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. Here, we introduce a novel Protein–Protein Language Model (PPLM) that jointly encodes paired sequences, enabling direct learning of interaction-aware representations beyond what single-chain models can provide. Building on this foundation, we develop PPLM-PPI, PPLM-Affinity, and PPLM-Contact for binary interaction, binding affinity, and interface contact prediction. 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 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[14-21], 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[22], and ESMDNN-PPI[23] 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[24, 25], with existing methods for protein–protein binding affinity prediction still depending heavily on complex structural models[26-30], 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)[31-34]. While methods such as GLINTER[35], CDPred[19], and DeepInter[20] employ ESM-MSA-1b[36] 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 present PPLM, a protein–protein language model specifically designed to learn contextual and relational features from paired protein sequences. To better capture inter-protein dependencies, we introduce a hybrid intra–/inter-protein attention mechanism that adapts transformer attention for co-represented protein pairs. Specifically, this mechanism integrates rotary embeddings-based positional encoding[37] for intra-protein attention with non-positional embedding for inter-protein residue pairs to avoid spurious spatial priors. In addition, learnable inter-protein attention weights and an explicit inter-protein attention mask are incorporated to enhance the modeling of cross-protein interactions while preserving a clear separation between intra- and inter-chain information flows. To support large-scale training, we further construct a comprehensive sequence-pair dataset comprising over 3.3 million high-quality protein pairs, enabling PPLM to learn robust interaction-aware representations.

Building upon this foundation, 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 all tasks, our model consistently outperforms existing methods. In inter-protein contact-map prediction, PPLM-Contact surpasses current state-of-the-art structure modelling-based methods, including AlphaFold2.3[38], AlphaFold3[39], and DMFold[40]. These results underscore PPLM's potential as a unified foundation for advancing computational PPI studies.

Results

PPLM comprises four coordinated pipelines that support both core language modeling and its downstream applications. In the primary language modeling pipeline (**Figure 1A**), each input sequence is independently tokenized with begin and end tokens and then concatenated to form the full paired input. The transformer blocks with a hybrid intra–/inter-protein attention mechanism is applied to jointly model intra- and inter-protein interactions. The resulting sequence embeddings and attention

matrices provide versatile representations that can be flexibly leveraged across PPI-related tasks.

To demonstrate the versatility of PPLM, we developed three task-specific models that address major challenges in PPI studies. For PPI prediction (**Figure 1B**), PPLM produces sequence embeddings together with intra- and inter-protein attention matrices for the sequence pair, which are aggregated through parallel max- and mean-pooling branches, each followed by a multilayer perceptron (MLP) for ensemble prediction. For binding affinity prediction (**Figure 1C**), the final layer of PPLM is fine-tuned on a curated 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. 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 used to predict contact maps and identify interface residues between the input proteins.

PPLM improves language modeling of protein pairs

Perplexity provides an objective, task-independent measure of a language model's confidence in predicting masked amino acids. Lower perplexity indicates higher predictive accuracy, with a value of 1 representing perfect prediction and ~20 reflecting near-random performance. We used perplexity to benchmark PPLM against ESM2[6], a state-of-the-art single-sequence language model that does not explicitly model inter-protein context. To more comprehensively evaluate model behavior on protein pairs, we implemented three complementary masking strategies (**Supplementary Fig. S1**).

Under a random masking strategy, perplexity was computed by iteratively masking non-overlapping 15% of residues in both sequences until all positions were covered. Identical positions in homomers were masked jointly to avoid trivial inference. As shown in the left panel of **Figure 2A**, PPLM achieved average perplexities of 7.29, 5.08, and 4.50 on homomeric PDB pairs, heteromeric PDB pairs, and STRING-derived pairs, corresponding to reductions of 13.1%, 24.5%, and 21.0% relative to ESM2 (8.40, 6.73, and 5.70). Median perplexity showed consistent improvements of 15.1%, 28.5%, and 19.0%. These differences were highly statistically significant (Wilcoxon signed-rank test p-values 2.96×10^{-139}, 8.07×10^{-177}, and $<10^{-300}$). A head-to-head comparison (**Figure 2B**) further showed that PPLM outperformed ESM2 on 91.9% (4,298/4,678) of sequence pairs.

To evaluate PPLM's sensitivity to interface residues, we next masked only residues located at the experimentally defined interaction interface. For heteromeric pairs, we introduced Dual mode (masking all interface residues from both chains) and Single mode (masking all interface residues from one chain at a time); homomeric used only the Dual mode to avoid trivial inference. As shown in the right panel of **Figure 2A**, In the Dual mode, PPLM achieved an average perplexity of 6.79 (median 5.78), significantly lower than ESM2's 8.50 (7.75) (p-value = 9.90×10^{-223}). In the Single mode, PPLM again outperformed ESM2 (7.81 vs 10.36 in average perplexity; 5.62 vs 8.89 in median; p-value = 1.17×10^{-208}). Corresponding pairwise comparisons (**Figure 2C**) showed that PPLM achieved lower perplexity in 80.7% (Dual) and 78.9% (Single) of cases.

To examine whether PPLM captures biologically meaningful dependencies, we analyzed the homodimer complex 1Y9B, a conserved transcription factor from Vibrio cholerae O1 biovar El Tor (**Figure 2D–F**). The inter-protein attention matrix from the final PPLM layer showed high-scoring regions that closely matched the experimental contact map. Among the top 20 residue pairs ranked by attention, 90% corresponded to true heavy-atom contacts (80% to C_β–C_β contacts). Of the 62 experimentally determined interface residues, 52 (83.9%) were recoverable from attention-derived residue rankings. The corresponding 3D visualization demonstrated that these predicted residues cluster around the physical interface, indicating that PPLM's attention mechanism naturally focuses on interaction-relevant regions during unsupervised learning. A more detailed representation-level comparison between PPLM and ESM2, including unsupervised and linear-probe analyses of inter-

protein attention matrices, is provided in **Supplementary Text S1**.

As expected for paired-sequence modeling, smaller PPLM variants exhibited higher perplexity, whereas deeper models showed substantially improved performance (**Figure 2G**). To assess potential pair-level redundancy and species-specific biases, we conducted a comprehensive bias audit by comparing all validation pairs to the training set. Most validation pairs had no detectable match, and PPLM's performance remained stable across identity intervals (**Figure 2H**). PPLM consistently outperformed ESM2 across the entire identity spectrum, and this trend remained robust across alternative coverage thresholds (**Supplementary Figure S3**).

Collectively, these results demonstrate that PPLM enhances language modeling of protein pairs, exhibits increased sensitivity to residue-level interactions at protein–protein interfaces, and effectively captures biologically meaningful inter-chain dependencies. These capabilities, enabled by paired-sequence training and a hybrid attention architecture, provide strong foundations for downstream tasks such as interface residue identification and structure-guided interface design, including antibody CDR optimization.

PPLM-PPI outperforms existing methods in binary protein interaction prediction

To evaluate the effectiveness of PPLM-PPI on protein–protein interaction predictions, we benchmarked it across five species, including *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, and *E. coli*. PPLM-PPI was compared against four state-of-the-art LM-based PPI predictors, including TUnA[41], ESMDNN-PPI[23], D-SCRIPT[15], and Topsy-Turvy[22]. All methods were installed and executed locally for reproducibility, except ESMDNN-PPI, which we reimplemented following the original publication due to the lack of publicly available source code.

Comprehensive evaluation results, including precision, recall, accuracy, F1-score, area under the receiver operating characteristic curve (AUROC), and area under the precision–recall curve (AUPRC) for all species, are provided in **Supplementary Table S1**. PPLM-PPI achieved the highest performance on nearly all metrics across all datasets. Since AUPRC is especially informative for imbalanced PPI datasets, we use it as the primary metric. As summarized in **Figure 3A**, PPLM-PPI achieved AUPRC scores of 0.920, 0.906, 0.882, 0.5745, and 0.784 on the five species, corresponding to improvements of 4.0%, 6.2%, 7.2%, 17.6%, and 12.9% over the second-best method, TUnA. Similarly, PPLM-PPI achieved the highest F1-score for all species, surpassing TUnA by 4.8%, 8.6%, 4.8%, 16.9%, and 15.5% (**Figure 3B**). Precision–recall curves across species further show that PPLM-PPI consistently dominates other methods across most recall thresholds (**Figure 3C**). In addition, the curves of PPLM-PPI are markedly smoother than those of competing methods, indicating superior stability and predictive consistency, particularly in high-precision and high-recall regimes.

An overall analysis across the five species test sets showed that PPLM-PPI demonstrated superior stability, with a mean ± standard deviation of 0.848 ± 0.078, compared to 0.778 ± 0.109 for TUnA, 0.762 ± 0.106 for ESMDNN-PPI, 0.532 ± 0.072 for D-SCRIPT, and 0.548 ± 0.118 for Topsy-Turvy (**Supplementary Table S2**). **Supplementary Table S3** details the statistical comparisons. Relative to TUnA and ESMDNN-PPI, PPLM-PPI improves F1-score by 10.1–10.5% and AUPRC by 9.6–11.9%, with large effect sizes (Cohen's d ≈ 1.742–6.586). Against D-SCRIPT and Topsy-Turvy, the improvements are even larger: 32.5–72.3% (F1) and 58.7–60.9% (AUPRC), also with consistently large effect sizes (Cohen's d ≈ 2.062–6.183).

To elucidate the contributions of different components, we conducted ablations by modifying pooling modules and selectively removing PPLM-derived features (**Supplementary Text S2)**. Multi-seed results (mean ± 95% CI) are presented in **Supplementary Table S4-5 and Figure S4**. Among the individual pooling strategies, mean pooling achieved the highest accuracy, while the combined mean–max pooling strategy produced the most robust and consistent performance across species. In the feature-level analysis, incorporating all three PPLM-derived feature types, including inter-protein attention, intra-protein attention, and embedding, yielded the best results, with embedding features contributing the strongest predictive signal. These findings highlight the importance of both effective

pooling and the integration of complementary attention-based and embedding-based representations.

Finally, to assess inductive generalization and evaluate whether shortcut mechanisms might arise in our paired-input setting[42, 43], we performed a sequence-similarity–stratified evaluation. Test interactions were grouped by the maximum single-sequence identity of either protein to any human protein in the training set, and all methods were compared within each identity interval. As detailed in **Supplementary Text S3**, PPLM-PPI maintains strong performance across all homology ranges and delivers the largest gains over existing methods, demonstrating that its predictive accuracy is not driven by shortcut learning and that the model generalizes robustly to unseen proteins under a strict cross-species inductive setting.

PPLM-Affinity achieves superior performance in protein binding affinity prediction

Although protein–protein interaction and binding affinity are related, they address distinct questions: PPLM-PPI estimates the probability that two proteins interact, whereas affinity prediction quantifies the continuous strength of binding. To explicitly target quantitative binding strength, we developed PPLM-Affinity by fine-tuning the last transformer block of PPLM to predict binding affinity values. We evaluated PPLM-Affinity on the PPB-Affinity dataset[28], a large and comprehensive benchmark for protein–protein binding affinity prediction. To avoid data leakage, we regrouped the five-fold splits according to structural similarity and performed five-fold cross-validation on the revised partitions. To assess the effectiveness of language models for this task, we implemented ESM2-Affinity by concatenating the receptor and ligand sequences and fine-tuning the model using the same architecture and training procedure as PPLM-Affinity. In addition, we retrained the structure-based PPB-Affinity model using its released source code under the same cross-validation setting. Detailed fold-wise results, mean ± standard deviation values and statistical comparison are summarized in **Supplementary Tables S7–S9**.

The left panel of **Figure 3D** reports the Pearson (PCC) and Spearman (SRCC) correlations between predicted and experimental ΔG values. These two metrics capture complementary aspects of predictive performance: PCC quantifies linear association, whereas SRCC measures rank consistency. PPLM-Affinity achieves mean ± standard deviation PCC and SRCC values of 0.643 ± 0.058 and 0.636 ± 0.082, representing improvements of 17.3% and 16.4% over ESM2-Affinity (0.548 ± 0.061 and 0.547 ± 0.053; Cohen's d = 1.274 and 0.958) and 18.0% and 17.8% over the structure-based PPB-Affinity model (0.545 ± 0.072 and 0.540 ± 0.088; Cohen's d = 1.326 and 1.064). **Supplementary Figure S6** depicts the distribution of absolute errors relative to experimental ΔG values. All three models show smooth, unimodal distributions without marked skewness. PPLM-Affinity exhibits the most concentrated error profile, with the lowest mean absolute error (1.68) and the smallest fitted dispersion (σ = 1.44), compared with higher values for ESM2-Affinity (1.85 and 1.52) and PPB-Affinity (1.85 and 1.63). These trends are consistent with the overall RMSE comparisons, where PPLM-Affinity again attains the lowest value (2.312 ± 0.297) relative to ESM2-Affinity (2.476 ± 0.376) and PPB-Affinity (2.463 ± 0.394). **Figure 3E-F** further illustrate the predicted versus experimental ΔG values for PPLM-Affinity and ESM2-Affinity, respectively. While both models show positive correlations with experimental measurements, the predictions from PPLM-Affinity cluster more tightly around the diagonal line, indicating improved accuracy, stronger correlations, and reduced systematic bias.

The middle and right panels of **Figure 3D** further analyze two biologically important subsets—antibody–antigen and TCR–pMHC interactions—which play central roles in immune recognition and therapeutic development[44, 45]. On the antibody–antigen subset, PPLM-Affinity achieves PCC and SRCC values of 0.380 and 0.404, corresponding to improvements of 117.1% and 111.5% over ESM2-Affinity, and 60.3% and 57.2% over PPB-Affinity. On the TCR–pMHC subset, PPLM-Affinity attains PCC and SRCC values of 0.366 and 0.312, yielding gains of 144.0% and 127.7% relative to ESM2-Affinity, and 39.2% and 36.8% relative to PPB-Affinity.

Because PPLM-Affinity and ESM2-Affinity share identical architectures and fine-tuning procedures, the substantial improvements delivered by PPLM-Affinity highlight the advantage of

using representations explicitly tailored to protein–protein interactions. Its strong performance relative to the structure-based PPB-Affinity model further indicates that PPLM can capture binding energetics directly from sequence, offering a scalable and structure-independent alternative for affinity prediction. Together, these results suggest that PPLM-Affinity provides improved accuracy and robustness across both general protein complexes and immunologically challenging systems, supporting its potential utility in antibody engineering, TCR optimization, and other sequence-based therapeutic design applications.

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

We evaluated PPLM-Contact on four benchmark datasets: Homodimer300 and Heterodimer99 from DeepInter20, and CASP_Homodimer43 and CASP_Heterodimer20. Comparisons were made against five state-of-the-art methods: PLMGraph-Inter[34], DeepInter[20], CDPred[19], GLINTER[35], and DeepHomo2.0[32]. All baseline methods were installed and executed locally to ensure reproducibility. Inter-protein contact precision was assessed at multiple cutoffs (top 1, 10, 50, $L/10$, $L/5$, and L), where L is the length of the shorter protein chain in the complex. Summary statistics and paired comparisons are provided in **Supplementary Table S10–13**, with per-protein results in **Supplementary Data 1–2**.

Figure 4A reports the top L contact precision on homodimer test sets using either experimental or AlphaFold2-predicted monomer structures. On Homodimer300 with experimental structures, PPLM-Contact achieves 77.8% precision, substantially outperforming DeepInter (68.9%), PLMGraph-Inter (50.6%), CDPred (63.0%), DeepHomo2.0 (51.4%), and GLINTER (34.5%) by margins of 12.8%, 53.7%, 23.5%, 51.3%, and 125.3%. All comparisons are statistically significant (Wilcoxon signed-rank test $p = 3.63\times10^{-21}$ to 1.29×10^{-47}). Using AlphaFold2-predicted monomer structures, PPLM-Contact maintains strong performance (66.6%) and surpasses the second-best method, DeepInter, by 10.4% ($p = 3.96\times10^{-13}$). A similar trend is observed on CASP_Homodimer43. With experimental monomer structures, PPLM-Contact achieves 65.2% precision and improves over baseline methods by 6.0%–157.7%. When using AlphaFold2-predicted monomer structures, the corresponding improvements are 12.9%, 39.5%, 46.3%, 72.3%, and 140.0%, respectively.

Figure 4B presents results on heterodimer test sets. On Heterodimer99, PPLM-Contact achieves 48.9% precision using experimental monomers and 45.1% using AlphaFold2-predicted monomers. These correspond to improvements of 37.0%–182.7% over competing methods, all statistically significant. On CASP_Heterodimer20, PPLM-Contact attains 45.6% precision with experimental monomer structures and 40.5% with AlphaFold2 monomers, again outperforming all baseline approaches. Comprehensive statistical analyses, including paired Cohen's d effect sizes and Benjamini–Hochberg false-discovery-rate (FDR) correction across the four test sets, are provided in **Supplementary Text S4**.

Figure 4C provides head-to-head comparisons between PPLM-Contact and each baseline method across all 434 homodimer and 119 heterodimer proteins using experimental monomer structures. Relative to DeepInter, PPLM-Contact achieves higher / lower top L contact precision on 214 / 48 homodimer proteins (62.4% / 14.0%) and 79 / 22 heterodimer proteins (66.4% / 18.5%). The corresponding higher / lower proportions are 82.5% / 12.0% and 63.0% / 24.4% for PLMGraph-Inter, 80.8% / 12.2% and 65.5% / 20.2% for CDPred, 87.2% / 8.2% for DeepHomo2.0, and 91.0% / 6.7% and 78.2% / 13.4% for GLINTER. These results indicate that PPLM-Contact achieves higher performance on a clear majority of targets, while also making it explicit that a non-trivial fraction of predictions is tied rather than strictly lower. **Supplementary Figure S7** presents the corresponding comparisons using AlphaFold2-predicted monomer structures, showing a similar trend in which PPLM-Contact outperforms baseline methods for most targets.

Figure 4D–F summarize the mean and standard deviation of top L precision across three random seeds for PPLM-Contact and all ablation variants. These results show that removing any major network component or feature of the model reduces precision, with the largest performance drops observed when excluding the triangle-multiplication module, PPLM inter-protein attention, MSA features, or

monomer distance maps. Detailed ablation analyses and statistical comparisons are provided in **Supplementary Text S5**. The effect of MSA depth is further examined in **Supplementary Text S6**.

Figures 5A–B illustrate an example on the EF-hand domain of human RASEF (PDB ID: 2PMY), consisting of two 91-residue identical chains (72 resolved structurally). When relying solely on MSA-derived features and monomer structures, PPLM-Contact without (w/o) PPLM features yields heavily incorrect top *N* predictions, achieving only 16.7% precision (20.8% for top *L*). In contrast, replacing MSA-derived features with PPLM-generated features allows PPLM-Contact w/o MSA to correctly identify 201 of 257 ground-truth contacts (78.2%) with 100% top *L* precision. Combining both PPLM and MSA features further improves performance to 81.3% top *N* precision while maintaining 100% top *L* precision. This example illustrates the strong ability of PPLM-derived features to capture geometric and interface-level spatial patterns critical for accurate inter-protein contact prediction.

Enhanced inter-protein contact prediction using predicted complex structures

Building on the performance of PPLM-Contact, we next examined whether incorporating structural information from modern complex-prediction models could further enhance inter-protein contact prediction. Recent advances in multimer structure prediction now provide highly accurate models of protein complexes, capturing detailed geometric organization at interfaces that is difficult to infer from sequence alone. These predicted structures therefore offer an opportunity to supply language-model pipelines with precise spatial context. To leverage this, we developed PPLM-Contact2, an enhanced variant of PPLM-Contact that integrates inter-protein distance maps extracted from predicted complex structures. We benchmarked PPLM-Contact2 against AlphaFold2.3[38], AlphaFold3[39], and DMFold[40] across all 434 homodimer and 119 heterodimer proteins (**Supplementary Tables S14–S15; Supplementary Data 3–4**).

Figure 5C–D summarize the top *L* precision of all general contact-prediction methods (using AlphaFold2-predicted monomer structures) and the complex structure-based methods (AlphaFold2.3, AlphaFold3, DMFold, and PPLM-Contact2) on the 434 homodimer and 119 heterodimer proteins. As expected, the integration of complex-structure features allows PPLM-Contact2 to markedly improve upon PPLM-Contact, increasing top *L* precision from 65.0% to 85.1% on homodimers and from 44.3% to 88.0% on heterodimers. Overall, all complex structure-based methods outperform general contact-prediction methods, consistent with the fact that AlphaFold leverages deep 3D structure-prediction pipelines, extensive geometric reasoning, and large-scale template and MSA signals, thereby providing much richer structural context than purely sequence- or MSA-based contact predictors. Compared with structure-modeling approaches, PPLM-Contact2 achieves improvements of 4.8% and 7.6% over AlphaFold2.3, 8.7% and 5.6% over AlphaFold3, and 5.6% and 8.1% over DMFold on homodimers and heterodimers. All improvements are statistically significant (**Supplementary Table S15**).

Head-to-head comparisons across the full benchmark (**Figure 6E–G**) further highlight the advantage of PPLM-Contact2. Relative to AlphaFold2.3, PPLM-Contact2 achieves higher / lower top *L* precision on 207 / 44 homodimer proteins (60.3% / 12.8%) and 75 / 10 heterodimer proteins (63.0% / 8.4%). The corresponding higher / lower proportions are 62.1% / 11.7% and 56.3% / 18.5% for AlphaFold3, and 62.7% / 11.1% and 64.7% / 10.1% for DMFold. **Supplementary Figure S9** presents the distributions of precision, recall, and F1-score across all test proteins. Compared with the structure-based modeling methods, PPLM-Contact2 shows a clear shift toward higher ranges for all three metrics. It achieves an F1-score of 0.780, exceeding 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 inter-protein contact prediction.

Together, these results demonstrate that even with the remarkable advances in end-to-end complex structure prediction, additional gains can be achieved by integrating predicted structural models with interaction-aware language-model features. The resulting hybrid framework not only improves contact-map prediction but also provides more reliable localization of inter-chain binding sites. This integrated information may serve as spatial constraints for downstream quaternary structure assembly

and offers a principled strategy for further improving protein complex modeling.

PPLM-Contact enhances interface residue recognition

Beyond predicting inter-protein contacts, accurately identifying interface residues is essential for understanding complex formation, signaling, and functional regulation of protein assemblies[46]. We therefore evaluated the interface-recognition performance of all general contact-prediction methods (using AlphaFold2-predicted monomer structures; corresponding results using experimental monomer structures are provided in **Supplementary Table S17** and **Figure S10**) as well as complex structure-based approaches on the 434 homodimer and 119 heterodimer proteins. Interface residues were defined as those with any heavy atom within 8 Å of a heavy atom from the partner chain in the experimental complex structure. For each target, we assessed the top *N* predicted residues, where *N* equals the number of ground-truth interface residues, independent of their specific contact partners. **Figure 6A–B** report the interface identification precision for all methods (**Supplementary Data 5–6**).

Among general contact-prediction methods, PPLM-Contact achieves the highest average precision, reaching 0.824 for homodimers and 0.695 for heterodimers. These values represent improvements of 5.0% and 12.1% over DeepInter, 7.6% and 7.1% over PLMGraph-Inter, 5.3% and 14.2% over CDPred, 19.3% over DeepHomo2.0 (homomers), and 28.7% and 39.7% over GLINTER. PPLM-Contact also attains the highest median precision (0.916 and 0.752), indicating consistently strong predictions across targets. All improvements are statistically significant (**Supplementary Table S16**). Integrating complex-structure features further boosts performance: PPLM-Contact2 markedly improves upon PPLM-Contact, increasing average and median precision to 0.897 and 0.962 for homodimers and 0.904 and 0.945 for heterodimers. Compared with structure-based methods, PPLM-Contact2 achieves the highest average and median precision across both homodimer and heterodimer test sets. Comprehensive statistical analyses confirm that these improvements are significant (**Supplementary Table S18**). The head-to-head comparisons between PPLM-Contact and general contact-prediction methods, as well as between PPLM-Contact2 and complex structure modelling approaches, further show that both PPLM-Contact and PPLM-Contact2 outperform competing methods on a clear majority of targets, underscoring their robustness and broad applicability (**Supplementary Figure S11-12**).

Figures 6C–D present an illustrative example using the crystal structure of the DNMT3A–DNMT3L complex (PDB ID: 4U7P), consisting of 426 and 197 experimentally resolved residues in two distinct chains. Interface residues identified by each method are highlighted on the 3D structure, and **Figure 6E** compares ground-truth and predicted residues along sequence positions. PPLM-Contact accurately identifies 40 of 43 (93.0%) interface residues on chain A and 40 of 44 (90.9%) on chain B, substantially outperforming baseline approaches. In contrast, DeepInter identifies only 4 (9.3%) and 26 (59.1%) residues; PLMGraph-Inter identified 34 (79.1%) and 31 (70.5%), CDPred identified 29 (67.4%) and 31 (70.5%), and GLINTER misidentifies all interface residues on chain A and recovers only 8 (18.2%) on chain B. Detailed inspection reveals that baseline methods tend to overpredict residues near the N-terminus of chain A and the C-terminus of chain B, whereas PPLM-Contact produces a more balanced residue distribution along both chains, yielding an overall precision of 0.92, markedly higher than DeepInter (0.35), PLMGraph-Inter (0.75), CDPred (0.69), and GLINTER (0.09).

These results demonstrate that PPLM-Contact achieves superior performance in interface residue recognition. 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.

Discussion

Computational modeling of protein–protein interactions (PPIs) is essential for elucidating cellular mechanisms and accelerating therapeutic discovery. Although protein language models have recently emerged as scalable tools for PPI analysis, they often fail to capture inter-protein context explicitly. To address this limitation, we introduced PPLM, a paired-sequence language model designed to learn inter-protein co-representation through a dedicated attention mechanism. Together with task-specific downstream architectures, PPLM supports interaction prediction, binding affinity estimation, and inter-protein contact prediction. Trained on a large composite dataset, PPLM achieves substantially lower perplexity than ESM2 at protein interfaces and consistently improves performance across the three downstream tasks. When combined with predicted structural features, the extended PPLM-Contact2 model also outperforms state-of-the-art structure-based approaches, including AlphaFold2.3, AlphaFold3, and DMFold.

Despite its strong performance, PPLM has several limitations that warrant further investigation. Its training relies on a composite sequence-pair dataset that, while diverse, may not fully capture the breadth of protein interactions across organisms and cellular states. In particular, transient, weak, or condition-dependent interactions remain underrepresented. A representative failure case is shown in **Supplementary Figure S13**, involving a nanobody–antigen complex with a very small, loop-mediated interface and weak co-evolutionary signal. Although PPLM-derived features partially compensate for the lack of MSA or structural information, they remain insufficient to recover these small, flexible, and weakly co-evolving interfaces, highlighting an intrinsic limitation of sequence-based contact prediction.

While this study focuses on three core PPI-related tasks, PPLM offers broader opportunities for integration with structural and functional modeling frameworks. Combining its interface-aware representations with AI-driven complex prediction tools such as AlphaFold3 may further improve the modeling of multimeric assemblies. Likewise, the strong performance of PPLM-Affinity on antibody–antigen and TCR–pMHC interactions suggests potential applications in CDR modeling and therapeutic design. By capturing fine-grained interface features, PPLM may provide a foundation for more accurate modeling of binding interactions and for guiding the development of next-generation biologics.

Methods

Datasets

To train PPLM, we constructed a composite protein dataset of protein interaction sequence pair by integrating protein complexes released before January 1, 2024 from the Protein Data Bank (PDB)[47] and interaction sequences from the STRING[48] database. After filtering for physical interfaces and removing redundancy through pair-level clustering using MMseqs2[49] together with a custom clustering procedure, the dataset comprised 25,245 heteromeric clusters and 23,082 homomeric clusters from PDB, as well as 629,045 clusters from STRING, covering over 3.3 million protein sequence pairs. For model validation, we randomly selected 5,000 singleton clusters and used the remaining clusters for training. After removing validation pairs that were redundant to any pairs in the training clusters, the final validation set consisted of 1,204 heteromeric, 1,003 homomeric, and 2,471 STRING-derived pairs (**Supplementary Text S7**). During the training of PPLM, sequence pairs were sampled from the PDB and STRING clusters at a 1:2 ratio.

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 (**Supplementary Text S8**). 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[28], , a large

curated resource compiled from multiple public affinity databases. The dataset provides experimentally determined binding affinities, crystal structures, mutation profiles, and annotations of receptor and ligand chains, comprising 12,062 interaction samples from 3,032 distinct PDB entries, including subgroups such as antibody–antigen and TCR–pMHC complexes. To prevent potential homology-based leakage, all PDB entries were clustered with US-align[50] using a complex-level TM-score cutoff of 0.8. Complexes with different numbers of chains were clustered separately to maintain meaningful structural comparability. All complexes within each structural cluster were assigned to the same fold in the five-fold cross-validation scheme. The resulting five folds contained 606, 606, 605, 605, and 605 PDB entries, respectively, and the list of PDB IDs for each fold is provided in **Supplementary Data 7**.

For inter-protein contact prediction, we adopted the same datasets used in 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). We also constructed two independent test sets comprising 43 homodimer targets (CASP_Homodimer43) and 20 heterodimer targets (CASP_Heterodimer20) collected from CASP13 to CASP16 to further assess model robustness and real-world applicability (**Supplementary Data 1 and 2**). Two sets of PPLM-Contact models were trained separately for homodimers and heterodimers: the homodimer model was trained exclusively on homodimers, whereas the heterodimer model was trained on the full combined training set.

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. The preprocessing workflows for paired sequences in PPLM and in ESM2's single-chain approach are illustrated in **Supplementary Figure S15**.

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)[37] 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.

PPLM was initialized from the 650M-parameter ESM2 model[6] and trained using four NVIDIA A100 GPUs with a gradient accumulation step of 32, for a total of 50,000 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 applied independently along the sequence dimension, forming two parallel representation branches. Each pooled representation is projected into a unified dimensional space through a dedicated linear layer and subsequently passed into its own multilayer perceptron (MLP) followed by a sigmoid activation to produce an interaction probability. Each MLP consists of five linear layers, where the first four layers are followed by layer normalization and a ReLU activation. The final interaction score is obtained by averaging the outputs of the two branches. The dimensionality parameters used at each transformation step are summarized in **Tables S20–21**.

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 $y = 0\ or\ 1$ 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[51] database using HHblits[52] with hyperparameters set to '-e 1e-3 -cov 0.4 -id 99'. For heterodimers, paired MSAs are constructed based on species annotation[40], 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)[53], row scores and average product corrections from direct coupling analysis (DCA)[54], as well as MSA embeddings and attention matrices derived from ESM-MSA-1b[55]. Lastly, the monomer distance maps are extracted from either experimentally determined (when available) or predicted monomer structures (by AlphaFold2[56]). The dimensions of all features are detailed in **Table S22**. 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 S16**). 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 S17**). Specifically, for a given inter-protein residue pair z_{ij}, the module aggregates signals from an intermediate residue n, utilizing intra-chain interactions r_{in} and cross-chain interactions z_{nj}, 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:

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

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

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 S18**). 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.,

$$q^h_{ij} = \mathcal{L}(z_{ij});\ \ k^h_{in} = \mathcal{L}(r_{in});\ v^h_{in} = \mathcal{L}(r_{in}); b^h_{ij} = \mathcal{L}(z_{ij}); g^h_{ij} = \sigma\left(\mathcal{L}(z_{ij})\right) \tag{6}$$

$$a^h_{ijk} = \mathrm{softmax}\left(\frac{1}{\sqrt{c}} q^h_{ij} \cdot k^h_{in} + b^h_{ij}\right) \tag{7}$$

$$\tilde{z}_{ij} = \mathcal{L}\left(g^h_{ij} \odot \sum_{n=1}^{L} a^h_{ijk} \cdot v^h_{in}\right) \tag{8}$$

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 S19**). Specifically, it attends over all inter-protein pairs z_{im} that share residue *i* (or z_{jm} 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:

$$q^h_{ij} = \mathcal{L}(z_{ij});\ k^h_{im} = \mathcal{L}(z_{im});\ v^h_{in} = \mathcal{L}(z_{im});\ g^h_{ij} = \sigma\left(\mathcal{L}(z_{ij})\right) \tag{9}$$

$$a^h_{ijk} = \mathrm{softmax}\left(\frac{1}{\sqrt{c}} q^h_{ij} \cdot k^h_{im} \cdot d_{im}\right) \tag{10}$$

$$\tilde{z}_{ij} = \mathcal{L}\left(g^h_{ij} \odot \sum_{m-1}^{L} a^h_{ijk} \cdot v^h_{im}\right) \tag{11}$$

where d_{im} 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[57]. 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[58] 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/. Source Data are provided with this paper.

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 was supported in part by the Ministry of Education, Singapore (T1251RES2309 and T2EP20125-0039), the Agency for Science, Technology and Research (A*STAR), Singapore (IAF-PP H25J6a0034), and the National Medical Research Council Open Fund – Young Individual Research Grant, Singapore (MOH-OFYIRG25jan-0011). 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. Bernett, J., Blumenthal, D.B. & List, M. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics* **25**, bbae076 (2024).

22. 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).
23. 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).
24. Meda, R.S. & Farimani, A.B. BAPULM: Binding Affinity Prediction using Language Models. *arXiv preprint arXiv:2411.04150* (2024).
25. Gorantla, R. et al. Learning Binding Affinities via Fine-tuning of Protein and Ligand Language Models. *bioRxiv*, 2024.2011. 2001.621495 (2024).
26. Siebenmorgen, T. & Zacharias, M. Computational prediction of protein–protein binding affinities. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **10**, e1448 (2020).
27. Guo, Z. & Yamaguchi, R. Machine learning methods for protein-protein binding affinity prediction in protein design. *Frontiers in Bioinformatics* **2**, 1065703 (2022).
28. Liu, H. et al. PPB-Affinity: Protein-Protein Binding Affinity dataset for AI-based protein drug discovery. *Scientific Data* **11**, 1-11 (2024).
29. 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).
30. 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).
31. Zeng, H. et al. ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic acids research* **46**, W432-W437 (2018).
32. 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).
33. 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).
34. Si, Y. & Yan, C. Protein language model-embedded geometric graphs power inter-protein contact prediction. *Elife* **12**, RP92184 (2024).
35. Xie, Z. & Xu, J. Deep graph learning of inter-protein contacts. *Bioinformatics* **38**, 947-953 (2022).
36. Rao, R.M. et al. in International Conference on Machine Learning 8844-8856 (PMLR, 2021).
37. Su, J. et al. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
38. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *biorxiv*, 2021.2010. 2004.463034 (2021).
39. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1-3 (2024).
40. 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).
41. Ko, Y.S., Parkinson, J., Liu, C. & Wang, W. TUnA: an uncertainty-aware transformer model

for sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics* **25** (2024).

42. Chatterjee, A. et al. Improving the generalizability of protein-ligand binding predictions with AI-Bind. *Nature communications* **14**, 1989 (2023).
43. Wang, Y. et al. ZeroBind: a protein-specific zero-shot predictor with subgraph matching for drug-target interactions. *Nature Communications* **14**, 7861 (2023).
44. 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).
45. 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).
46. 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).
47. Burley, S.K. et al. Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, 627-641 (2017).
48. 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).
49. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026-1028 (2017).
50. Zhang, C., Shine, M., Pyle, A.M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature methods* **19**, 1109-1115 (2022).
51. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* **45**, D170-D176 (2017).
52. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics* **20**, 1-15 (2019).
53. 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).
54. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130 (2014).
55. 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).
56. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *nature* **596**, 583-589 (2021).
57. Studer, G., Tauriello, G. & Schwede, T. Assessment of the assessment—All about complexes. *Proteins: Structure, Function, and Bioinformatics* **91**, 1850-1860 (2023).
58. Lin, T. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:1708.02002* (2017).

Figures

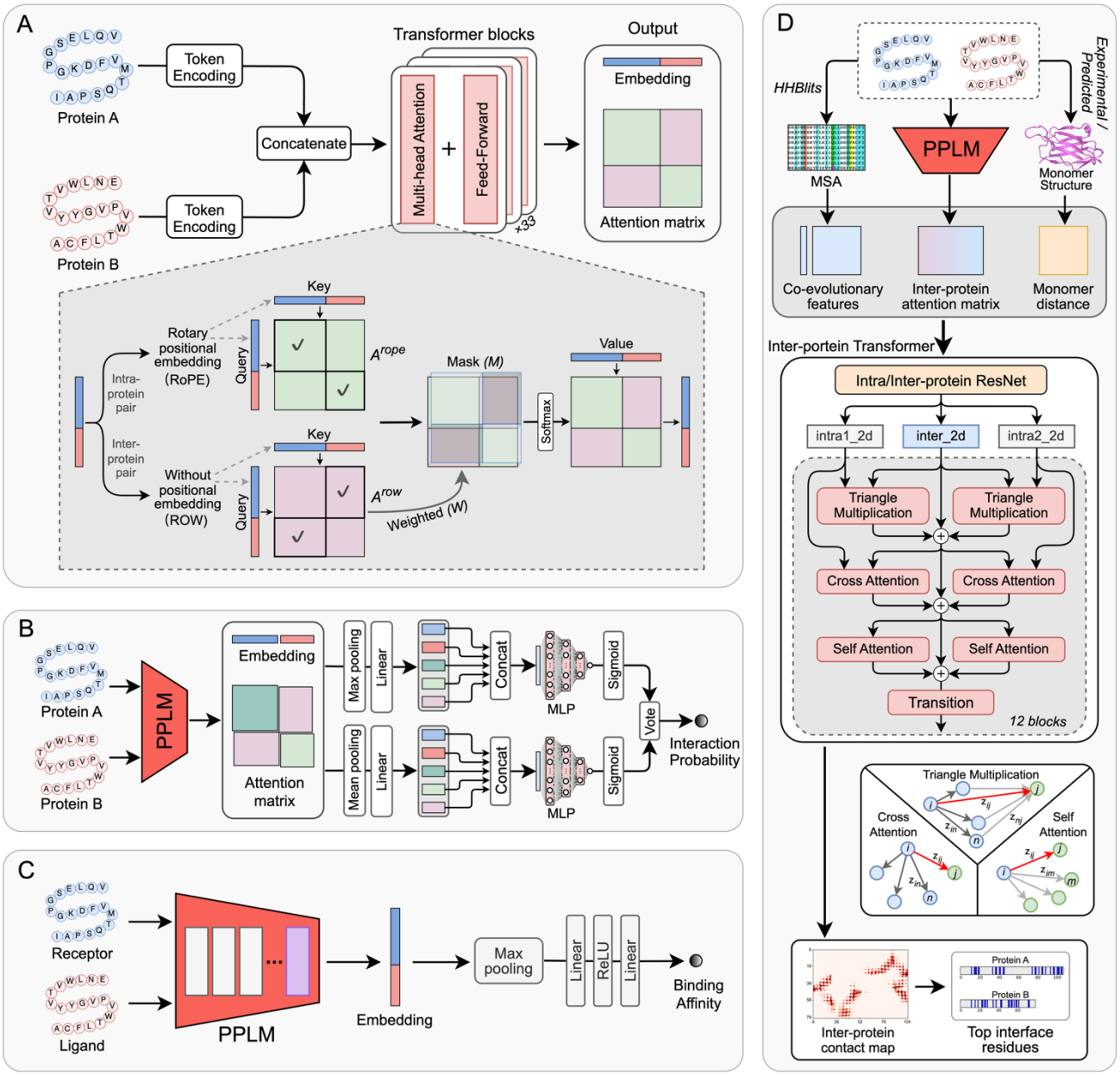

Figure 1. Overview of the PPLM framework and its downstream applications. (A) PPLM architecture. The input protein sequence pair are independently tokenized and then processed through transformer blocks equipped with a cross-protein attention mechanism to capture intra- and inter-protein interactions. The resulting embeddings and attention matrixes can be leveraged for diverse PPI tasks. **(B)** PPLM-PPI pipeline. For a given sequence pair, PPLM is used to generate the embeddings, intra- and inter-protein attention matrix. These representations are aggregated using max and mean pooling, each followed by a multilayer perceptron to estimate the interaction probability. The final prediction is obtained by averaging the probabilities from the two pooling branches. **(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 pair, inter-protein attention matrix from PPLM is integrated with MSA-derived features and monomer distance maps to capture both evolutionary and structural information. These combined 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 module adopts a parallel architecture to update inter-protein representations. The top interface residues (top 50 displayed) were extracted from the top-ranked predicted contact pairs.

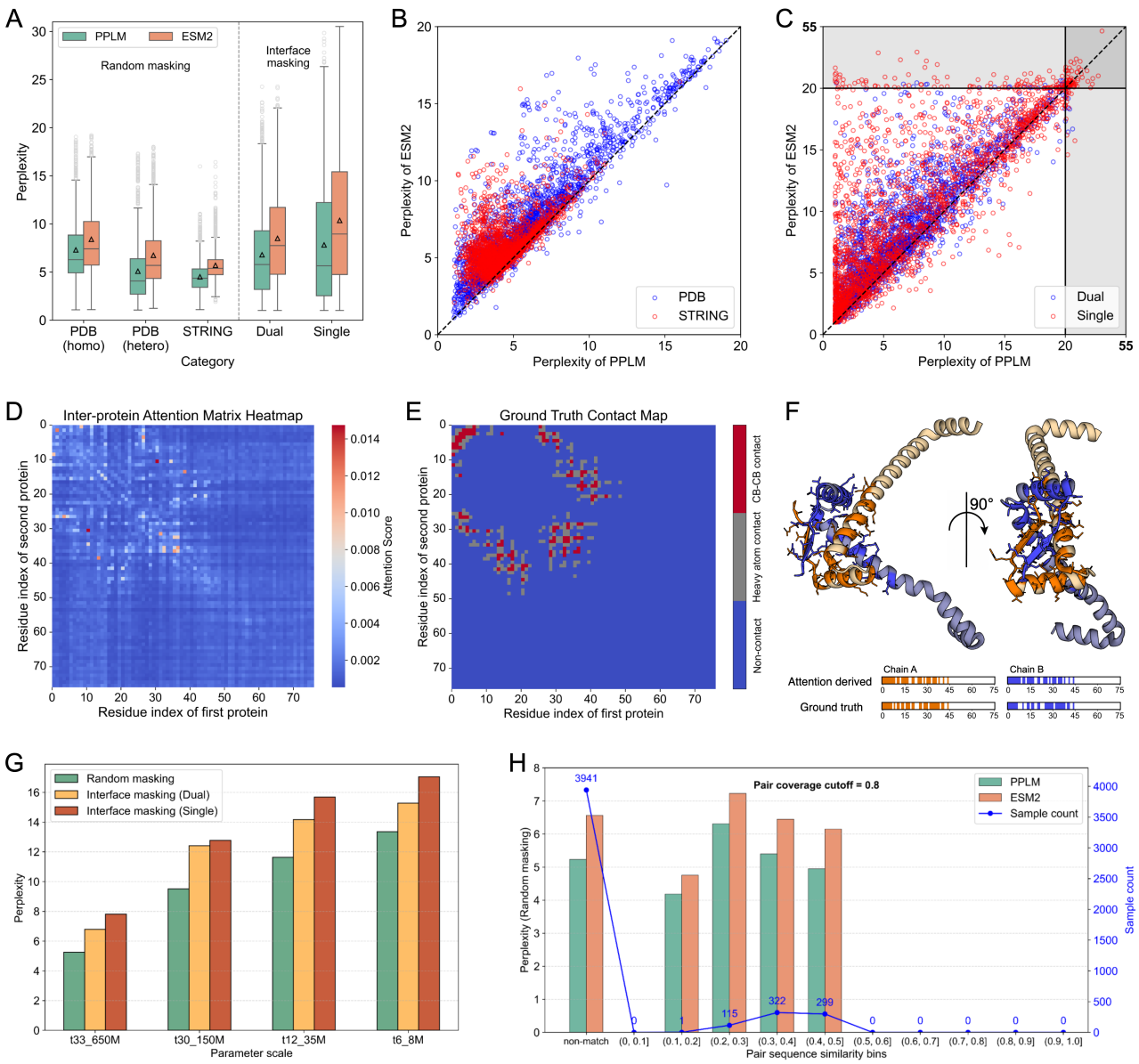

Figure 2. Perplexity of PPLM and ESM2 on protein sequence pairs. **(A)** Perplexity distribution for sequence pairs from the PDB and STRING databases under random masking and interface masking strategies. Interface masking is applied only to sequence pairs from PDB database, with single-chain interface 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. **(D-F)** Example of a conserved putative transcription factor from Vibrio cholerae O1 biovar El Tor (PDB ID: 1Y9B). **(D)** Heatmap of the inter-protein attention matrix generated by PPLM. **(E)** Gound-truth contact map extracted from the experimental structure. **(F)** Three-dimensional virtualization of the complex, with interface residues identified from the inter-protein attention matrix highlighted. **(G)** Perplexities of PPLM across different parameter scales. **(H)** Perplexities of PPLM and ESM2 under varying pairwise sequence similarity to the training set. The number of samples within each similarity bin is indicated.

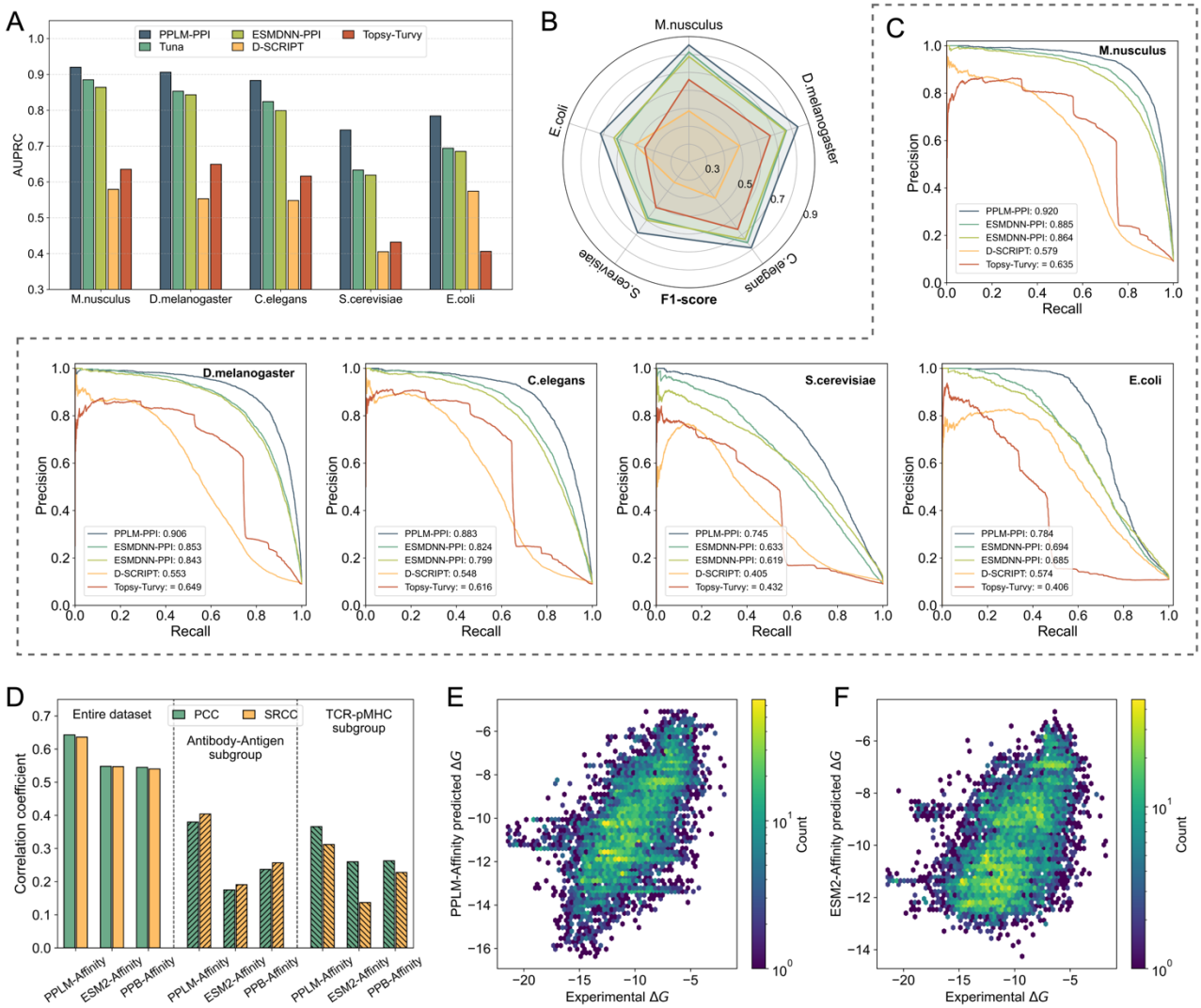

Figure 3. Performance of PPLM-PPI and PPLM-Affinity on protein–protein interaction and binding affinity prediction. (A) AUPRC of PPLM-PPI, TUnA, ESMDNN-PPI, D-SCRIPT, and Topsy-Turvy across five species test sets. **(B)** F1-score of the five PPI prediction methods on the five species test sets. **(C)** Precision-recall curves of the five 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-F)** Head-to-head comparison between experimental and predicted binding affinity for PPLM-Affinity and ESM2-Affinity, respectively. Plots show hexagonal density maps to improve visibility at high point density. Color intensity reflects the local log-transformed point density.

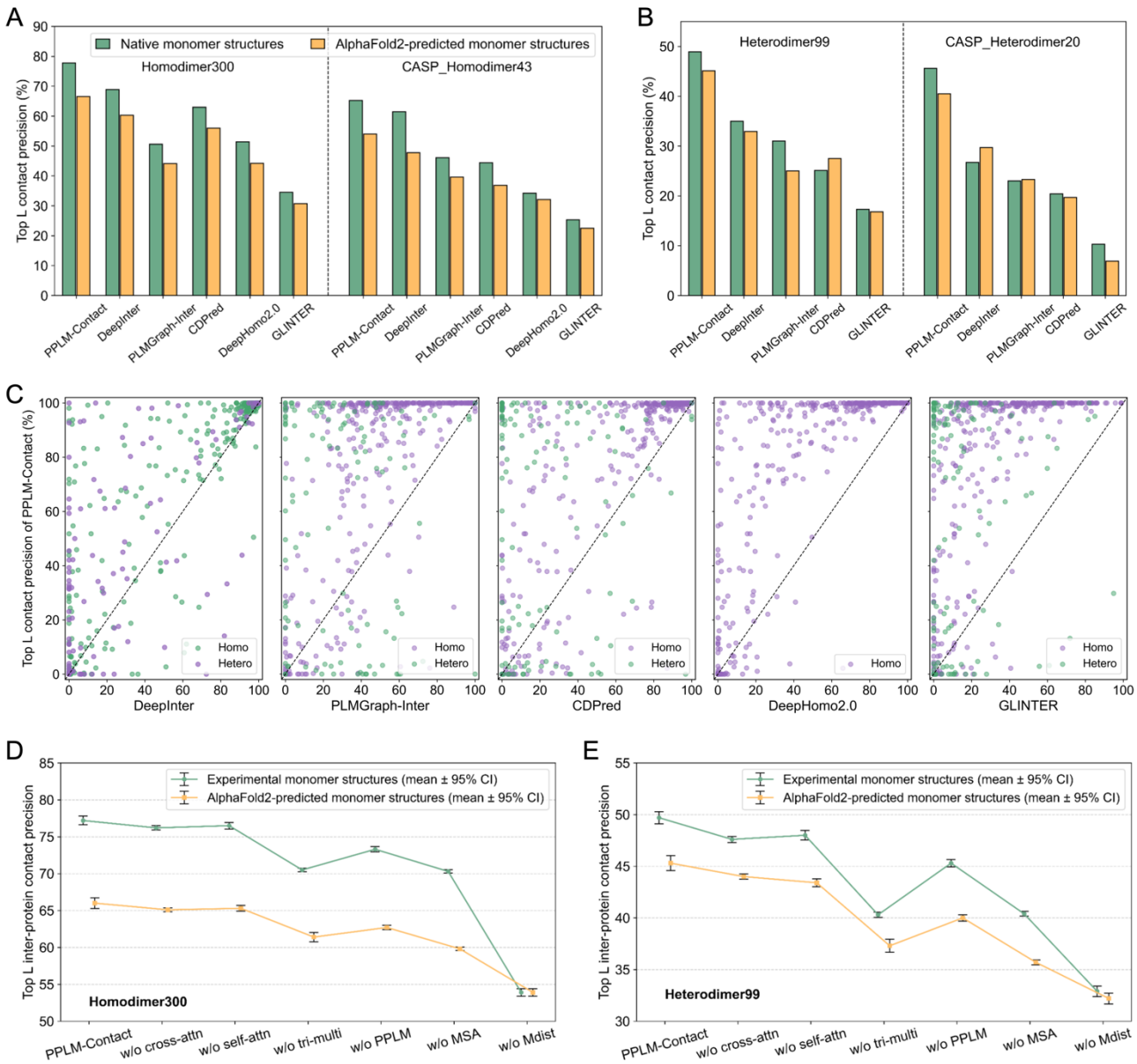

Figure 4. Performance of PPLM-Contact and the comparison methods on the inter-protein contact prediction. (A) Top *L* contact precision of PPLM-Contact, DeepInter, CDPred, DeepHomo2.0, and GLINTER on the Homodimer300 and CASP_Homodimer43 test sets using both experimental monomer structures and AlphaFold2-predicted 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–E)** Mean top *L* contact precision and 95% confidence intervals (CI) of PPLM-Contact and its ablation variants across three seeds on the Homodimer300 and Heterodimer99 test sets, respectively.

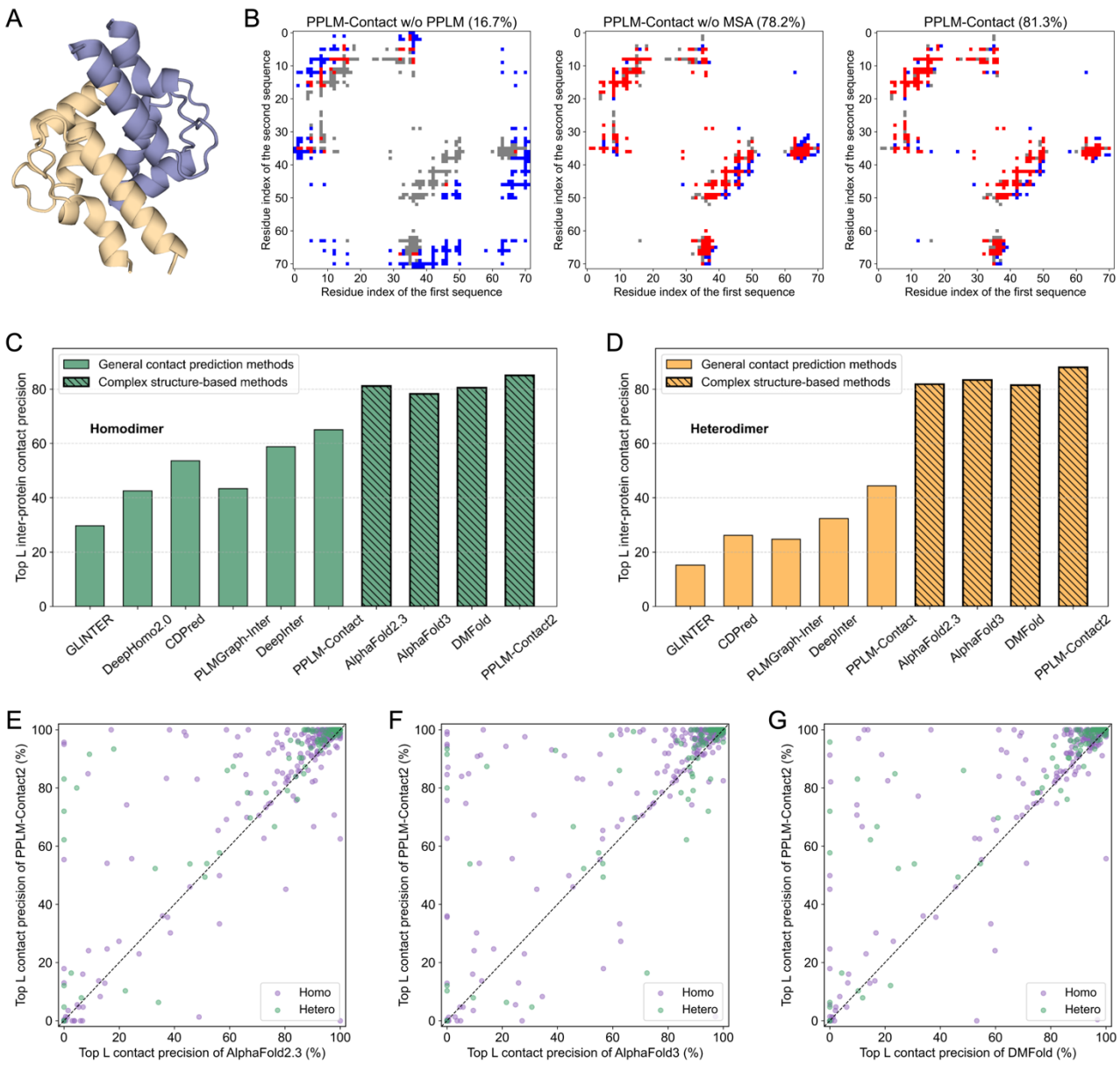

Figure 5. (A) Example of inter-protein contact prediction for the EF-hand domain of human RASEF (PDB ID: 2PMY). **(B)** Contact maps for 2PMY predicted by PPLM-Contact w/o PPLM, PPLM-Contact w/o MSA, and the full PPLM-Contact models, respectively. Each contact map shows the alignment between ground-truth 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. **(C–D)** top *L* contact precision of general contact-prediction methods (GLINTER, DeepHomo2.0, CDPred, PLMGraph-Inter, DeepInter, and PPLM-Contact) and complex structure–based approaches (AlphaFold2.3, AlphaFold3, DMFold, and PPLM-Contact2) on 343 homodimers and 119 heterodimers, respectively. **(E–G)** Pairwise comparisons of top *L* contact precision between PPLM-Contact2 and AlphaFold2.3, AlphaFold3, and DMFold.

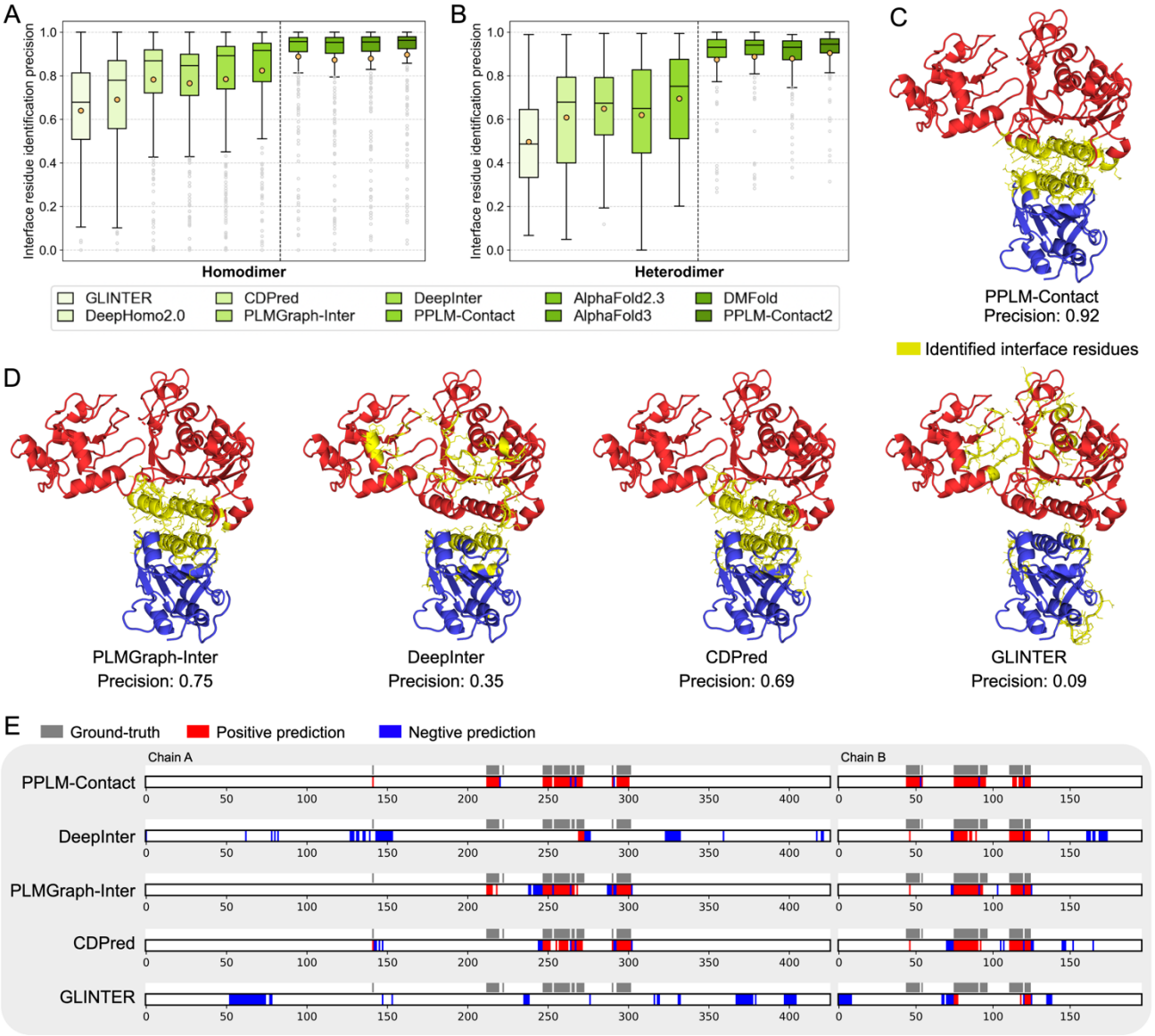

Figure 6. Comparative analysis of protein–protein interface residue identification. (A-B) Precision of interface-residue identification by general contact-prediction methods (GLINTER, DeepHomo2.0, PLMGraph-Inter, CDPred, DeepInter, and PPLM-Contact) and complex structure–based approaches (AlphaFold2.3, AlphaFold3, DMFold, and PPLM-Contact2) on 343 homodimers and 119 heterodimers, respectively. **(C-D)** Crystal structure of the DNMT3A–DNMT3L complex (PDB: 4U7P), with identified interface residues from inter-protein contact-prediction methods highlighted in yellow and shown in stick representation. **(E)** Residue-level annotation maps for 4U7P, where grey bars indicate ground-truth interface residues, red bars denote true positives, and blue bars denote false positives.