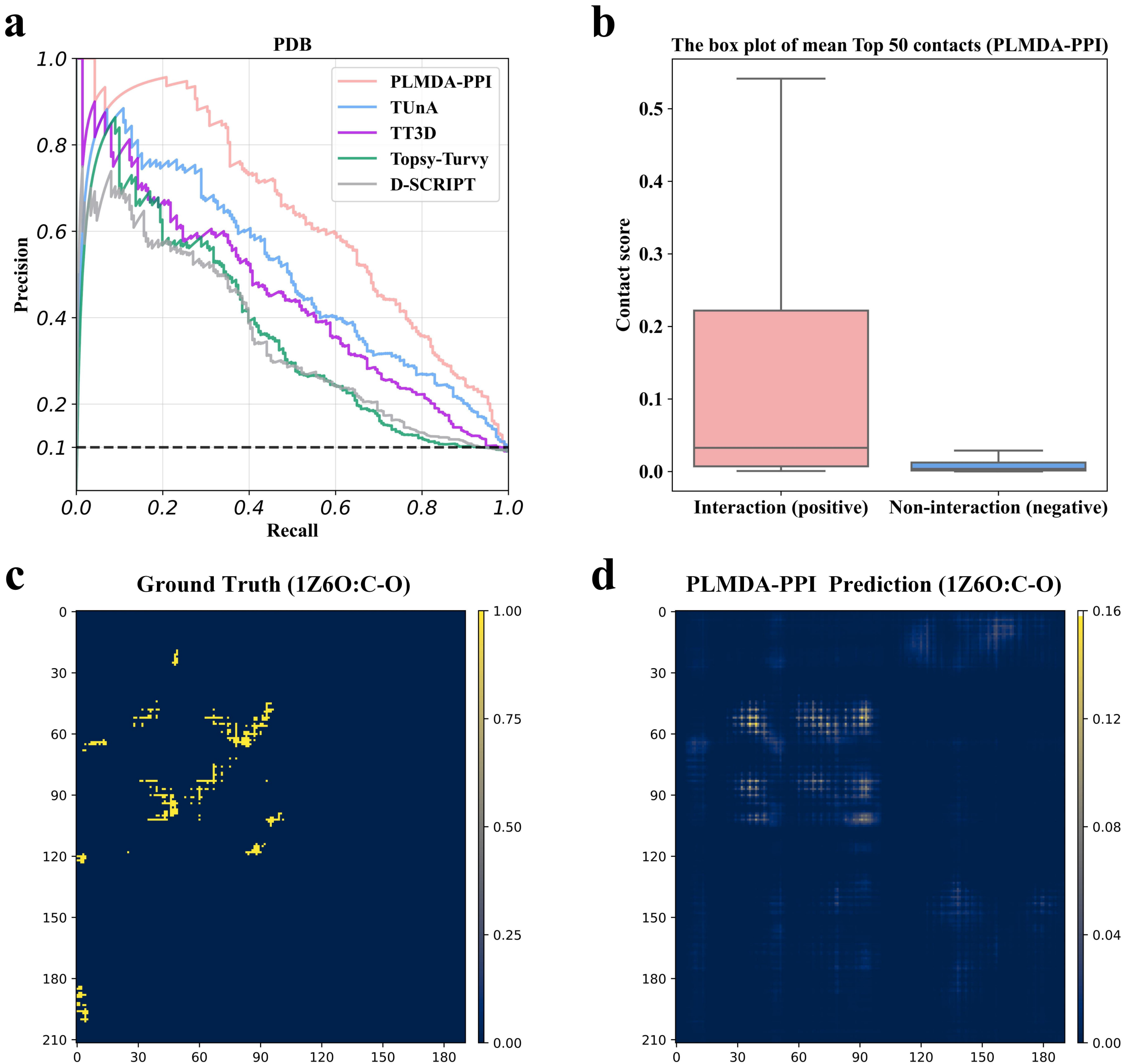
Top-performing metrics are highlighted in bold.

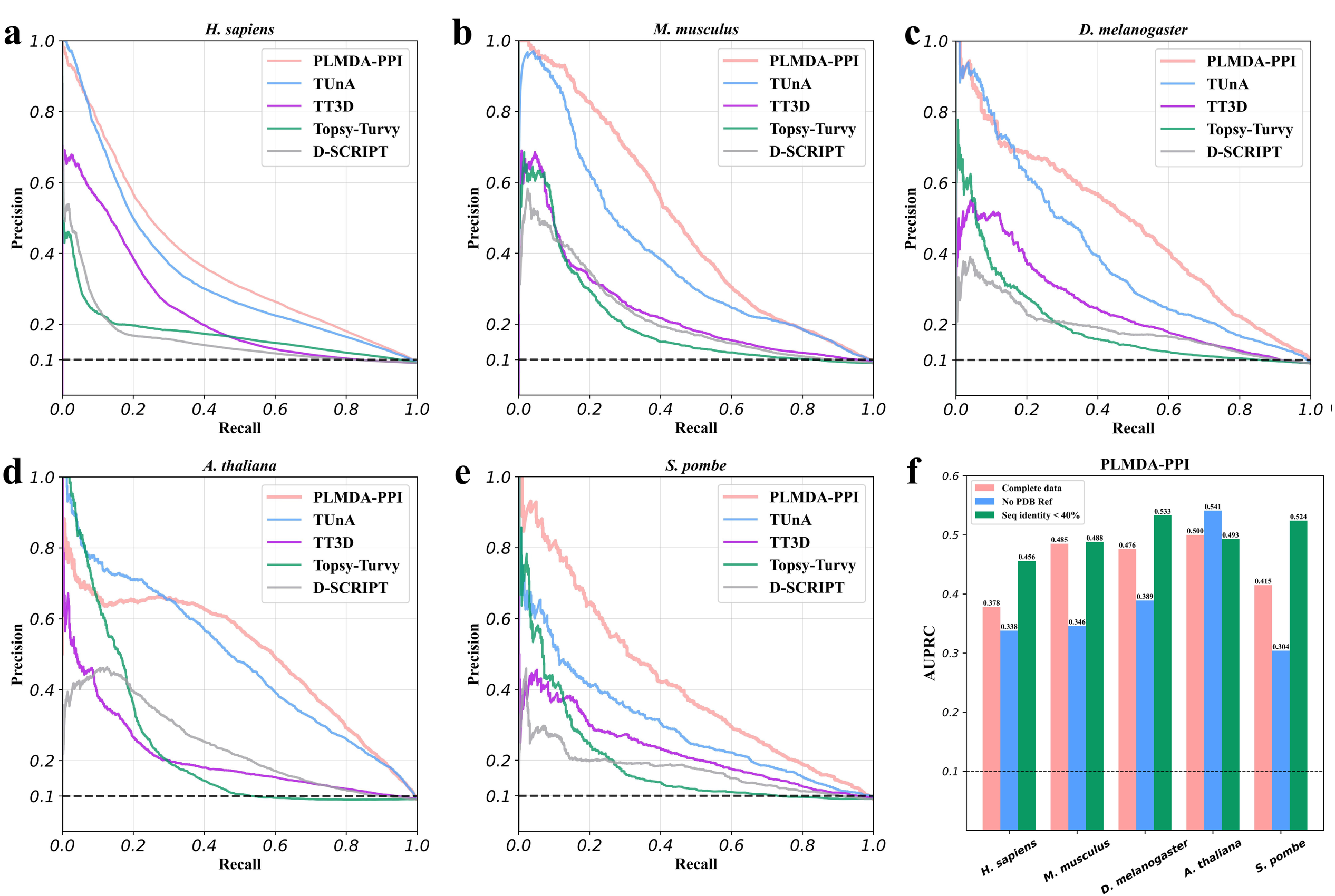
Table S2. Performances of the models when trained directly on the HINT *H. sapiens* PPI training set

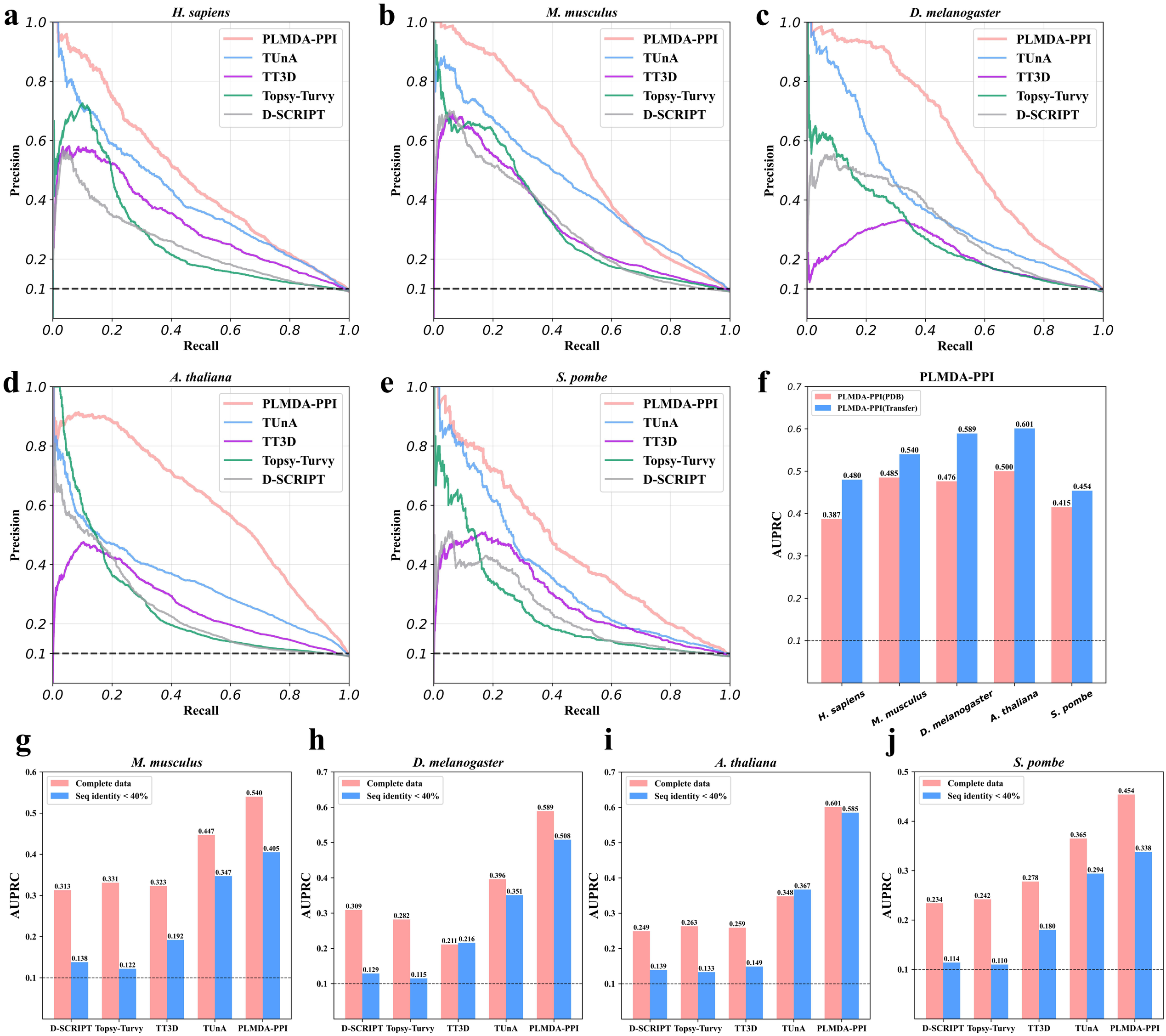
Species	Model	AUPRC	AUROC
H. sapiens	D-SCRIPT	0.197	0.74
	Topsy-Turvy	0.212	0.709
	TT3D	0.227	0.706
	TUnA	0.313	0.791
	PLMDA-PPI	**0.400**	**0.812**
M. musculus	D-SCRIPT	0.169	0.713
	Topsy-Turvy	0.16	0.63
	TT3D	0.196	0.688
	TUnA	0.311	0.803
	PLMDA-PPI	**0.393**	**0.838**
A. thaliana	D-SCRIPT	0.134	0.635
	Topsy-Turvy	0.127	0.576
	TT3D	0.153	0.593
	TUnA	0.238	0.738
	PLMDA-PPI	**0.313**	**0.829**
D. melanogaster	D-SCRIPT	0.123	0.615
	Topsy-Turvy	0.141	0.563
	TT3D	0.239	0.685
	TUnA	0.210	0.748
	PLMDA-PPI	**0.378**	**0.828**
S. pombe	D-SCRIPT	0.119	0.597
	Topsy-Turvy	0.141	0.573
	TT3D	0.204	0.66
	TUnA	0.238	0.684
	PLMDA-PPI	**0.377**	**0.770**

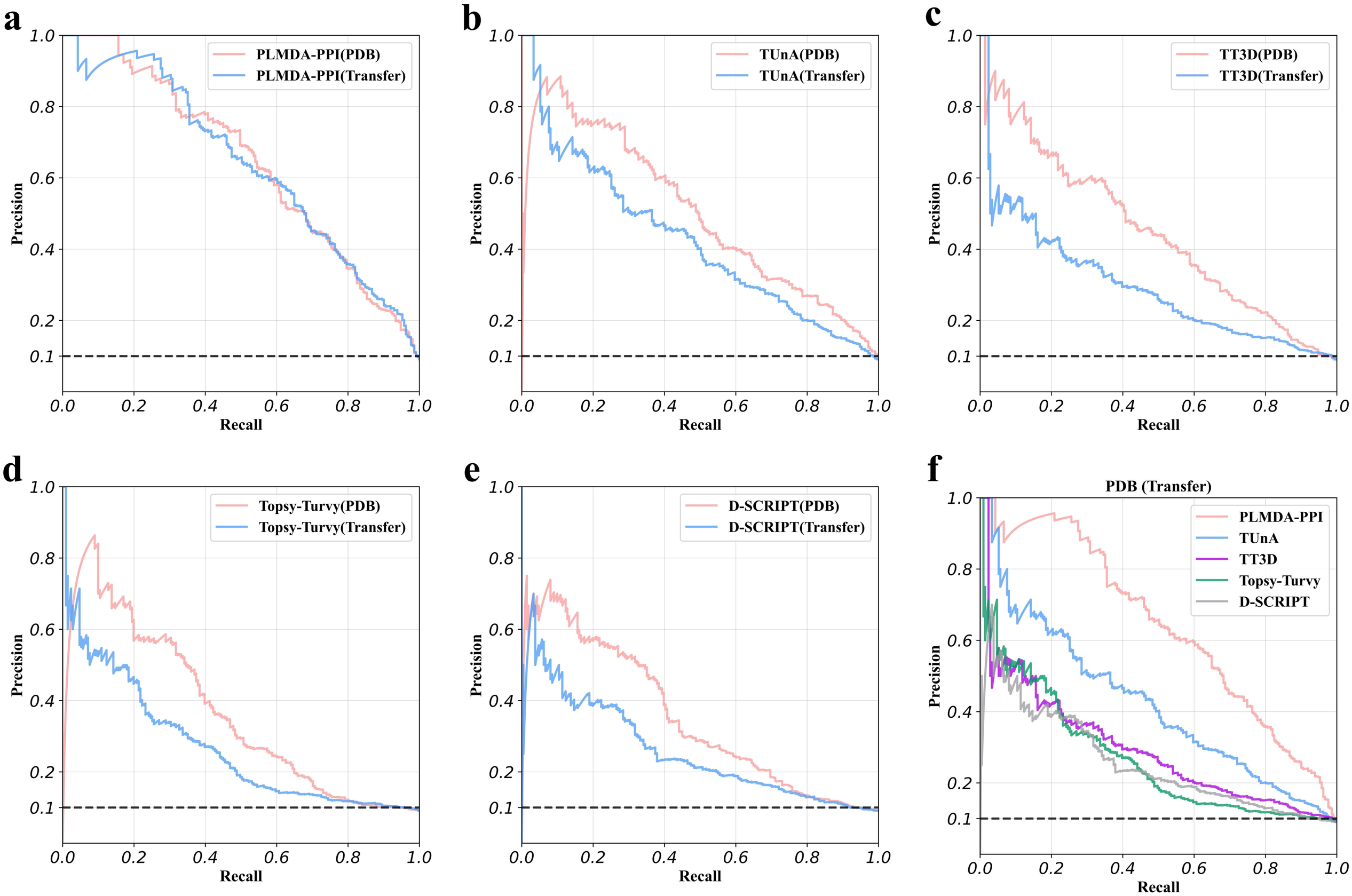
Top-performing metrics are highlighted in bold.

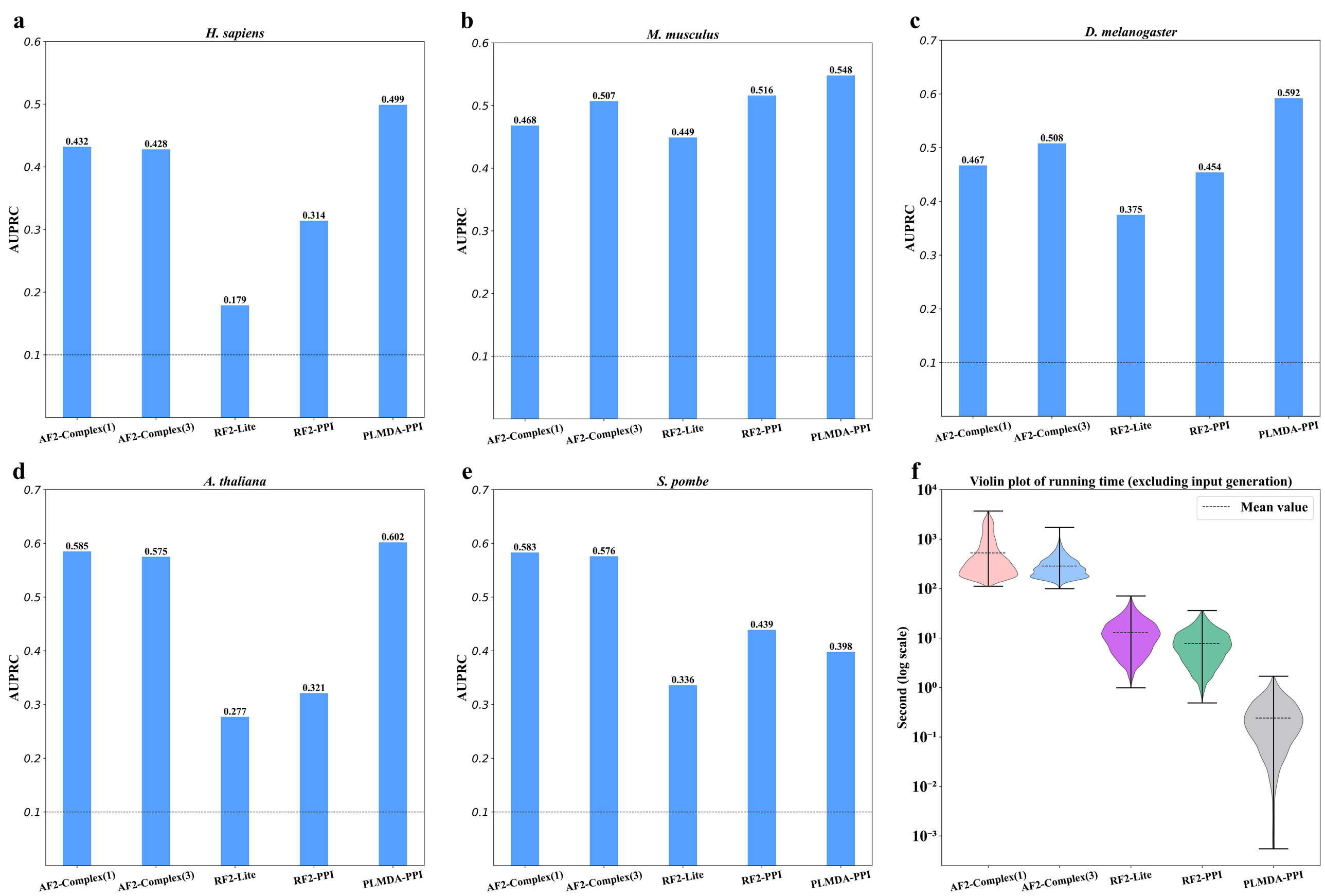

a
Non-interaction (negative)
Protein A
Protein B
Energy
Entropy
Unsatable state
Probabilities of residue-residue contacts
Interaction (positive)
Protein A
Protein B
Energy
Entropy
Native state
Probabilities of residue-residue contacts
b
Protein Graph
MSA Embedding
Structure Embedding
Protein Graph
MSA Embedding
Structure Embedding
Protein Encoder
Protein Embedding
Protein Embedding
2D Features
Outer concatenation
ResNet
Conv 15×1
Conv 3x3
Conv 1×15
12 Blocks
Protein Sequence Embedding
Predicted contact map
Protein Sequence Embedding
Interaction Module
PPI Score
c
Protein Sequence Embedding
p0
Memory0
p0
Protein Sequence Embedding
Protein Sequence Embedding
Memoryn-1
Protein Sequence Embedding
Predicted contact map
Protein Attention
pn
GRU Cell
Memoryn
Protein Attention
pn
Predicted contact map
pfin
pfin
PPI Score
FNN Blocks
d
Protein Data Bank
Train set
Val set
PLMDA-PPI (PDB)
Model Evaluation
H. sapiens
M. musculus
A. thaliana
D. melanogaster
S. pombe
Transfer Learning
PLMDA-PPI (Transfer)
Model Evaluation
H. sapiens
M. musculus
A. thaliana
D. melanogaster
S. pombe
High-quality Interactomes
H. sapiens
Train set
Val set

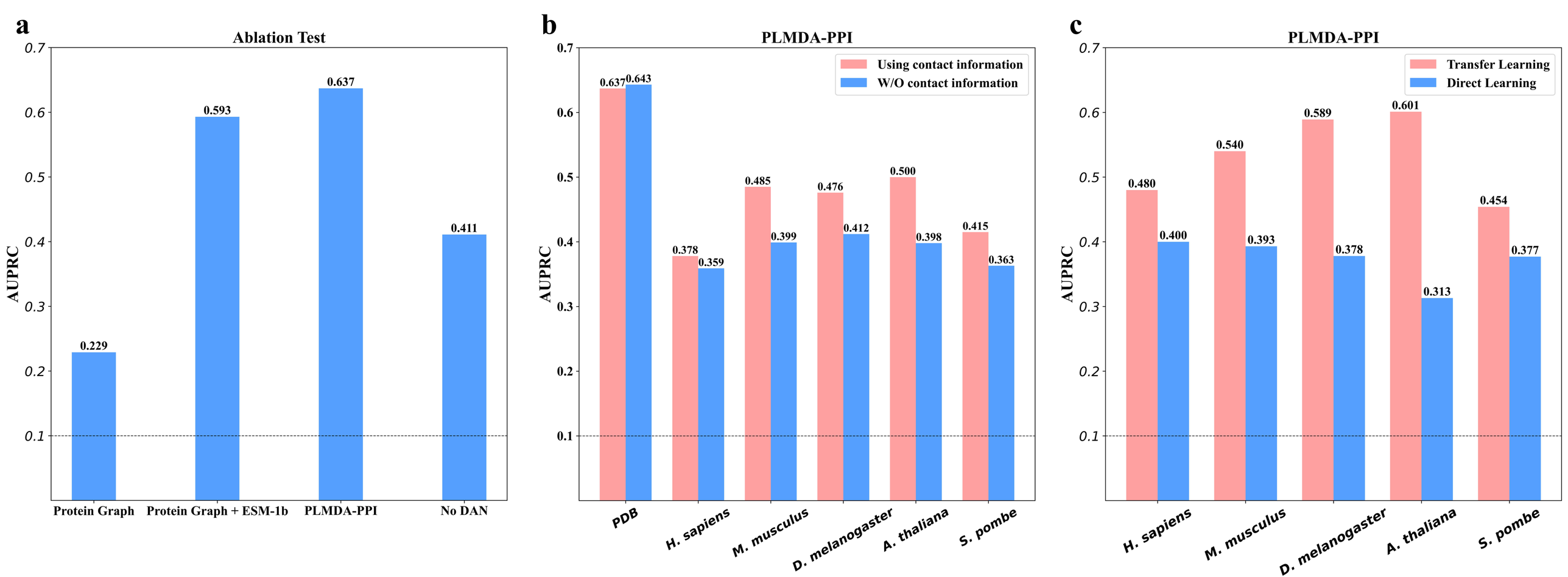

a
PDB
PLMDA-PPI
TUnA
TT3D
Topsy-Turvy
D-SCRIPT
Precision
Recall
b
The box plot of mean Top 50 contacts (PLMDA-PPI)
Contact score
Interaction (positive)
Non-interaction (negative)
c
Ground Truth (1Z6O:C-O)
d
PLMDA-PPI Prediction (1Z6O:C-O)


a
H. sapiens
b
M. musculus
c
D. melanogaster
d
A. thaliana
e
S. pombe
PLMDA-PPI
TUnA
TT3D
Topsy-Turvy
D-SCRIPT
Precision
Recall
f
PLMDA-PPI
Complete data
No PDB Ref
Seq identity < 40%
AUPRC
0.378
0.338
0.456
0.485
0.346
0.488
0.476
0.389
0.533
0.500
0.541
0.493
0.415
0.304
0.524
H. sapiens
M. musculus
D. melanogaster
A. thaliana
S. pombe


a
H. sapiens
b
M. musculus
c
D. melanogaster
d
A. thaliana
e
S. pombe
PLMDA-PPI
TUnA
TT3D
Topsy-Turvy
D-SCRIPT
Precision
Recall
f
PLMDA-PPI
PLMDA-PPI(PDB)
PLMDA-PPI(Transfer)
AUPRC
0.387
0.480
0.485
0.540
0.476
0.589
0.500
0.601
0.415
0.454
H. sapiens
M. musculus
D. melanogaster
A. thaliana
S. pombe
g
M. musculus
h
D. melanogaster
i
A. thaliana
j
S. pombe
Complete data
Seq identity < 40%
AUPRC
0.313
0.138
0.331
0.122
0.323
0.192
0.447
0.347
0.540
0.405
0.309
0.129
0.282
0.115
0.211
0.216
0.396
0.351
0.589
0.508
0.249
0.139
0.263
0.133
0.259
0.149
0.348
0.367
0.601
0.585
0.234
0.114
0.242
0.110
0.278
0.180
0.365
0.294
0.454
0.338
D-SCRIPT
Topsy-Turvy
TT3D
TUnA
PLMDA-PPI


a
PLMDA-PPI(PDB)
PLMDA-PPI(Transfer)
b
TUnA(PDB)
TUnA(Transfer)
c
TT3D(PDB)
TT3D(Transfer)
d
Topsy-Turvy(PDB)
Topsy-Turvy(Transfer)
e
D-SCRIPT(PDB)
D-SCRIPT(Transfer)
f
PDB (Transfer)
PLMDA-PPI
TUnA
TT3D
Topsy-Turvy
D-SCRIPT
Precision
Recall
0.0
0.1
0.2
0.4
0.6
0.8
1.0


a
H. sapiens
AUPRC
0.432
0.428
0.179
0.314
0.499
AF2-Complex(1)
AF2-Complex(3)
RF2-Lite
RF2-PPI
PLMDA-PPI
b
M. musculus
AUPRC
0.468
0.507
0.449
0.516
0.548
AF2-Complex(1)
AF2-Complex(3)
RF2-Lite
RF2-PPI
PLMDA-PPI
c
D. melanogaster
AUPRC
0.467
0.508
0.375
0.454
0.592
AF2-Complex(1)
AF2-Complex(3)
RF2-Lite
RF2-PPI
PLMDA-PPI
d
A. thaliana
AUPRC
0.585
0.575
0.277
0.321
0.602
AF2-Complex(1)
AF2-Complex(3)
RF2-Lite
RF2-PPI
PLMDA-PPI
e
S. pombe
AUPRC
0.583
0.576
0.336
0.439
0.398
AF2-Complex(1)
AF2-Complex(3)
RF2-Lite
RF2-PPI
PLMDA-PPI
f
Violin plot of running time (excluding input generation)
Mean value
Second (log scale)
AF2-Complex(1)
AF2-Complex(3)
RF2-Lite
RF2-PPI
PLMDA-PPI


a
Ablation Test
AUPRC
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.229
0.593
0.637
0.411
Protein Graph
Protein Graph + ESM-1b
PLMDA-PPI
No DAN
b
PLMDA-PPI
Using contact information
W/O contact information
0.637
0.643
0.378
0.359
0.485
0.399
0.476
0.412
0.500
0.398
0.415
0.363
PDB
H. sapiens
M. musculus
D. melanogaster
A. thaliana
S. pombe
c
PLMDA-PPI
Transfer Learning
Direct Learning
0.480
0.400
0.540
0.393
0.589
0.378
0.601
0.313
0.454
0.377
H. sapiens
M. musculus
D. melanogaster
A. thaliana
S. pombe