A Corporative Language Model for Protein–Protein Interaction, Binding Affinity, and Interface Contact Prediction

Jun Liu[1], Hungyu Chen[2], Yang Zhang[1,2,3,*]

[1]*Cancer Science Institute of Singapore, National University of Singapore, 117599, Singapore.*
[2]*Department of Computer Science, School of Computing, National University of Singapore, 117417, Singapore.*
[3]*Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117596, Singapore.*
[*] *Correspondence should be addressed to (zhang@nus.edu.sg)*

Abstract

Understanding protein–protein interactions (PPIs) is crucial for deciphering cellular processes and guiding therapeutic discovery. While recent protein language models have advanced sequence-based protein representation, most are designed for individual chains and fail to capture inherent PPI patterns. We introduce a novel Protein–Protein Language Model (PPLM) that jointly encodes interacting protein pair representations. Building on this foundation, we develop three specialized predictors—PPLM-PPI, PPLM-Affinity, and PPLM-Contact—targeting binary interaction, binding affinity, and interface contact predictions, respectively. Large-scale experiments show that PPLM-PPI achieves state-of-the-art performance across different species on binary interaction prediction, while PPLM-Affinity outperforms both ESM2 and structure-based methods on binding affinity modeling, particularly on challenging cases including antibody–antigen and TCR–pMHC complexes. PPLM-Contact further surpasses existing PPI contact predictors on inter-protein contact prediction and interface residue recognition, including those deduced from cutting-edge complex structure predictions. Together, these results highlight the potential of co-represented language models to advance computational modeling of PPIs.

Introduction

Protein–protein interactions (PPIs) are central to virtually all cellular processes, ranging from signal transduction and immune recognition to metabolic regulation and molecular assembly[1] Understanding PPIs involves three fundamental and interrelated questions: Do two proteins interact? How strong is their binding? And which residues mediate the interaction? These questions map directly onto the key computational challenges of binary interaction prediction, binding affinity estimation, and interface residue identification[2-4].

Large-scale protein language models (PLMs), such as ESM[5-7] and ProtBERT[8], have demonstrated strong capabilities in learning protein sequence representations, driving progress in structure prediction[9, 10], function annotation[11], protein design[12], and mutational effect estimation[13]. Although PLMs have been widely applied in PPI-related studies[14-20], these models are inherently trained on single-protein sequences and lack the ability to model inter-protein relationships. For protein interaction prediction, for example, methods such as D-SCRIPT[15], Topsy-Turvy[21], and ESMDNN-PPI[22] leverage either LSTM-based encoders or ESM2-derived embeddings from single-sequence for binary classification, yet struggle to capture interface-level context. In the task of binding affinity prediction, most PLM-based approaches are designed for protein-ligand targets[23, 24], with existing methods for protein–protein binding affinity prediction still depending heavily on complex structural models[25-29], highlighting a key opportunity for developing scalable, structure-free solutions based on sequence-derived representations. For inter-protein contact prediction, most methods are built on multiple sequence alignments (MSAs)[30-33]. While methods such as GLINTER[34], CDPred[19], and DeepInter[20] employ ESM-MSA-1b[35] to extract both intra- and inter-protein features, their performance is tightly coupled to the quality and availability of paired MSAs, as well as the fact that ESM-MSA-1b has been pretrained solely on monomeric sequences. These limitations, shared across all three major PPI-related tasks, underscore the need for a unified paired protein language model that can directly learn inter-protein relationships from sequences.

To address these limitations, we introduce PPLM, a novel protein–protein language model specifically designed to learn contextual and relational features from paired protein sequences. To model inter-protein relationships accurately, we design a new hybrid attention mechanism that employes Rotary embeddings[36] for intra-protein positional encoding and non-positional embedding for inter-protein residue pairs to avoid spurious spatial priors. Learnable inter-protein attention weights and an inter-protein attention mask are then introduced to enhance the learning of inter-protein attention and explicitly separate intra- from inter-protein dependencies. To support large-scale training, we construct a comprehensive paired sequence dataset comprising over 3.3 million high-quality protein pairs curated from the Protein Data Bank (PDB)[37] and STRING[38] databases. This dataset allows robust training of PPLM to capture interaction-aware representations.

Building upon PPLM, we develop three task-specific methods for key PPI-related challenges: PPLM-PPI for binary interaction prediction, PPLM-Affinity for binding affinity estimation, and PPLM-Contact for inter-protein contact prediction and interface residue identification. Across these tasks, PPLM consistently outperforms existing methods, with notable gains in challenging scenarios such as antibody–antigen and TCR–pMHC complexes. In inter-protein contact-map prediction, PPLM-Contact surpasses current state-of-the-art structure modelling approaches, including AlphaFold2.3[39], AlphaFold3[40], and DMFold[41]. These results underscore PPLM's potential as a unified foundation for advancing computational PPI studies.

Results

PPLM consists of four distinct pipelines designed for the core language model and its downstream applications. In the primary language modeling pipeline (**Figure 1A**), input sequences are independently tokenized with begin and end tokens, then concatenated to form the full-size encoding. A specialized attention mechanism is then employed to capture both inter- and intra-protein residue interactions in parallel. The resulting embeddings and attention matrix from PPLM can be utilized in

various ways for PPI-related research, offering flexible features for a range of downstream tasks.

To illustrate the versatility of PPLM, we developed three task-specific models targeting major challenges in PPI studies. For PPI prediction (**Figure 1B**), PPLM generates interaction features including intra-protein embeddings, intra-protein attention maps, and inter-protein attention maps. These are processed independently and then concatenated into a single representation, which is passed through a multilayer perceptron (MLP) with sigmoid activation to predict the probability of interaction between the input sequence pairs. For binding affinity prediction (**Figure 1C**), the final layer of PPLM is fine-tuned on a small, high-quality dataset of experimentally measured affinities. The full-sequence embeddings of the receptor and ligand proteins are aggregated using max pooling and then passed through linear layers to estimate their binding strength. Finally, for inter-protein contact-map prediction (**Figure 1D**), the inter-protein attention matrix from PPLM is integrated with MSA-derived features and monomer distance maps to capture both evolutionary and structural information. A dedicated inter-protein transformer, composed of three specialized modules designed to update inter-protein representations, is trained to predict the inter-protein contacts and identify interface residues between the input monomer proteins.

To rigorously evaluate PPLM's performance, we compiled independent benchmark datasets for each downstream task: (1) For language model evaluation, 1,194 homodimer and 1,306 heterodimer protein sequence pairs were collected from the PDB[37], along with 2,500 high-confidence sequence pairs from STRING[38]; (2) For protein–protein interaction prediction, 22,000 positive and 216,183 negative PPI samples spanning 5 species were used[15]; (3) For binding affinity prediction, 12,062 interaction samples from 3,032 distinct PDB entries were used in five-fold cross-validation setting[27]; and (4) For inter-protein contact prediction and interface residue identification, 300 homodimer and 99 heterodimer proteins from DeepInter[42], as well as 43 homodimer and 20 heterodimer proteins from the Critical Assessment of Structure Prediction (CASP13-16) experiments. Both the training and test datasets are made publicly available at https://zhanggroup.org/PPLM/benchmark.html.

Detailed methodological descriptions and dataset specifications are provided in the **Methods** section.

PPLM improves language modeling of protein pairs

Perplexity serves as a task-independent and objective metric for evaluating the predictive performance of protein language models, capturing the model's uncertainty in predicting amino acid residues. Lower perplexity values correspond to higher confidence and accuracy, with a score of 1 indicating perfect prediction and value ~20 representing near-random performance. In this study, we first employed perplexity to benchmark the performance of PPLM against ESM2[6], a state-of-the-art single-sequence protein language model trained on large-scale individual protein sequences without explicitly modeling inter-protein context, on the validation dataset of PPLM.

The left panel of **Figure 2A** presents the perplexity results under a random masking strategy, where each residue from both input sequences was sequentially masked to compute pseudo perplexity. PPLM achieves average perplexities of 3.00 and 4.01 for protein sequence pairs from PDB and STRING, respectively, representing reductions of 23.5% and 22.6% compared to ESM2 (3.92 and 5.18). These improvements are statistically significant, with Wilcoxon signed-rank test p-values < 1E-300 for both datasets. The median perplexities are 1.83 (PDB) and 3.88 (STRING), which are 22.5% and 20.3% lower than those of ESM2 (2.36 and 4.87). Notably, the observed median perplexities of PPLM are considerably lower than the corresponding means, indicating that PPLM performs well on the majority of samples, with a small number of challenging outliers contributing to higher perplexity values. **Figure 2B** further presents a head-to-head comparison of perplexity under the random masking, where PPLM outperforms ESM2 in 4,491 out of 5000 cases (89.8%), while ESM2 achieves lower perplexity in only 419 (8.4%) cases.

To further examine PPLM's sensitivity to interfacial residues, we evaluated perplexity under an interface masking strategy using PDB-derived protein pairs. In this setup, only residues known to

participate in protein–protein interfaces were masked. Here, two masking modes were defined for heteromeric pairs: the Both mode, which masks interface residues from both sequences, and the Single mode, which masks interface residues from only one sequence. For homomeric pairs, due to sequence identity, only the Both mode was applicable.

As shown in **Figure 2A**, PPLM achieves an average perplexity of 8.79 and a median 7.59 in the 'Both' mode, significantly lower than those of ESM2 (10.60 and 10.01), with a Wilcoxon signed-rank test p-value of 5.96E-251. In the 'Single' mode, PPLM again outperforms ESM2, with the average and median perplexities of 8.24 and 6.34, compared to 10.72 and 10.36 for ESM2 (with p-value = 4.35E-237). **Figure 2C** further displays the head-to-head comparison of perplexity under the two modes, where PPLM demonstrated a lower perplexity than ESM2 on 79.3% cases in the 'Both' mode and 78.8% cases in the 'Single' mode.

Interestingly, in heteromeric pairs, both the average and median perplexities of PPLM under the 'Single' mode were lower than those under the 'Both' mode, an effect not observed with ESM2. This suggests that PPLM is capable of capturing meaningful inter-protein interface signals and effectively leveraging the partner protein's context to recover masked interface residues.

Overall, these results highlight PPLM's improved sensitivity to residue-level interactions at protein–protein interfaces. This advantage is primarily attributed to its incorporative training framework for paired sequences, augmented by a hybrid attention network architecture which explicitly captures intra- and inter-protein interaction patterns. Such capabilities offer important benefits for downstream applications, including interface residue identification and structure-guided interface design, such as antibody CDR optimization.

PPLM-PPI outperforms existing methods in binary protein interaction prediction

To examine the effectiveness of PPLM-PPI on protein–protein interaction predictions, we tested it on five PPI datasets across five species, including *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, and *E. coli*. PPLM-PPI was compared against three state-of-the-art LM-based PPI predictors, including ESMDNN-PPI[22], D-SCRIPT[15], and Topsy-Turvy[21]. To ensure reproducibility, all methods were installed and executed locally, except for ESMDNN-PPI, which lacks publicly available source code, and we reproduced it locally based on the implementation details provided in its original publication.

The area under the precision-recall curve (AUPRC) is a performance metric that emphasizes the model's ability to correctly identify positive samples, which is particularly challenging for the imbalanced class distributions such as genome-wide PPI identifications. **Figure 3A** summarizes the AUPRC of PPLM-PPI against the three control methods. PPLM-PPI achieves AUPRC scores of 0.895, 0.880, 0.854, 0.708, and 0.759 on the five test species, corresponding to improvements of 11.7%, 6.7%, 7.4%, 18.8%, and 11.3% over ESMDNN-PPI; 54.3%, 59.1%, 55.8%, 74.8%, and 32.0% over D-SCRIPT; and 40.9%, 35.6%, 38.6%, 63.9%, and 86.9% over Topsy-Turvy, respectively. These results demonstrate the superior precision-recall performance of PPLM-PPI in distinguishing true PPIs across diverse species.

To test the balanced ability for recognizing both positive and negative PPI samples, **Figure 3B** reports the area under the receiver operating characteristic curve (AUROC) results, where PPLM-PPI achieves the highest AUROC score across all species among all methods. In **Figure 3C**, we further list the precision-recall curves of all methods across the five species datasets. PPLM-PPI consistently outperforms other methods across most recall thresholds. Meanwhile, the PPLM-PPI curves are noticeably smoother than those of other methods, reflecting its strong stability and predictive consistency, particularly in high-precision or high-recall regions. **Table S1** provides a summary of different metrics, including precision, recall, accuracy, F1-score, AUROC, and AUPRC, across all species, confirming the overall superiors of PPLM-PPI over all control methods in nearly every metric. This consistent robustness across diverse species and evaluation metrics underscores the potential of PPLM-PPI for real-world, high-confidence PPI discovery applications.

To elucidate the contributions of different components within PPLM-PPI, we conducted ablation experiments by replacing the pooling module and selectively removing specific features generated by PPLM. These experiments were designed to assess how each component impacts interaction prediction performance (**Tables S2 and S3**). As shown in **Figure S1A**, the choice of pooling strategy had a substantial impact on model performance. Among the single pooling approaches, the mean pooling consistently delivered the best results, while min pooling produced the worst performance across species due to its sensitivity to weak or noisy signals. Notably, a combined strategy integrating both mean and max pooling achieved the most robust and effective overall performance. We further investigated the relative contributions of the three feature types: embedding, intra-protein attention, and inter-protein attention (**Figure S1B**). Among them, embedding features contributed most substantially to predictive performance, likely due to their ability to capture both intra- and inter-protein contextual information. Incorporating all three feature types resulted in the highest overall accuracy, underscoring the complementary strengths of attention-based and embedding-based representations in modeling protein interaction propensity. These findings highlight the importance of both robust pooling strategies and multi-view feature integration in capturing the complex determinants of protein–protein interactions. In particular, the synergy between embedding and attention-based representations proved critical for achieving high predictive accuracy in cross-species PPI tasks.

PPLM-Affinity achieves superior performance in protein binding affinity prediction

We assessed the performance of PPLM-Affinity on the recently constructed PPB-Affinity dataset[27], a large and comprehensive benchmark for protein–protein binding affinity prediction. Evaluation was conducted using five-fold cross-validation, following the original protocol defined for the benchmark. To examine the effectiveness of language models in this task, we implemented ESM2-Affinity, where the receptor and ligand sequences were concatenated and used as input to a model trained with the same architecture and fine-tuning procedure as PPLM-Affinity. In addition, we also compared PPLM-Affinity with the PPB-Affinity model, which was trained by Liu et al[27] using the original PPB-Affinity dataset built on structural features of the protein–protein complexes.

The left panel of **Figure 3D** presents the Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SRCC) between the experimentally measured and predicted binding affinities (ΔG) for PPLM-Affinity, ESM2-Affinity, and PPB-Affinity across the entire dataset. The corresponding average values are summarized in **Table S4**. These two metrics are complementary: PCC quantifies the strength of linear relationships between predicted and experimental values, while SRCC assesses the strength of monotonic relationships based on their rank order. PPLM-Affinity achieves a PCC of 0.764 and an SRCC of 0.752, representing improvements of 14.2% and 15.9% over ESM2-Affinity (0.669 and 0.649), and 8.99% and 8.83% over the structure-based PPB-Affinity model (0.701 and 0.691), respectively.

Figure S2A further illustrates the distribution of prediction errors relative to experimental ΔG values. All three models show approximately Gaussian error distributions centered around zero, suggesting stable predictive behavior with limited systematic bias. Notably, the error distribution of PPLM-Affinity is more tightly centered around zero, with the lowest standard deviation $\sigma = 1.898$, compared to 2.179 by ESM2-Affinity and 2.083 by PPB-Affinity, indicating smaller and more consistent prediction errors across the dataset.

Figures 3E and **3F** display scatter plots of predicted versus experimental ΔG values for PPLM-Affinity and ESM2-Affinity, respectively. While both models exhibit positive correlations with experimental measurements, consistent with **Figure 3D**, the predictions from PPLM-Affinity lie closer to the diagonal line, reflecting greater accuracy. This is further supported by a lower RMSE (1.898 vs. 2.179), indicating improved predictive precision and reduced variance. Given that both models share the same architecture and training pipeline, the substantial gains achieved by PPLM-Affinity highlight the advantage of PPLM as a language model specifically optimized for protein–protein interactions.

Furthermore, its superior performance over the structure-based PPB-Affinity model underscores its capability to accurately infer binding affinities directly from sequence, providing a scalable and structure-independent alternative for affinity prediction.

The middle and right panels of **Figure 3D** further evaluate performance on two biologically significant PPI subgroups: antibody–antigen and TCR–pMHC, which are critical for immunological recognition and therapeutic development[43, 44]. On the antibody–antigen subset, PPLM-Affinity achieves PCC and SRCC values of 0.548 and 0.553, representing improvements over 95% compared to ESM2-Affinity (0.280 and 0.281), and 66.6% and 85.6% over PPB-Affinity (0.329 and 0.298). On the TCR–pMHC subset, PPLM-Affinity achieves PCC and SRCC values of 0.565 and 0.505, corresponding to gains of 54.0% and 60.3% over ESM2-Affinity (0.367 and 0.315), and 27.5% and 11.7% over PPB-Affinity (0.443 and 0.452), respectively.

Figures S2B/D and **S2C/E** display scatter plots of predicted versus experimental ΔG values for the antibody–antigen and TCR–pMHC subsets, respectively. In both cases, PPLM-Affinity produces a more pronounced upward trend, with predictions tightly clustered around the diagonal, indicating a stronger alignment with experimental values. In contrast, ESM2-Affinity shows a more diffuse pattern, with notably higher RMSE values (1.683 and 1.450 for antibody–antigen and TCR–pMHC, respectively, compared to 1.924 and 1.690 for PPLM-Affinity). These results highlight PPLM-Affinity's improved robustness and accuracy in modeling both general and immunologically complex interactions, suggesting its potential utility in antibody engineering, TCR optimization, and other sequence-based immunotherapeutic design tasks.

PPLM-Contact outperforms existing methods in inter-protein contact prediction

The performance of PPLM-Contact is evaluated on four benchmark test sets: Homodimer300 and Heterodimer99 from DeepInter[20], and CASP_Homodimer43 and CASP_Heterodimer20 collected from CASP13-16 experiments. It was compared against four state-of-the-art methods: DeepInter[20], CDPred[19], GLINTER[34], and DeepHomo2.0[31]. To access reliability, we measured inter-protein contact precision at various cutoffs of top 1, 10, 50, $L/10$, $L/5$, and L inter-protein contact precision sorted by the confidence score, where L is the length of the shorter protein chain in the complex. A summary of the result is provided in Supplementary **Table S5–6**, with detailed outcomes available in Supplementary **Data 1–2**. To ensure reproducibility, all comparison methods were installed and executed locally.

Figure 4A presents the top L contact precision of each method on homodimer test sets, using experimental or AlphaFold2-predicted monomer structures. On Homodimer300 with experimental monomer structures, PPLM-Contact achieves a top L contact precision of 77.8%, outperforming DeepInter (68.9%), CDPred (63.0%), DeepHomo2.0 (51.4%), and GLINTER (34.5%) by margins of 12.9%, 23.5%, 51.4%, and 125.5%, respectively. These improvements are statistically significant, with p-values of 3.65E-21, 5.58E-31, 3.17E-43, and 1.30E-47 based on the Wilcoxon signed-rank test. When using AlphaFold2-predicted monomer structures, PPLM-Contact maintains strong performance with a top L precision of 66.6%, outperforming the second-best method, DeepInter, by 10.4% (p-value = 4.07E-13). A similar trend is observed on the CASP_Homodimer43 test set, where PPLM-Contact achieves a precision of 65.2% when using experimental monomer structures, outperforming DeepInter by 6.0%, CDPred by 46.8%, DeepHomo2.0 by 90.6%, and GLINTER by 157.7%. When using AlphaFold2-predicted monomer structures, the corresponding improvements are 13.0%, 46.3%, 68.2%, and 140.0%, respectively.

Figure 4B further lists the top L contact precision of each method on heterodimer test sets. On the Heterodimer99 test set, PPLM-Contact achieves a top L contact precision of 48.0% using experimental monomer structures and 43.5% with AlphaFold2-predicted monomers. These represent improvements of 37.1% and 32.2% over DeepInter, 91.2% and 58.2% over CDPred, and 177.5% and 158.9% over GLINTER, respectively. All improvements are statistically significant supported by Wilcoxon signed-rank tests (p-values reported in **Table S5**). On the CASP_Heterodimer20 test set, PPLM-Contact

achieves a top L contact precision of 46.8% with experimental monomer structures and 40.6% with AlphaFold2-predicted monomers, which are also significantly higher than that of all control methods (Left panel of **Figure 4B**).

Figures 4C provide head-to-head comparisons between PPLM-Contact and other methods across the four benchmark test sets using experimental monomer structures. On the Homodimer300 test set, PPLM-Contact achieves the highest top L contact precision among all methods on 239 out of 300 proteins (79.7%) and leads on 30 out of 43 proteins in the CASP_Homodimer43 test set. For the Heterodimer99 test set, PPLM-Contact outperforms DeepInter, CDPred, and GLINTER on 54, 61, and 72 out of 99 proteins, respectively, while underperforming only on 28, 22, and 15 proteins, and showing equal precision on the remaining cases. On CASP_Heterodimer20 test set, PPLM-Contact surpasses the three methods on 16, 14, and 16 proteins, respectively, with inferior performance observed on only 3, 5, and 2 cases. **Figure S3** list head-to-head comparisons based on AlphaFold2-predicted monomer structures, showing a similar trend which PPLM-Contact outperforms the control methods in the majority of cases.

To evaluate the contribution of individual feature types, we trained three ablated variants of PPLM-Contact, each excluding one key component while keeping all other settings unchanged: (i) PPLM-Contact w/o MSA, which excludes MSA-derived features (PSSM, DCA, and ESM-MSA-1b outputs); (ii) PPLM-Contact w/o PPLM, which removes inter-protein attention features generated by PPLM; and (iii) PPLM-Contact w/o Mdist, which omits monomer distance maps extracted from either experimental or AlphaFold-predicted monomer structures. **Figure 4D** illustrates the changes in top 1 to L contact precisions for these variants on Homodimer300 and Heterodimer99 test sets using experimental monomer structures. The removal of monomer distance maps led to the most substantial drop, the top L contact precision decrease by 31.2% and 30.8% for homodimers and heterodimers, respectively, underscoring the critical role of monomer geometry. Excluding MSA-derived features led to a precision drop of 9.7% and 18.7%, reflecting the valuable contribution of co-evolutionary information. Conversely, incorporating PPLM-generated features boosted top L contact precision by 6.6% and 6.1%, demonstrating PPLM's effectiveness in capturing essential inter-protein geometric relationships. The results based on AlphaFold2-predicted monomer structures are also listed in **Table S7** and **Datasets 3–4**, which demonstrate similar trend as observed for that on experimental monomer structures. These comparisons emphasize the complementary nature and essential roles of each feature type in accurate inter-protein contact prediction.

As an illustration, **Figure 4E-F** showcase an example of the EF-hand domain of human RASEF (PDB ID: 2PMY), which consists of two identical protein chains, each with 91 amino acids, 72 of which have experimentally resolved structures. When using only MSA-derived features and monomer structures, similar to most existing methods, PPLM-Contact w/o PPLM yields largely incorrect top N contact predictions (where N equals the number of ground-truth contacts), resulting in a low contact precision of 16.7% (20.8% for top L contacts). In contrast, replacing the MSA-derived features with PPLM-generated features allows the PPLM-Contact w/o MSA to correctly identify 201 out of 257 ground-truth contacts (78.2%), achieving 100% precision for top L contacts. When combining both MSA and PPLM features, the full PPLM-Contact model further improves top N contact precision to 81.3%, while maintaining 100% precision for top L contacts. The result of this example highlights again the effectiveness of PPLM in capturing underlying geometric patterns and interface-level spatial relationships, which are crucial for accurate inter-protein contact prediction.

PPLM-Contact enhances interface residue recognition

In addition to protein–protein contact prediction, recognizing what residues participate in the PPIs is of practical usefulness in determining complex formation, signaling, and biological activity[45]. In **Figure 5**, we report the precision of PPLM-Contact for interface residue recognitions, compared to the same set of four control methods, based on experimentally determined monomer structures across 343 homodimer and 119 heterodimer test proteins (see **Figure S4** for results using AlphaFold2-predicted

monomer structures). Here, we defined interface residues as those whose heavy atoms are within 8 Å of any heavy atom from the opposite chain in the experimental structure, where only the top *M* predictions are assessed with *M* being the number of ground-truth interface residues for each target, regardless of their specific contact partners (Supplementary **Data 5–6**).

As shown in **Figure 5A**, PPLM-Contact achieves the highest interface recognition precision of 0.860 for homodimers and 0.729 for heterodimers, exceeding DeepInter (0.810 and 0.633) by 6.3% and 15.2%, CDPred (0.805 and 0.607) by 6.9% and 20.2%, DeepHomo2.0 (0.720) by 19.5%, and GLINTER (0.651 and 0.498) by 32.2% and 46.3%, respectively. Wilcoxon signed-rank tests confirm the statistical significance of these performance improvements (**Table S8**). **Figure 5B** further provides head-to-head comparisons of identification precision between PPLM-Contact and each control method on homodimer and heterodimer proteins, respectively. PPLM-Contact outperforms competing methods for 74.6%–92.7% homodimer proteins and 73.9%–90.8% heterodimer proteins. Compared to the second-best method (DeepInter), PPLM-Contact improves precision by more than 0.2 in 57 proteins (11.5%), while the corresponding numbers are 91 proteins (19.7%) for CDPred, 94 proteins (27.4%) for DeepHomo2.0, and 230 proteins (49.8%) for GLINTER.

Figures 5C showcases an example of the Crystal structure of DNMT3A-DNMT3L complex (PDB ID: 4U7P), which consists of two distinct chains with experimentally resolved structures for 426 and 197 amino acids, respectively. **Figure 5D** visually highlights the interface residues identified by each method on the complex structure, while **Figure 5E** further compares the positions of both ground-truth and predicted interface residues. It was found that PPLM-Contact successfully recognized 40 out of 43 (93.0%) interface residues in chain A and 40 out of 44 (90.9%) interface residue in chain B, demonstrating a substantial improvement over the control methods. In comparison, DeepInter identified 4 (9.3%) and 26 (59.1%) interface residues, CDPred identified 29 (67.4%) and 31 (70.5%), and GLINTER missed all interface residues in chain A and only identified 8 (18.2%) interface residues in chain B. Detailed analyses show that all the control methods showed an overprediction of the interface residues on the regions close to N-terminal for chain A, while for chain B the overpredictions are mostly located on C-terminal region. In contrast, PPLM-Contact demonstrated a much more balanced interface residue predictions along both chains A and B, resulting in an excellent overall precision of 0.92 for chains A and B which is significantly higher than that by DeepInter (0.35), CDPred (0.69), and GLINTER (0.09), respectively.

These results demonstrate that PPLM-Contact achieves superior performance in interface residue recognition compared to existing methods. This improvement can be primarily attributed to two key innovations. First, PPLM effectively captures the PPI signals that complement co-evolutionary information from MSA, providing reliable features for the deep learning model. Second, the inter-protein specific transformer network, designed with a parallel architecture, enables efficient learning of geometric interactions between residues across protein interfaces.

Enhancing LM-based interface contact prediction with predicted complex structures

Recent advances in protein complex structure prediction have enabled access to highly accurate models of protein assemblies, offering valuable insights into interfacial organization. These predicted structures provide an opportunity to inform interface prediction with rich geometric context. To leverage this progress, we developed PPLM-Contact2, an enhanced version of PPLM-Contact that incorporates inter-protein distance maps extracted from complex structure models predicted by AlphaFold2.3[39], AlphaFold3[40], and DMFold[41] as additional features. We evaluated PPLM-Contact2 against the state-of-the-art structure-based methods on inter-protein contact prediction (Supplementary **Table S9–11**, **Data 7–9**).

Figure 6A presents the top *L* contact precisions of PPLM-Contact2 in comparison with AlphaFold2.3, AlphaFold3, and DMFold across four benchmark test sets: Homodimer300, Heterodimer99, CASP_Homodimer43, and CASP_Heterodimer20. PPLM-Contact2 achieves precisions of 86.4%, 89.9%, 75.4%, and 79.1% on these test sets, respectively, which represent

consistent improvements over AlphaFold2.3 by 4.0%, 4.8%, 13.5%, and 28.4%, over AlphaFold3 by 8.1%, 4.4%, 17.0%, and 15.7%, and over DMFold by 5.5%, 8.3%, 7.3%, and 7.7%. The improvements are statistically significant, as confirmed by Wilcoxon signed-rank tests (p-values reported in **Table S9**). **Figure 6C–E** illustrate head-to-head comparisons between PPLM-Contact2 and the structure-based methods across all 462 test proteins, where PPLM-Contact2 achieves a higher precision on 408 proteins (88.3%) compared to AlphaFold2.3, on 400 proteins (86.6%) compared to AlphaFold, and on 412 proteins (89.2%) compared to DMFold. **Figure S5** presents the distributions of precision, recall, and F1-score for inter-protein contact prediction across all test proteins. Compared to the complex structure modeling methods, PPLM-Contact2 demonstrates a clear shift toward higher score ranges in all three metrics. It achieves an F1-score of 0.780, surpassing AlphaFold2.3 (0.748), AlphaFold3 (0.739), and DMFold (0.743) by 4.3%, 5.5%, and 5.0%, respectively, highlighting its robustness and consistency in predicting inter-protein contacts.

We further evaluated the ability of these methods to identify interface residues. **Figure 6B** presents the precision of interface residue identification by PPLM-Contact2 and the structure-based methods. Again, PPLM-Contact2 achieves the highest average and median precision for both homodimer and heterodimer proteins. Specifically, its average and median precision reached 0.897 and 0.962 for homodimers, and 0.904 and 0.945 for heterodimers, respectively. Wilcoxon signed-rank tests confirm that these improvements over the control methods are statistically significant (**Table S11**). **Figure S6** illustrate head-to-head comparisons of precision between PPLM-Contact2 and the control methods across all proteins, highlighting PPLM-Contact2's superior performance on most proteins.

These results indicate that, despite the remarkable progress in end-to-end complex structure prediction, additional gains can be obtained by incorporating structural models with paired language model-based pipelines to enhance inter-chain contact map and PPI binding sites predictions. This integrated information can be further used as spatial constraints to guide quaternary structure assembly simulations for more accurate protein complex structure predictions.

Discussion

Computational modeling of PPIs is essential for elucidating cellular mechanisms and accelerating therapeutic discovery. While protein language models offer a scalable solution for high-throughput PPI analysis, current models often fall short in capturing the inter-protein context and interaction-specific features critical for accurate prediction. To address these limitations, we developed PPLM, a protein–protein language model that learns directly from paired sequences through a novel attention architecture, enabling explicit modeling of inter-protein context. Built upon PPLM, we introduced three task-specific downstream methods, PPLM-PPI, PPLM-Affinity and PPLM-Contact, targeting three progressively key PPI tasks: binary interaction prediction, binding affinity estimation, and interface contact prediction (including inter-chain contact map and interface residue recognition).

A key advantage of PPLM over existing language models lies in its attention mechanism, which disentangles intra- and inter-protein positional encoding while incorporating learnable inter-protein weights and masking strategies. Coupled with training on a comprehensive paired-sequence dataset, these innovations allow PPLM to effectively model both intra- and inter-protein interaction patterns. Benchmarking results show that PPLM achieves significantly lower perplexity than ESM2 on both general residues and interface regions, and its versatility across downstream tasks, combined with customized architectures, leads to substantial performance improvements over existing approaches.

By leveraging both learned embeddings and intra-/inter-protein attention matrices, PPLM-PPI achieves substantial performance improvements in binary interaction prediction, surpassing existing protein language model-based methods by an average of over 11% across five test species. For binding affinity prediction, fine-tuning the final layer of PPLM enables PPLM-Affinity to outperform ESM2-Affinity by more than 14% in terms of Pearson and Spearman correlations with experimental measurements. Remarkably, on two biomedically critical subgroups of antibody–antigen and TCR–pMHC interactions, PPLM-Affinity achieves gains of 50% and 10%, respectively, over the best-

performing alternative methods. For inter-protein contact prediction, PPLM-Contact integrates the inter-protein attention matrices derived from PPLM with a specialized transformer architecture, achieving over 10% higher top L precision on homodimers and more than 30% improvement on heterodimers compared to existing approaches. When further enhanced with inter-protein distance features extracted from predicted complex structures, the extended model PPLM-Contact2 consistently outperforms the cutting-edge structure-based methods, including AlphaFold2.3[39], AlphaFold3[40], and DMFold[41], in inter-chain contact prediction.

While this study demonstrates the effectiveness of PPLM across key PPI-related tasks, its potential applications extend beyond the scope evaluated here. One promising direction involves integrating PPLM's interface-aware representations with AI-driven structural prediction tools such as AlphaFold3[40] to improve the modeling of protein complex assemblies. In addition, the transferability of PPLM to antibody–antigen and TCR–pMHC systems merits further exploration. Given its lower perplexity at interaction interfaces relative to existing language models, PPLM may serve as a valuable foundation for complementarity-determining region (CDR) modeling and design. Its ability to capture fine-grained interface features may support the accurate modeling of binding interactions, thereby facilitating the development and optimization of therapeutic antibodies and TCRs.

Despite its strong performance, PPLM has some limitations that warrant further investigation. Its training relies on a composite PPI dataset that, while diverse, may not fully capture the breadth of protein interactions across organisms and cellular states. In particular, transient, weak, or condition-specific interactions remain underrepresented, potentially limiting generalizability to less common PPI modes. Moreover, PPLM focuses primarily on modeling sequence-derived representations, without explicitly incorporating biophysical constraints or functional annotations. Incorporating complementary signals—such as thermodynamics, mutational effects, or co-evolutionary cues—may further improve predictive accuracy and interpretability, particularly for context-specific or mechanistically complex interactions.

Methods

Datasets

To train PPLM, we constructed a composite protein dataset of protein interaction sequence pair by integrating protein complex structures from the Protein Data Bank (PDB)[37] and PPI sequences from the STRING[38] database. From the PDB, we extracted high-quality sequence pairs from protein complexes released before January 1, 2024, selecting only those in which at least one pair of residues from different chains had a Cβ–Cβ distance within 8Å. This yielded 273,295 heteromeric and 86,915 homomeric sequence pairs. These pairs are clustered using MMseqs2[46] at a 50% pairwise sequence identity threshold with 80% alignment coverage, resulting in 25,245 heteromeric and 23,082 homomeric clusters. To further expand the diversity and coverage of interaction types, we incorporated high-confidence interaction pairs from the STRING database, which contains over one billion protein–protein associations. For this, we selected 2,995,604 sequence pairs with the highest confidence scores (>0.999) and clustered them with the same sequence identity cutoff and coverage criteria used for clustering the PDB-derived pairs. Clusters redundant with those from the PDB were removed, yielding 629,045 clusters. For model validation, we randomly sampled one representative pair from each of 1,306 heteromeric, 1,194 homomeric, and 2,500 STRING-derived clusters to construct the validation set. The remaining clusters were used as the training set.

To train and evaluate PPLM-PPI for PPI prediction, we adopted benchmark datasets from D-SCRIPT[15] spanning six species: *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, and *E. coli*. Given the relative rarity of true PPIs, each dataset maintains a 10:1 ratio of negative to positive pairs. To ensure data integrity, we identified and removed duplicate, erroneous, and invalid samples arising from the random negative sampling process (see **Text S1** and **Figure S7**). The PPLM-PPI model was trained and validated on the *H. sapiens* dataset and tested on the remaining five species datasets. Detailed dataset statistics are listed in **Table S12**.

For protein–protein binding affinity prediction, we adopted the PPB-Affinity dataset[27], a large and curated resource compiled from multiple public affinity databases. The dataset provides experimentally determined binding affinities, crystal structures, mutation patterns, and annotations of receptor and ligand chains. It comprises 12,062 interaction samples from 3,032 distinct PDB entries, encompassing subgroups such as antibody–antigen and TCR–pMHC complexes. Notably, 24.3% of entries feature single-point mutants, while 21.1% include multiple mutation types, offering broad coverage of mutational landscapes relevant to affinity modeling. PPLM-Affinity was trained and evaluated using five-fold cross-validation, following the benchmark protocol evaluation settings established in the original PPB-Affinity study.

For inter-protein contact prediction, we adopted the same datasets as DeepInter[20] to ensure fair comparison with existing methods. These datasets consist of non-redundant homodimers and heterodimers curated from the PDB. The training set includes 3,504 homodimers and 1,881 heterodimers, while the validation set comprises 296 homodimers and 96 heterodimers. The test set contain 300 homodimers (Homodimer300) and 99 heterodimers (Heterodimer99). To account for structural differences between homodimers and heterodimers, we trained separate models: the homodimer models was trained exclusively on homodimers, whereas the heterodimer model was trained on the full combined training set. To further assess robustness and real-world applicability, we assessed PPLM-Contact on an independent test set comprising 43 homodimer targets (CASP_Homodimer43) and 20 heterodimer targets (CASP_Heterodimer20) collected from CASP13 to CASP16 (see Supplementary Data 1 and 2).

Protein–protein language model

PPLM is a pretrained protein–protein language model that can capture potential protein interaction information by providing informative representations of protein pairs (**Figure 1A**). Given a protein sequence pair as input, the sequences are first tokenized independently and then concatenated with special beginning-of-sequence (BOS) and end-of-sequence (EOS) tokens to explicitly mark the boundaries of each chain. The concatenated initial representations will be used as the input of a stack of 33 serially connected transformer blocks, with each block consisting of a tailored multi-head attention module and a feed-forward network.

In the multi-head attention module, we introduce a novel attention mechanism to capture both intra- and inter-protein residue interactions. To distinguish these two interaction types, two separate attention matrices are computed: one with rotary positional embeddings (RoPE)[36] for intra-chain residue pairs, and another without positional encoding (ROW) for inter-protein residue pairs. These two matrices are then integrated into a full-size attention matrix. To enhance PPLM's ability to capture diverse inter-protein interaction patterns, we introduce a set of learnable weights (W_{lh}) across transformer layers (l) and attention heads (h), applied specifically to the inter-protein attention components. Additionally, an inter-protein binary mask (M_{ij}) is employed to distinguish inter-protein from intra-chain residues pairs (i, j) and to better guide the model's focus on inter-protein interactions. Thus, the full attention matrix is defined as:

$$A_{lhij} = W_{lh} \cdot M_{ij} \cdot A_{lhij}^{row} + \left(1 - M_{ij}\right) \cdot A_{lhij}^{rope} + M_{ij} \tag{1}$$

where A_{lhij}^{row} and A_{lhij}^{rope} represent the attention matrices computed without and with RoPE, respectively, and M_{ij} is a binary mask matrix with value 1 for inter-protein residue pairs and 0 for intra-protein pairs.

The masking strategy plays a pivotal role in masked language model training. In PPLM, we adopt distinct masking strategies for sequence pairs derived from different data resources. For high-quality sequence pairs from the PDB database, greater emphasis is placed on interface residues: 30% of interface residues are randomly masked, compared to 15% of non-interface residues. In contrast, for sequence pairs from the STRING database, a uniform masking strategy is applied, where 15% of residues in each sequence are randomly selected for masking. To accommodate GPU memory

limitations, sequence pair with full length exceeding 1,024 residues are cropped. For sequence pairs from PDB database, cropping is biased toward retaining fragments include most interface residues. For sequence pairs from STRING database, continuous fragments are cropped from each sequence, with fragment lengths proportional to the original sequence lengths.

The final objective function is a weighted average of cross-entropy losses computed separately over masked residues in each sequence:

$$\mathcal{L}_{MLM} = -\frac{1}{|M_1|}\sum_{i\in M_1}\log p(x_i^{(1)}|x\backslash M) - \frac{1}{|M_2|}\sum_{i\in M_2}\log p(x_i^{(2)}|x\backslash M) \quad (2)$$

where $x^{(1)}$ and $x^{(2)}$ represent the amino acid sequences of the two proteins, and $x = (x^{(1)}, x^{(2)})$ represents all residues from both sequences. M_1 and M_2 denote the sets of masked residue positions in each sequence, and $M = M_1 \cup M_2$ denotes all masked positions across both sequences. The notation $x\backslash M$ refers to the unmasked residues from both sequences.

During the training procedure, sequence pairs are sampled from the PDB and STRING clusters at a 1:2 ratio. PPLM was initialized from the 650M-parameter ESM2 model[6] and trained using four NVIDIA A100 GPUs with a gradient accumulation step of 32, totaling 6.4 million training steps. Optimization was performed using the AdamW optimizer with exponential decay rates β_1=0.9 and β_2=0.98. The learning rate was linearly warmed up to 1×10^{-6} over the first 2,000 steps, followed by a linear decay to 5×10^{-7} over the remainder of the training.

Interaction prediction using PPLM

Building upon PPLM, we developed PPLM-PPI for protein–protein interaction prediction by leveraging the protein pair representations generated by PPLM (**Figure 1B**). Given a pair of input protein sequences, PPLM is first used to extract several features: the embeddings of each sequence, the intra-protein attention matrices for both sequences, and the inter-protein attention matrix between sequences. For each feature, max pooling and mean pooling are independently applied along the sequence dimension to obtain compact representations. These representations are then projected into a unified dimensional space using separate linear layers and subsequently concatenated into a single joint representation. An MLP, followed by a sigmoid activation layer, is then used to predict the interaction probability. The MLP consists of five linear layers, with each of the first four layers followed by layer normalization and a ReLU activation. The dimensionality parameters used at each transformation step are summarized in **Tables S13–14**.

PPLM-PPI was trained using 10-fold cross-validation on the *H. sapiens* dataset. The data was randomly divided into ten subsets, ensuring a balanced ratio of positive and negative samples in each. For each fold, one subset was held out for validation, while the remaining nine were used for training. Models were trained for 12 epochs per fold, and the checkpoint with the highest AUPRC on the validation set was selected. The five best-performing models across all folds were then ensembled to generate final predictions during inference on the test sets. Training was performed on a single NVIDIA A100 GPU with a batch size of 32. The model was optimized using the AdamW optimizer with a learning rate of 5×10^{-5}. The loss function was binary cross-entropy between predicted probabilities and ground-truth interaction labels, defined as:

$$\mathcal{L}_{PPI} = -(y\cdot\log p + (1-y)\cdot\log(1-p)) \quad (3)$$

where $y \in \{0, 1\}$ denotes the ground-truth interaction label, with 1 for positive and 0 for negative, and $p \in (0, 1)$ is the predicted probability of interaction.

Binding affinity prediction using PPLM

We further developed a sequence-based method, PPLM-Affinity, for protein–protein binding affinity prediction by fine-tuning the final layer (i.e., the last transformer block) of PPLM on a curated affinity dataset (**Figure 1C**). Given the limited size of the dataset—many samples involve mutations and include multi-chain receptor or ligand proteins, especially within antibody–antigen and TCR–

pMHC subgroups—we fine-tuned the final layer to better adapt PPLM's contextual representations for capturing sequence-level determinants of molecular interaction strength. The full-length embeddings output by the fine-tuned PPLM are aggregated along the sequence dimension using max pooling, and the resulting vector is passed through two fully connected layers with an intermediate ReLU activation to predict the binding affinity value. In cases where the receptor or ligand consists of multiple chains, their sequences are concatenated into a single sequence prior to input into PPLM.

PPLM-Affinity was trained and evaluated using five-fold cross-validation. To prevent data leakage, samples sharing the same PDB ID were grouped into the same fold. Each fold was trained for 15 epochs on a single NVIDIA A100 GPU. Optimization was performed using the AdamW optimizer with a learning rate of 1×10^{-4}, and the loss function was the mean squared error (MSE) between predicted and experimentally determined binding affinities:

$$\mathcal{L}_{Affinity} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{3}$$

where y_i and $\hat{y}_i$ represent the ground-truth and predicted binding affinity, respectively; N is the batch size, which equals 1 in our setting.

Inter-protein contact prediction using PPLM

In addition to predicting interaction and binding strength, we further developed PPLM-Contact for inter-protein contact prediction and identify interaction interfaces. As illustrated in **Figure 1D**, PPLM-contact consists of two main components: (1) feature extraction from PPLM, MSA, and monomer structures; and (2) a novel inter-protein transformer architecture incorporating parallel triangle multiplication, cross-attention and self-attention mechanisms.

For each input protein pair, features are extracted from three sources. First, PPLM is used to create the inter-protein attention matrix, which captures cross-chain dependencies and serves as a key inter-protein feature. Second, MSAs for each monomer sequence is collected by searching against the UniRef30_2021_03[47] database using HHblits[48] with hyperparameters set to '-e 1e-3 -cov 0.4 -id 99'. For heterodimers, paired MSAs are constructed based on species annotation[41], while for homodimers, the monomer MSA is directly reused as the paired MSA. A variety of features are then extracted from both monomer and paired MSAs, including position-specific scoring matrices (PSSMs)[49], row scores and average product corrections from direct coupling analysis (DCA)[50], as well as MSA embeddings and attention matrices derived from ESM-MSA-1b[51]. Lastly, the monomer distance maps are extracted from either experimentally determined (when available) or predicted monomer structures (by AlphaFold2[52]). The dimensions of all features are detailed in **Table S15**. For the enhanced version, PPLM-Contact2, inter-protein distance map extracted from the complex structure predicted by AlphaFold2.3 is incorporated as an additional feature.

In the inter-protein transformer architecture, two ResNet modules are used to independently encode intra-protein and inter-protein features before passing into the inter-protein transformer blocks (**Figure S8**). The intra-protein features of both sequences are processed by a shared intra-protein ResNet module with tied parameters. This is followed by 12 inter-protein transformer blocks, each containing two parallel triangle multiplication modules, two parallel cross-attention modules, two parallel self-attention modules, and a transition layer (see middle box of **Figure 1D**).

The triangle multiplication module is designed to update the inter-protein pair representation by integrating information from both intra- and inter-protein pairs (**Figure S9**). Specifically, for a given inter-protein residue pair z_{ij}, the module aggregates signals from an intermediate residue k, utilizing intra-chain interactions r_{ik} and cross-chain interactions z_{kj}, while ensuring geometric triangle constraints are maintained. A parallel operation is simultaneously applied to both proteins, allowing inter-protein features to be updated using intra-protein information from each chain. The detailed update equations are:

$$z'_{ij} = \mathcal{L}(z_{ij}) \odot \sigma\left(\mathcal{L}(z_{ij})\right);\ r'_{ik} = \mathcal{L}(r_{ik}) \odot \sigma(\mathcal{L}(r_{ik}));\ \ g_{ij} = \sigma\left(\mathcal{L}(z_{ij})\right) \tag{4}$$

$$\tilde{z}_{ij} = g_{ij} \odot \varphi\left(\sum_{k=1}^{L} r'_{ik} \cdot z'_{ij}\right) \tag{5}$$

where z_{ij} and r_{ik} are the inter- and intra-protein pair representations, respectively, and $\tilde{z}_{ij}$ is the updated z_{ij} based on one protein; $\mathcal{L}$ presents a linear transformation, σ the sigmoid activation, and φ a LayerNorm followed by a linear transformation. L is the monomer sequence length. Two triangle multiplication operations, with one per protein, are performed in parallel, and the resulting outputs are combined to update the original z_{ij}.

The cross-attention module is designed to update the inter-protein pair representation z_{ij} by attending to all intra-protein pair r_{in} that share residue *i* (or r_{jn} that share residue *j*), thereby capturing rich contextual information from the corresponding monomer protein (**Figure S10**). In this mechanism, the inter-protein pair representation acts as the query, while the intra-protein pairs serve as the key and value in the attention operation, i.e.,

$$q_{ij}^{h} = \mathcal{L}(z_{ij});\ \ k_{in}^{h} = \mathcal{L}(r_{in});\ v_{in}^{h} = \mathcal{L}(r_{in}); b_{ij}^{h} = \mathcal{L}(z_{ij}); g_{ij}^{h} = \sigma\left(\mathcal{L}(z_{ij})\right) \tag{6}$$

$$a_{ijk}^{h} = \mathrm{softmax}\left(\frac{1}{\sqrt{c}} q_{ij}^{h} \cdot k_{in}^{h} + b_{ij}^{h}\right) \tag{7}$$

$$\tilde{z}_{ij} = \mathcal{L}\left(g_{ij}^{h} \odot \sum_{n=1}^{L} a_{ijk}^{h} \cdot v_{in}^{h}\right) \tag{8}$$

Two parallel cross-attention operations are performed across two proteins according to the equations above, and the resulting updates are then combined to update the inter-protein pair representation.

A self-attention module is applied to further update the inter-protein representation from its own context (**Figure S11**). Specifically, it attends over all inter-protein pairs z_{in} that share residue *i* (or z_{jn} that share residue *j*) to update z_{ij}, thereby enabling long-range information flow across the protein interface. During this process, intra-protein distances extracted from monomer structure are transformed using a Gaussian kernel and used to modulate the attention weights[20]. The detailed formulation is:

$$q_{ij}^{h} = \mathcal{L}(z_{ij});\ k_{in}^{h} = \mathcal{L}(z_{in});\ v_{in}^{h} = \mathcal{L}(z_{in});\ g_{ij}^{h} = \sigma\left(\mathcal{L}(z_{ij})\right) \tag{9}$$

$$a_{ijk}^{h} = \mathrm{softmax}\left(\frac{1}{\sqrt{c}} q_{ij}^{h} \cdot k_{in}^{h} \cdot d_{in}\right) \tag{10}$$

$$\tilde{z}_{ij} = \mathcal{L}\left(g_{ij}^{h} \odot \sum_{n-1}^{L} a_{ijk}^{h} \cdot v_{in}^{h}\right) \tag{11}$$

where d_{in} denotes the intra-protein distance transformed by the Gaussian kernel function. Parallel self-attention operations are performed independently along two monomer protein dimensions and then aggregated to update the inter-protein pair representation z_{ij}. Finally, a transition layer consisting of two linear transformations is applied at the end of each transformer block.

PPLM-Contact was trained on a single NVIDIA A100 GPU for a total of 100 epoch using the AdamW optimizer. The initial learning rate was set to 1×10^{-3} and decayed by a factor of 0.98 after each epoch. To handle long sequences during training, monomer proteins longer than 256 residues were cropped by randomly selecting a contiguous 256-residue fragment that contained the highest number of interface residues. Inter-protein residue pairs were defined as contacts if the minimum

distance between their heavy atoms was less than 8Å[53]. The top five models with the highest top L inter-protein contact precision on the validation set were then used for ensemble prediction during inference. Given the severe class imbalance between contact and non-contact residue pairs, we employed the Focal Loss function[54] to mitigate the influence of dominant non-contact pairs:

$$\mathcal{L}_{Contact} = -\frac{1}{L_1 L_2}\sum_{i=1}^{L_1}\sum_{j=1}^{L_2}\alpha\left(1-p_{ij}\right)^{\gamma}\log\left(p_{ij}\right) \tag{12}$$

where p_{ij} is the predicted contact probability; $\alpha = 0.25$ is a balance factor, and $\gamma = 1.5$ is a focusing factor; L_1 and L_2 are the protein lengths of the two monomer proteins, respectively.

Data availability

The dataset used in this study is deposited at both https://zhanggroup.org/PPLM/benchmark.html and our GitHub repository: https://github.com/junliu621/PPLM/tree/main/data/. Protein pairs used to train PPLM are collected from the Protein Data Bank (PDB, https://www.rcsb.org/) and STRING database (https://string-db.org/). The interaction dataset for PPLM-PPI is sourced from D-SCRIPT (https://github.com/samsledje/D-SCRIPT/), while the binding affinity dataset for PPLM-Affinity is taken from PPB-Affinity (https://github.com/ChenPy00/PPB-Affinity). The Uniref30_2021_03 database is available at https://wwwuser.gwdguser.de/~compbiol/uniclust/.

Code availability

The PPLM webserver and source codes, including the language model, interaction prediction, binding affinity prediction, and inter-protein contact prediction, are freely accessible at https://zhanggroup.org/PPLM/. The source codes are also available on GitHub at https://github.com/junliu621/PPLM.

Acknowledgements

We thank Dr. Yang Li for discussions. This work is supported in part by Singapore Ministry of Education (T1 251RES2309), National University of Singapore Startup Grants (#A-8001129-00-00, #A-0009651-30-00), Cancer Science Institute of Singapore, and Natural Science Foundation of Ningxia Province 2023AAC05036. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Y.Z. conceived and designed the project and supervised the work. J.L. developed the methods and performed. J.L and H.C. prepared the initial data, performed the benchmark and analyzed the data. J.L and Y.Z wrote the paper. All authors wrote the manuscript and approved the final version.

Reference

1. Barabasi, A.-L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nature reviews genetics* **5**, 101-113 (2004).
2. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. Principles of protein− protein interactions: what are the preferred ways for proteins to interact? *Chemical reviews* **108**, 1225-1244 (2008).
3. Siebenmorgen, T. & Zacharias, M. Computational prediction of protein–protein binding affinities. *Advanced Review* **10**, e1448 (2019).
4. Vangone, A. & Bonvin, A.M. Contacts-based prediction of binding affinity in protein–protein complexes. *elife* **4**, e07454 (2015).
5. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**,

e2016239118 (2021).
6. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
7. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024.2007. 2001.600583 (2024).
8. Elnaggar, A. et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **44**, 7112-7127 (2021).
9. Chowdhury, R. et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology* **40**, 1617-1623 (2022).
10. Fang, X. et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nature Machine Intelligence* **5**, 1087-1096 (2023).
11. Wang, S., You, R., Liu, Y., Xiong, Y. & Zhu, S. NetGO 3.0: protein language model improves large-scale functional annotations. *Genomics, Proteomics & Bioinformatics* **21**, 349-358 (2023).
12. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications* **13**, 4348 (2022).
13. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems* **34**, 29287-29303 (2021).
14. Jin, M. et al. ProLLM: protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction. *bioRxiv*, 2024.2004. 2018.590025 (2024).
15. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems* **12**, 969-982. e966 (2021).
16. Liu, J., Liu, D., He, G. & Zhang, G. Estimating protein complex model accuracy based on ultrafast shape recognition and deep learning in CASP15. *Proteins: Structure, Function, and Bioinformatics* **91**, 1861-1870 (2023).
17. Zhou, Z. et al. ProAffinity-GNN: A Novel Approach to Structure-Based Protein–Protein Binding Affinity Prediction via a Curated Data Set and Graph Neural Networks. *Journal of Chemical Information and Modeling* **64**, 8796-8808 (2024).
18. Romero-Molina, S. et al. PPI-affinity: A web tool for the prediction and optimization of protein–peptide and protein–protein binding affinity. *Journal of proteome research* **21**, 1829-1841 (2022).
19. Guo, Z., Liu, J., Skolnick, J. & Cheng, J. Prediction of inter-chain distance maps of protein complexes with 2D attention-based deep neural networks. *Nature Communications* **13**, 6963 (2022).
20. Lin, P., Tao, H., Li, H. & Huang, S.-Y. Protein–protein contact prediction by geometric triangle-aware protein language models. *Nature Machine Intelligence* **5**, 1275-1284 (2023).
21. Singh, R., Devkota, K., Sledzieski, S., Berger, B. & Cowen, L. Topsy-Turvy: integrating a global view into sequence-based PPI prediction. *Bioinformatics* **38**, i264-i272 (2022).
22. Li, Y., Wang, C., Gu, H., Feng, H. & Ruan, Y. ESMDNN-PPI: a new protein–protein interaction prediction model developed with protein language model of ESM2 and deep neural network.

Measurement Science and Technology **35**, 125701 (2024).
23. Meda, R.S. & Farimani, A.B. BAPULM: Binding Affinity Prediction using Language Models. *arXiv preprint arXiv:2411.04150* (2024).
24. Gorantla, R. et al. Learning Binding Affinities via Fine-tuning of Protein and Ligand Language Models. *bioRxiv*, 2024.2011. 2001.621495 (2024).
25. Siebenmorgen, T. & Zacharias, M. Computational prediction of protein–protein binding affinities. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **10**, e1448 (2020).
26. Guo, Z. & Yamaguchi, R. Machine learning methods for protein-protein binding affinity prediction in protein design. *Frontiers in Bioinformatics* **2**, 1065703 (2022).
27. Liu, H. et al. PPB-Affinity: Protein-Protein Binding Affinity dataset for AI-based protein drug discovery. *Scientific Data* **11**, 1-11 (2024).
28. Xue, L.C., Rodrigues, J.P., Kastritis, P.L., Bonvin, A.M. & Vangone, A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics* **32**, 3676-3678 (2016).
29. Wang, M., Cang, Z. & Wei, G.-W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence* **2**, 116-123 (2020).
30. Zeng, H. et al. ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic acids research* **46**, W432-W437 (2018).
31. Lin, P., Yan, Y. & Huang, S.-Y. DeepHomo2.0: improved protein–protein contact prediction of homodimers by transformer-enhanced deep learning. *Briefings in Bioinformatics* **24**, bbac499 (2023).
32. Hu, L., Wang, X., Huang, Y.-A., Hu, P. & You, Z.-H. A survey on computational models for predicting protein–protein interactions. *Briefings in bioinformatics* **22**, bbab036 (2021).
33. Si, Y. & Yan, C. Protein language model-embedded geometric graphs power inter-protein contact prediction. *Elife* **12**, RP92184 (2024).
34. Xie, Z. & Xu, J. Deep graph learning of inter-protein contacts. *Bioinformatics* **38**, 947-953 (2022).
35. Rao, R.M. et al. in International Conference on Machine Learning 8844-8856 (PMLR, 2021).
36. Su, J. et al. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
37. Burley, S.K. et al. Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, 627-641 (2017).
38. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51**, D638-D646 (2023).
39. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *biorxiv*, 2021.2010. 2004.463034 (2021).
40. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1-3 (2024).
41. Zheng, W. et al. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nature Methods* **21**, 279-289 (2024).

42. Si, Y. & Yan, C. Improved inter-protein contact prediction using dimensional hybrid residual networks and protein language models. *Briefings in Bioinformatics* **24**, bbad039 (2023).
43. Brekke, O.H. & Sandlie, I. Therapeutic antibodies for human diseases at the dawn of the twenty-first century. *Nature reviews Drug discovery* **2**, 52-62 (2003).
44. Szeto, C., Lobos, C.A., Nguyen, A.T. & Gras, S. TCR recognition of peptide–MHC-I: Rule makers and breakers. *International journal of molecular sciences* **22**, 68 (2020).
45. Swapna, L.S., Bhaskara, R.M., Sharma, J. & Srinivasan, N. Roles of residues in the interface of transient protein-protein complexes before complexation. *Scientific reports* **2**, 334 (2012).
46. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026-1028 (2017).
47. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* **45**, D170-D176 (2017).
48. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics* **20**, 1-15 (2019).
49. Gribskov, M., McLachlan, A.D. & Eisenberg, D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A* **84**, 4355-4358 (1987).
50. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130 (2014).
51. 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** (2021).
52. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *nature* **596**, 583-589 (2021).
53. Studer, G., Tauriello, G. & Schwede, T. Assessment of the assessment—All about complexes. *Proteins: Structure, Function, and Bioinformatics* **91**, 1850-1860 (2023).
54. Lin, T. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:1708.02002* (2017).

Figures

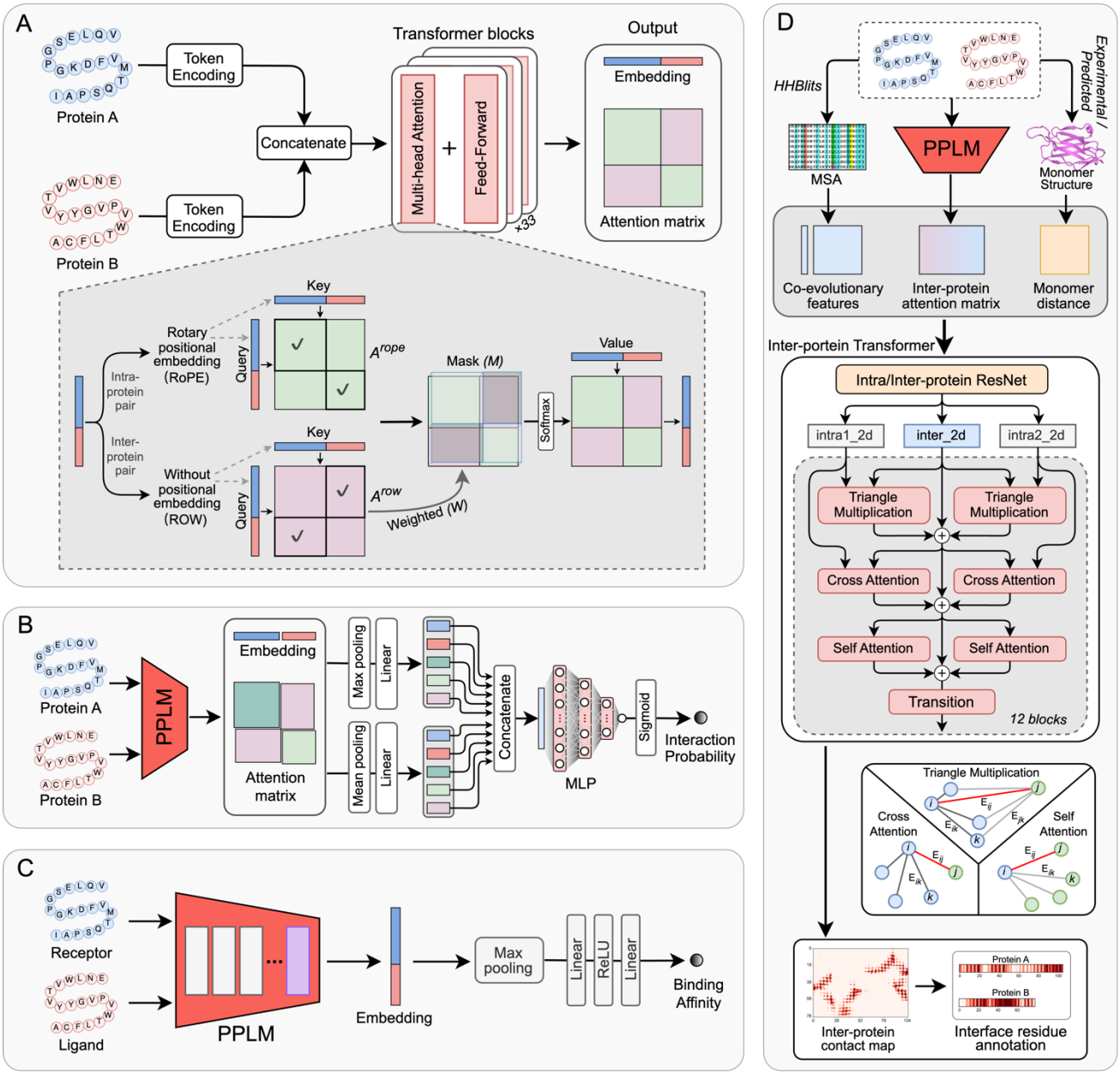

Figure 1. Overview of the PPLM framework and its downstream applications. (A) PPLM architecture. The input protein sequence pair are independently tokenized and then processed through transformer blocks equipped with a cross-protein attention mechanism to capture intra- and inter-protein interactions. The resulting embeddings and attention matrixes can be leveraged for diverse PPI tasks. **(B)** PPLM-PPI pipeline. For the input sequences, PPLM is used to generate the embeddings, intra-protein attention matrices and inter-protein attention matrix, which are aggregated using max and mean pooling. The aggregated features are concatenated and passed through a multilayer perceptron to predict the interaction probability between the input sequences. **(C)** PPLM-Affinity pipeline. The final transformer block of PPLM is finetuned on binding affinity data, where the resulting embeddings for paired sequences are aggregated via max pooling and passed through two fully connected layers to predict binding affinity value. **(D)** PPLM-Contact pipeline. For a given sequence pairs, inter-protein attention matrix from PPLM is integrated with MSA-derived features and monomer distance maps to capture both evolutionary and structural information. These features are processed by a series of inter-protein transformer blocks, with each composed of three core modules to predict the inter-protein contacts. Each core module adopts a parallel architecture to update inter-protein representations using both intra- and inter-protein features.

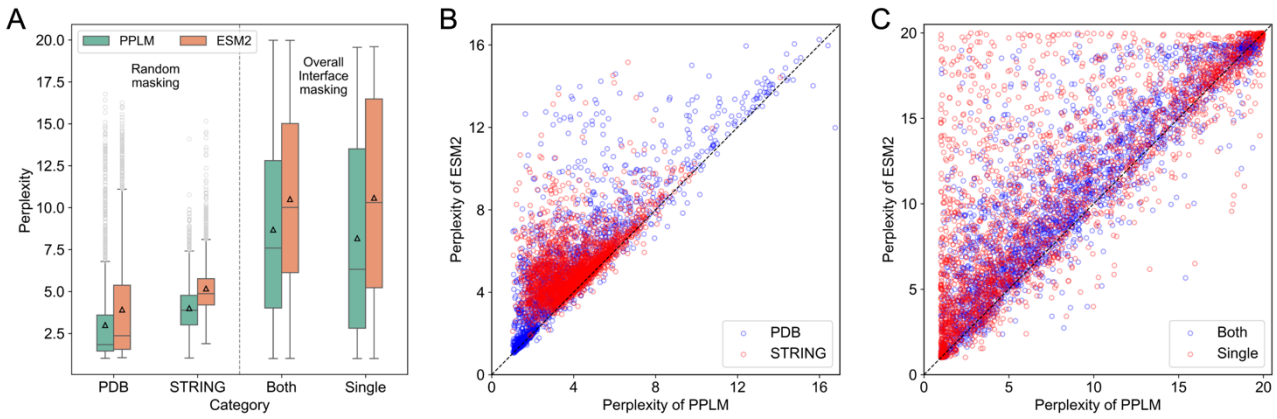

Figure 2. Perplexity of PPLM and ESM2 on protein sequence pairs. **(A)** Perplexity distribution for sequence pairs from the PDB and STRING databases under random masking and interface masking strategies. Interface masking is applied only to sequence pairs from PDB database, with single-chain interface masking used exclusively for heterodimers. **(B)** Head-to-head comparison of perplexity between PPLM and ESM2 under random masking. **(C)** Head-to-head comparison under overall interface masking, including both dual-chain and single-chain masking.

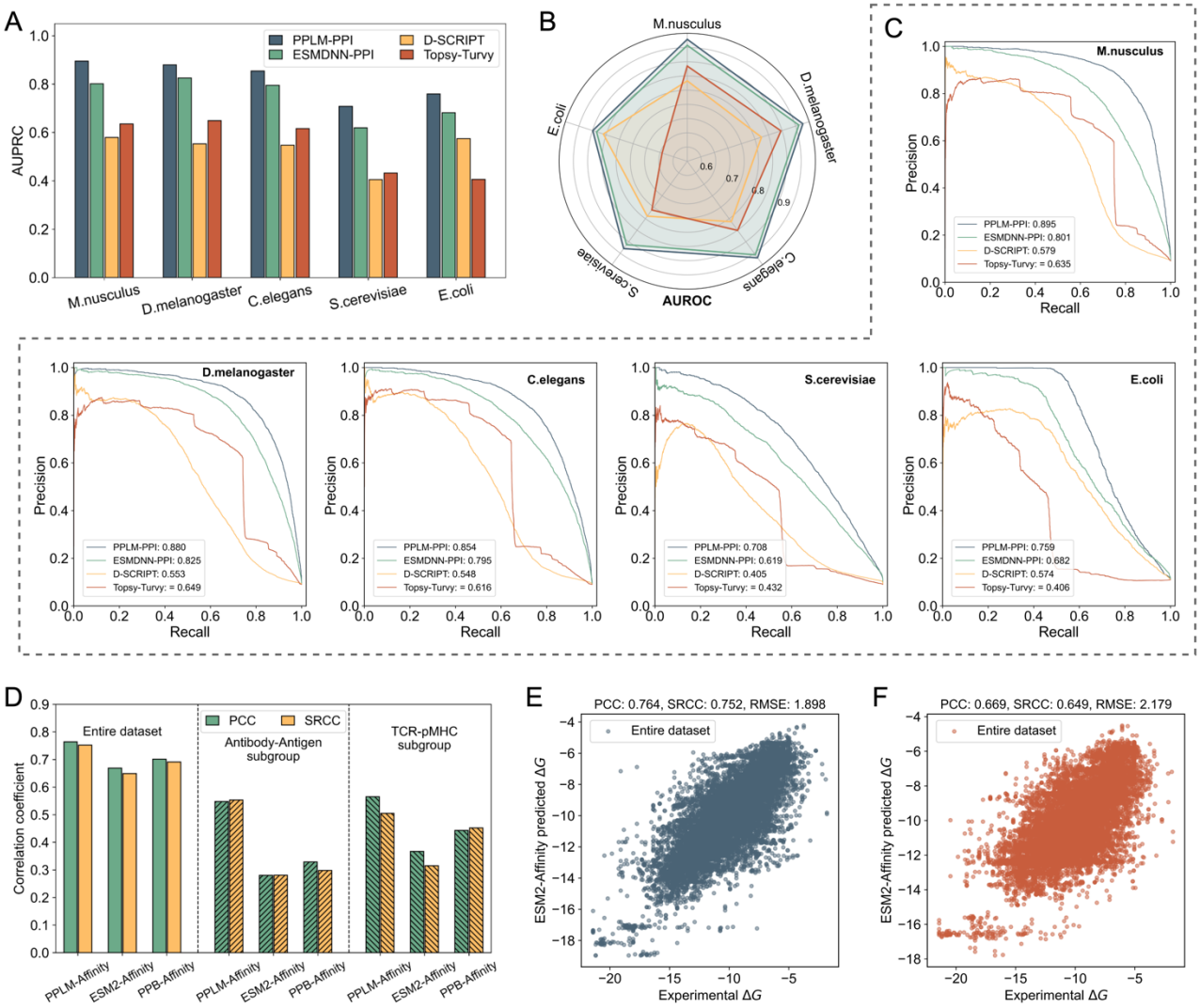

Figure 3. Performance of PPLM-PPI and PPLM-Affinity on protein–protein interaction and binding affinity prediction. (A) AUPRC of PPLM-PPI, ESMDNN-PPI, D-SCRIPT, and Topsy-Turvy across five species test sets. **(B)** AUROC of the four PPI prediction methods on the five species test sets. **(C)** Precision-recall curves of the four PPI prediction methods on the five species test sets. **(D)** Pearson (PCC) and Spearman (SRCC) correlation coefficients of PPLM-Affinity, ESM2-Affinity, and PPB-Affinity evaluated on the entire dataset, as well as the antibody–antigen and TCR–pMHC subgroups. **(E)** Comparison between experimental and PPLM-Affinity predicted binding affinities on the entire dataset. **(F)** Comparison between experimental and ESM2-Affinity predicted binding affinities on the entire dataset.

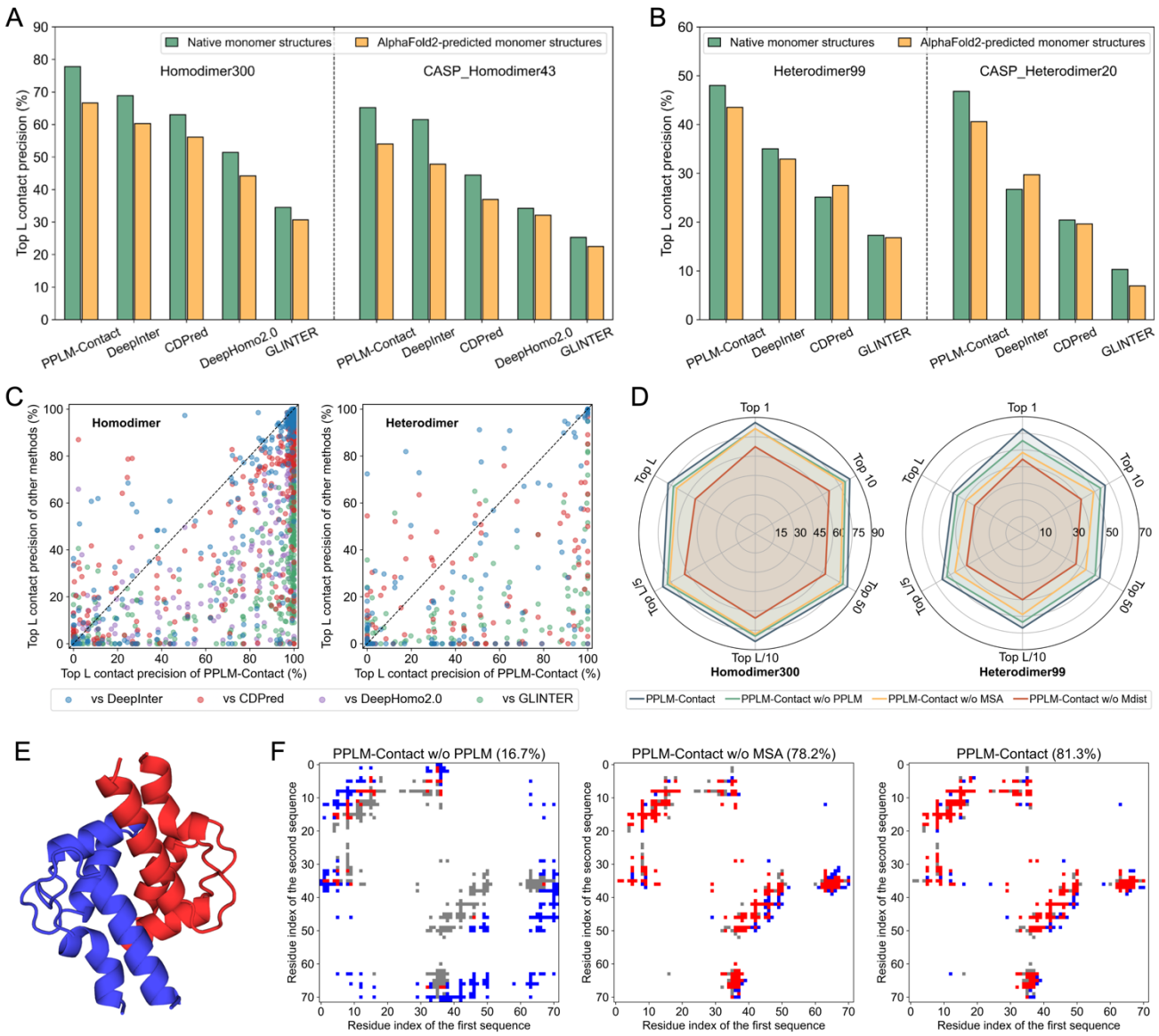

Figure 4. Performance of PPLM-Contact and the comparison methods on the inter-protein contact prediction. (A) Top *L* contact precision of PPLM-Contact, DeepInter, CDPred, DeepHomo2.0, and GLINTER on the Homodimer300 and CASP_Homodimer43 test sets using both experimental monomer structures and AlphaFold2-predicted monomer structures. **(B)** Top *L* contact precision of PPLM-Cxontact, DeepInter, CDPred, and GLINTER on the Heterodimer99 and CASP_Heterodimer20 test sets using both experimental monomer structures and AlphaFold2-predicted monomer structures. **(C)** Head-to-head comparison of top *L* contact precision between PPLM-Contact and other methods on the 343 homodimer and 119 heterodimer proteins. **(D)** Top contact precisions of PPLM-Contact and its ablation variants on the Homodimer300 and Heterodimer99 test sets. **(E)** An example of inter-protein contact prediction for the EF-hand domain of human RASEF (PDB ID: 2PMY). **(F)** Contact maps for 2PMY predicted by PPLM-Contact w/o PPLM, PPLM-Contact w/o MSA, and the full PPLM-Contact models, respectively. Each contact map shows the alignment between ground-truth contacts and top *N* predicted contacts, where *N* equals the number of ground-truth contacts. Grey dots representing ground-truth contacts, red dots indicate true positive predictions, and blue dots denote false positive predictions. The number in the parentheses denotes the top *N* contact precision.

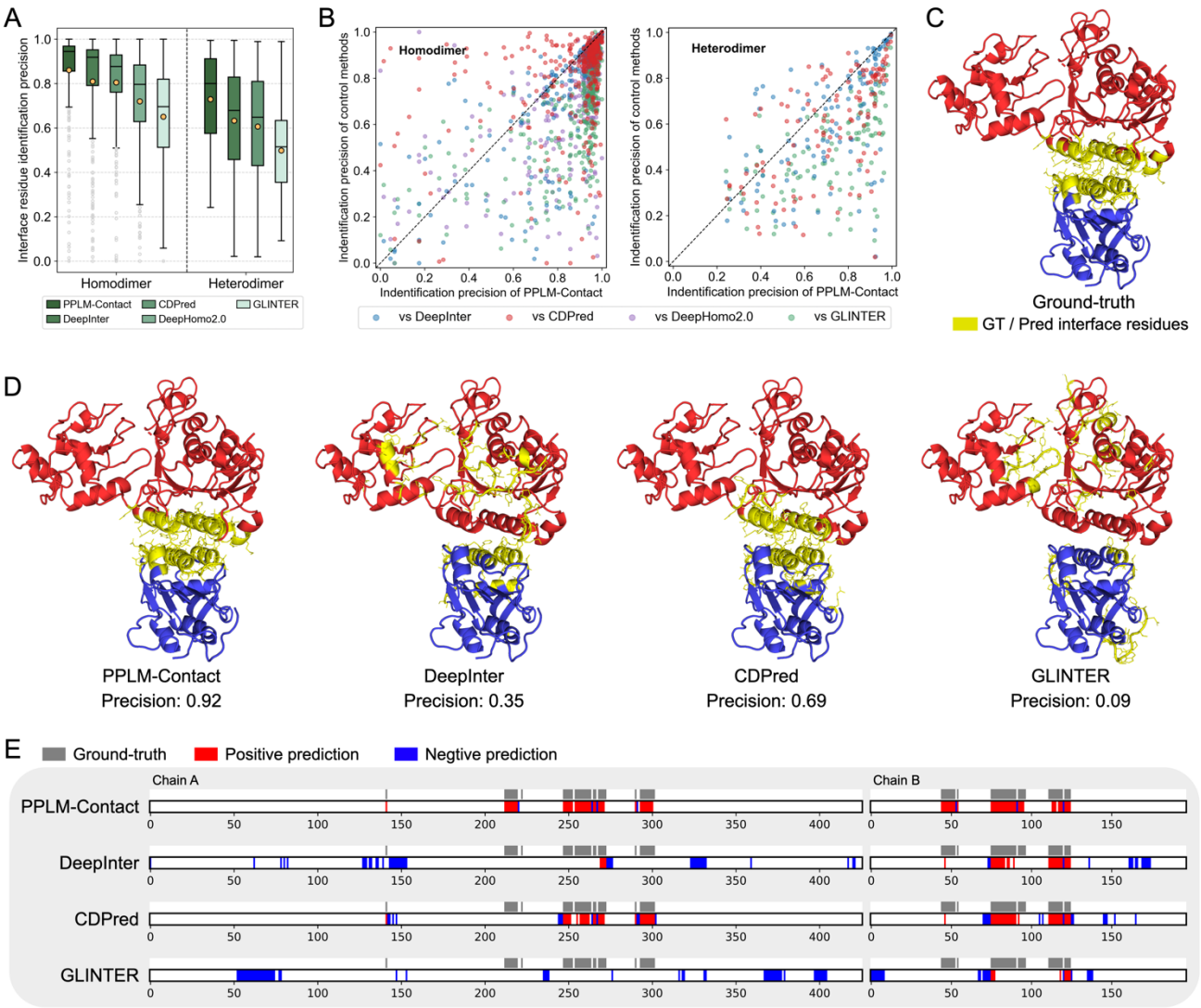

Figure 5. Comparative analysis of protein–protein interface residue identification. (A) Precision of interface residue identification by PPLM-Contact, DeepInter, CDPred, DeepHomo2.0, and GLINTER on the homodimer and heterodimer proteins in the test sets using experimental monomer structures. **(B)** Head-to-head comparison of identification precision between PPLM-Contact and each control method on the homodimer and heterodimer proteins. **(C)** 3D structure visualization of the Crystal structure of DNMT3A-DNMT3L complex (4U7P), with ground-truth interface residues are highlighted in yellow and shown in stick representation. **(D)** Visualization of predicted interface residues identified by each method, with predicted interface residues are highlighted in yellow and shown in stick representation **(E)** Residue-level annotation maps of each method for 7L15, with grey bars indicate ground-truth interface residues, red bars represent true positive predictions, and blue bars denote false positive predictions.

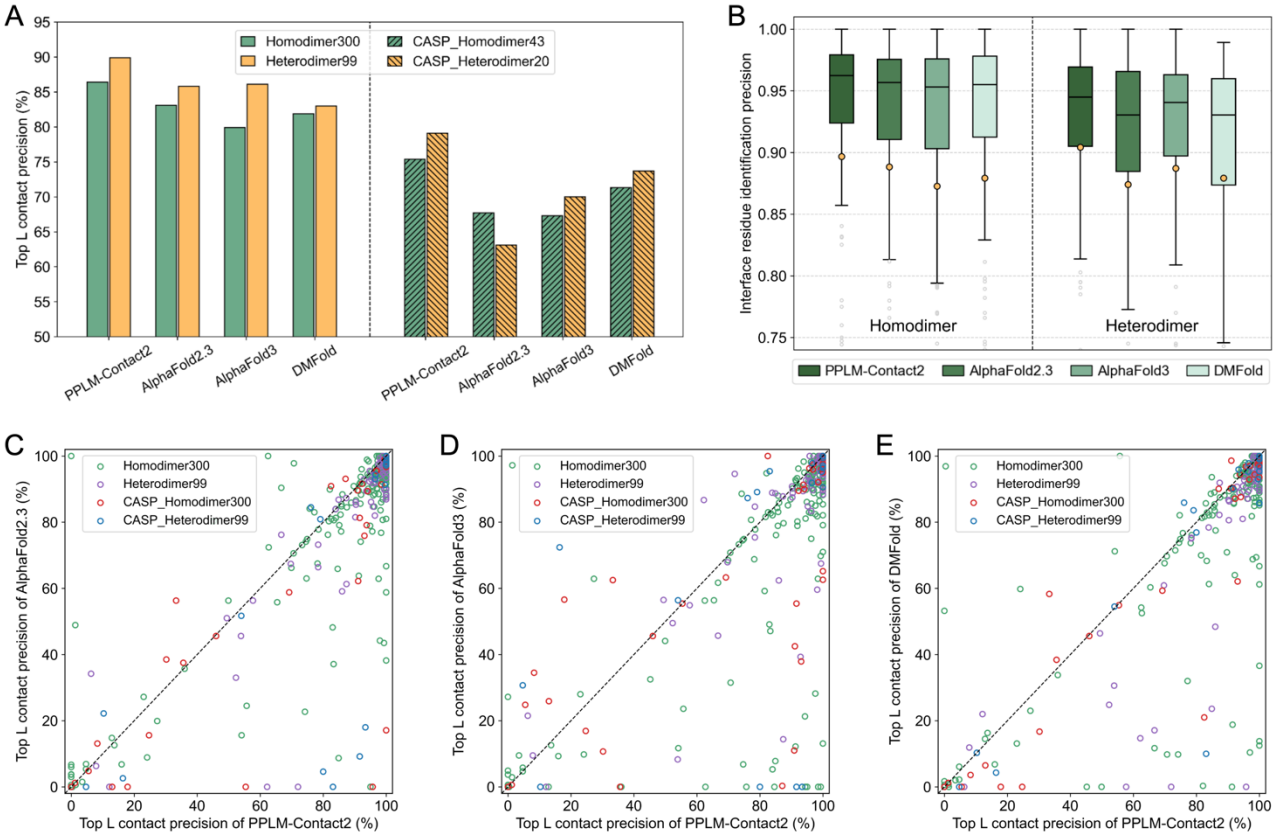

Figure 6. Performance of the enhanced PPLM-Contact2 and the state-of-the-art protein complex structure prediction methods on inter-protein contact prediction. (A) Top *L* contact precision of PPLM-Contact2, AlphaFold2.3, AlphaFold3, and DMFold on the Homodimer300, Heterodimer99, CASP_Homodimer43, and CASP_Heterodimer20 test sets. **(B)** Precision for interface residue identification by PPLM-Contact2 and control methods on all homodimer and heterodimer protiens. **(C–E)** Head-to-head comparison of top *L* contact precision between PPLM-Contact2 and AlphaFold2.3, AlphaFold3, and DMFold, respectively.