Computational and Systems Biology

Protein language model embedded geometric graphs power inter-protein contact prediction

Yunda Si, Chengfei Yan

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

https://en.wikipedia.org/wiki/Open_access

Copyright information

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
March 14, 2024 (this version)

Reviewed preprint version 1
December 12, 2023

Sent for peer review
September 4, 2023

Posted to preprint server
August 24, 2023

Abstract

Accurate prediction of contacting residue pairs between interacting proteins is very useful for structural characterization of protein-protein interactions (PPIs). Although significant improvement has been made in inter-protein contact prediction recently, there is still large room for improving the prediction accuracy. Here we present a new deep learning method referred to as PLMGraph-Inter for inter-protein contact prediction. Specifically, we employ rotationally and translationally invariant geometric graphs obtained from structures of interacting proteins to integrate multiple protein language models, which are successively transformed by graph encoders formed by geometric vector perceptrons and residual networks formed by dimensional hybrid residual blocks to predict inter-protein contacts. Extensive evaluation on multiple test sets illustrates that PLMGraph-Inter outperforms five top inter-protein contact prediction methods, including DeepHomo, GLINTER, CDPred, DeepHomo2 and DRN-1D2D_Inter by large margins. In addition, we also show that the prediction of PLMGraph-Inter can complement the result of AlphaFold-Multimer. Finally, we show leveraging the contacts predicted by PLMGraph-Inter as constraints for protein-protein docking can dramatically improve its performance for protein complex structure prediction.

eLife assessment

This study presents a **useful** deep learning-based inter-protein contact prediction method named PLMGraph-Inter which combines protein language models and geometric graphs. The evidence supporting the claims of the authors is **solid**. The authors show that their approach may be **useful** in some cases where AlphaFold-Multimer performs poorly. This work will be of interest to researchers working on protein complex structure prediction, particularly when accurate experimental structures are available for one or both of the monomers in isolation.

Introduction

Protein-protein interactions(PPIs) are essential activities of most cellular processes(Alberts, 1998; Spirin & Mirny, 2003). Structure characterization of PPIs is important for mechanistic investigation of these cellular processes and therapeutic development(Goodsell & Olson, 2000).

However, currently experimental structures of many important PPIs are still missing as experimental methods to resolve complex structures such as X-ray crystallography, nuclear magnetic resonance, cryo-electron microscopy are costly and time-consuming(Berman et al., 2000). Therefore, it is necessary to develop computational methods to predict protein complex structures(Bonvin, 2006). Predicting contacting residue pairs between interacting proteins can be considered as an intermediate step for protein complex structure prediction(Hopf et al., 2014; Ovchinnikov et al., 2014), as the predicted contacts can be integrated into protein-protein docking algorithms to assist protein complex structure prediction(Dominguez et al., 2003; H. Li & Huang, 2021; Sun et al., 2020). Besides, the predicted contacts can also be very useful to guide protein interfacial design(Martino et al., 2021) and the inter-protein contact prediction methods can be further extended to predict novel PPIs (Cong et al., 2019; Green et al., 2021).

Based on the fact that contacting residue pairs often vary co-operatively during evolution, coevolutionary analysis methods(Weigt et al., 2009) have been used in previous studies to predict inter-protein contacts(Hopf et al., 2014; Ovchinnikov et al., 2014). However, coevolutionary analysis methods do have certain limitations. For examples, effective coevolutionary analysis requires a large number of interolog sequences, which are often difficult to obtain, especially for heteromeric PPIs(R. M. Rao et al., 2021a); and it is difficult to distinguish inter-protein and intra-protein coevolutionary signals for homomeric PPIs(Uguzzoni et al., 2017). Inspired by its great success in intra-protein contact prediction(Hanson et al., 2018; Ju et al., 2021; Y. Li et al., 2019; Si & Yan, 2021; Wang et al., 2017), deep learning has also been applied to predict inter-protein contacts(Guo et al., 2022; Roy et al., 2022; Xie & Xu, 2022; Yan & Huang, 2021; Zeng et al., 2018). ComplexContact(Zeng et al., 2018), to the best of our knowledge, the first deep learning method for inter-protein contact prediction, has significantly improved the prediction accuracy over coevolutionary analysis methods. However, its performance on eukaryotic PPIs is still quite limited, partly due to the difficulty to accurately infer interologs for eukaryotic PPIs. In a later study, coming from the same group as ComplexContact, Xie et al. developed GLINTER(Xie & Xu, 2022), another deep learning method for inter-protein contact prediction. Comparing with ComplexContact, GLINTER leverages structures of interacting monomers, from which their rotational invariant graph representations are used as additional input features. GLINTER outperforms ComplexContact in the prediction accuracy, although there is still large room for improvement, especially for heteromeric PPIs. It is worth mentioning that CDPred(Guo et al., 2022), a recently developed method, further surpasses GLINTER in prediction accuracy with 2D attention-based neural networks. Apart from these methods developed to predict inter-protein contacts for both homomeric and heteromeric PPIs, inter-protein contact prediction methods specifically for homomeric PPIs were also developed(Roy et al., 2022; T. Wu et al., 2022; Yan & Huang, 2021), as predicting the inter-protein contacts for homomeric PPIs is generally much easier due to the symmetric restriction, relatively larger interfaces and the trivialness of interologs identification. For example, Yan et al. developed DeepHomo(Yan & Huang, 2021), a deep learning method specifically to predict inter-protein contacts of homomeric PPIs, which also significantly outperforms coevolutionary analysis-based methods. However, DeepHomo requires docking maps calculated from structures of interacting monomers, which is computationally expensive and is also sensitive to the quality of monomeric structures. Besides, coming from the same group, Lin et al. further developed DeepHomo2(Lin et al., 2023) for inter-protein contact prediction for homomeric PPIs by including the MSA (multiple sequence alignment) embeddings and attentions from an MSA-based protein language model (MSA transformer)(R. M. Rao et al., 2021b) in their prediction model, which further improved the prediction performance. In almost the same time with DeepHomo2, we proved that embeddings from protein language models(R. Rao et al., 2021; Rives et al., 2021) (PLMs) are very effective features to predict inter-protein contacts for both homomeric and heteromeric PPIs, and we further show the sequence embeddings (ESM-1b(Rives et al., 2021)), MSA embeddings (ESM-MSA-1b(R. M. Rao et al., 2021b) & Position-Specific Scoring Matrix (PSSM)) and the inter-protein coevolutionary information complement each other in the prediction, with which we

developed DRN-1D2D_Inter(Si & Yan, 2023). Extensive benchmark results show that DRN-1D2D_Inter significantly outperforms DeepHomo and GLINTER in inter-protein contact prediction, although DRN-1D2D_Inter makes the prediction purely from sequences.

In this study, we developed a structure-informed method to predict inter-protein contacts. Given the structures of two interacting proteins, we first build rotationally and translationally (SE(3)) invariant geometric graphs from the two monomeric structures, which encode both the inter-residue distance and orientation information of the monomeric structures. We further embedded the single sequence embeddings (ESM-1b), MSA embeddings (ESM-MSA-1b & PSSM) and structure embeddings (ESM-IF(Hsu et al., 2022a)) from PLMs in the graph nodes of the corresponding residues to build the PLM embedded geometric graphs, which are then transformed by graph encoders formed by geometric vector perceptrons to generate graph embeddings for interacting monomers. The graph embeddings are further combined with inter-protein pairwise features and transformed by residual networks formed by dimensional hybrid residual blocks (residual block hybridizing 1D and 2D convolutions) to predict inter-protein contacts. The developed method referred to as PLMGraph-Inter was extensive benchmarked on multiple tests with application of either experimental or predicted structures of interacting monomers as the input. The result shows that in both cases, PLMGraph-Inter outperforms other top prediction methods including DeepHomo, GLINTER, CDPred, DeepHomo2 and DRN-1D2D_Inter by large margins. In addition, we also compared the prediction results of PLMGraph-Inter with the protein complex structures generated by AlphaFold-Multimer(Evans et al., 2022a). The result shows that for many targets which AlphaFold-Multimer made poor predictions, PLMGraph-Inter yielded better results. Finally, we show leveraging the contacts predicted by PLMGraph-Inter as constraints for protein-protein docking can dramatically improve its performance for protein complex structure prediction.

Results

Overview of PLMGraph-Inter

The method of PLMGraph-Inter is summarized in **Figure 1a**. PLMGraph-Inter consists of three modules: the graph representation module (**Figure 1b**), the graph encoder module (**Figure 1c**) and the residual network module (**Figure 1d**). Each interacting monomer is first transformed into a PLM-embedded graph by the graph representation module, then the graph is passed through the graph encoder module to obtain a 1D representation of each protein. The two protein representations are transformed into 2D pairwise features through outer concatenation (horizontal & vertical tiling followed by concatenation) and further concatenated with other 2D pairwise features including the inter-protein attention maps and the inter-protein co-evolution matrices, which are then transformed by the residual network module to obtain the predicted inter-protein contact map.

The graph representation module

The first step of the graph representation module is to represent the protein 3D structure as a geometric graph, where each residue is represented as a node, and an edge is defined if the C_α atom distance between two residues is less than 18Å. For each node and edge, we use scalars and vectors extracted from the 3D structures as their geometric features. To make the geometric graph SE(3) invariant, we use a set of local coordinate systems to extract the geometric vectors. The SE(3) invariance of representation of each interacting monomer is important, as in principle, the inter-protein contact prediction result should not depend on the initial positions and orientations of protein structures. A detailed description can be found in the Methods section. The second step is to integrate the single sequence embedding from ESM-1b(Rives et al., 2021), the MSA embedding from ESM-MSA-1b(R. Rao et al., 2021), the Position-Specific Scoring Matrix (PSSM) calculated from the MSA and structure embedding from ESM-IF(Hsu et al., 2022b) for each interacting

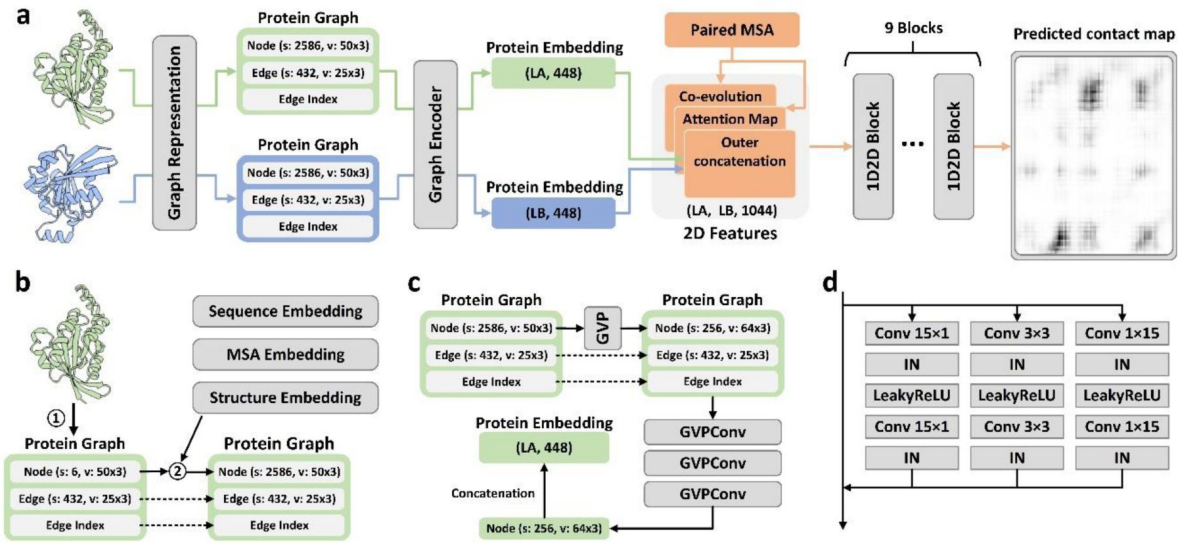


Figure 1.

Overview of PLMGraph-Inter. (a) The network architecture of PLMGraph-Inter. (b) The graph representation module. (c) The graph encoder module, s denotes scalar features, v denotes vector features. (d) The dimensional hybrid residual block (“IN” denotes Instance Normalization).

monomer using its corresponding geometric graph. Where ESM-1b and ESM-MSA-1b are pretrained PLMs learned from large datasets of sequences and MSAs respectively with masked language modeling tasks, and ESM-IF is a supervised PLM trained from 12 million protein structures predicted by AlphaFold2(Jumper et al., 2021) for fixed backbone design. The embeddings from these models contain high dimensional representations of each residue in the protein, which are concatenated and further combined with the PSSM to form additional features of each node in the geometric graph. Since the sequence embeddings, the MSA embeddings, the PSSM and the structure embeddings are all SE(3) invariant, the PLM-embedded geometric graph of each protein is also SE(3) invariant.

The graph encoder module

The graph encoder module is formed by geometric vector perceptron (GVP) and GVP convolutional layer (GVPConv)(Jing, Eismann, Soni, et al., 2021; Jing, Eismann, Suriana, et al., 2021). Where GVP is a graph neural network module consisting of a scalar track and a vector track, which can perform rotationally invariant transformations on scalar features and rotationally equivariant on vector features of nodes and edges; GVPConv follows the message passing paradigm of graph neural network and mainly consists of GVP, which updates the embedding of each node by passing information from its neighboring nodes and edges. A detailed description of GVP and GVPConv can be found in the Methods section and also in the work of GVP(Jing, Eismann, Soni, et al., 2021; Jing, Eismann, Suriana, et al., 2021). For each protein graph, we first use a GVP module to reduce the dimension of the scalar features of each node from 2586 to 256, which is then transformed successively by three GVPConv layers. Finally, we stitch the scalar features and the vector features of each node to form the 1D representation of the protein. Since the input protein graph is SE(3) invariant and the GVP and GVPConv transformations are rotationally equivariant, the 1D representation of each interacting monomer is also SE(3) invariant.

The residual network module

The residual network module is mainly formed by 9 dimensional hybrid residual blocks to transform the 2D feature maps to obtain the predicted inter-protein contact map. Our previous study illustrated the effective receptive field can be enlarged with the application of the dimensional hybrid residual block, thus helps improve the model performance(Si & Yan, 2021). A more detailed description of the transforming procedure can be found in the Methods section.

Evaluation of PLMGraph-Inter on HomoPDB and HeteroPDB test sets

We first evaluated PLMGraph-Inter on two self-built test sets which are non-redundant to the training dataset of PLMGraph-Inter: HomoPDB and HeteroPDB. Where HomoPDB is the test set for homomeric PPIs containing 400 homodimers and HeteroPDB is the test set for heteromeric PPIs containing 200 heterodimers. For comparison, we also evaluated DeepHomo, GLINTER, DeepHomo2, CDPred and DRN-1D2D_Inter on the same datasets. Since DeepHomo and DeepHomo2 was developed to predict inter-protein contacts only for homomeric PPIs, its evaluation was only performed on HomoPDB. It should be noted that since HomoPDB and HeteroPDB are not de-redundant with the training sets of DeepHomo, DeepHomo2, CDPred and GLINTER, the performances of the four methods may be overestimated.

In all the evaluations, the structural related features were drawn from experimental structures of interacting monomers separated from complex structures of PPIs after randomizing their initial positions and orientations (DRN-1D2D_Inter does not use structural information). Besides, we also used the AlphaFold2 predicted monomeric structures as the input, considering experimental structures of interacting monomers often do not exist. Since the interacting monomers in HomoPDB and HeteroPDB are not de-redundant with the training set of AlphaFold2, using default settings of AlphaFold2 may overestimate its performance. To mimic the performance of

AlphaFold2 in real practice and produce predicted monomeric structures with more diverse qualities, we only used the MSA searched from Uniref100(Suzek et al., 2015) protein sequence database as the input of AlphaFold2 and set to not use the template. The predicted structures yielded a mean TM-score 0.88, which is close to the performance of AlphaFold2 for CASP14 targets (mean TM-score 0.85)(Z. Lin et al., 2023).

Table 1 shows the mean precision of each method on the HomoPDB and HeteroPDB when top (5, 10, 50, L/10, L/5) predicted inter-protein contacts are considered, where L denotes the sequence length of the shorter protein in the PPI and (Note: GLINTER encountered errors for 81 (5 when using the predicted monomeric structures) targets in HomoPDB and 15 (3 when using the predicted monomeric structures) targets in HeteroPDB at run time and did not produce predictions, thus we removed these targets in the evaluation of the performance of GLINTER. The performances on HomoPDB and HeteroPDB for these methods after the removal of these targets which GLINTER failed in any case are shown in **Table S1**). As can be seen from the table, whenever the experimental or the predicted monomeric structures were used as the input, the mean precision of PLMGraph-Inter far exceeds those of other algorithms in each metric for both datasets. Particularly, the mean precision of PLMGraph-Inter is substantially improved in each metric on each dataset compared to our previous method DRN-1D2D_Inter which used most features of PLMGraph-Inter except these drawn from structures of the interacting monomers, illustrating the importance of the inclusion of structural information. Besides, GLINTER, CDPred and DeepHomo2 also use structural information and PLMs, but have much lower performance than PLMGraph-Inter, illustrating the efficacy of our deep learning framework.

It can also be seen from the result that all the methods tend to have better performances on HomoPDB than those on HeteroPDB. One possible reason is that the complex structures of homodimers are generally C2 symmetric, which largely restricts the configurational spaces of PPIs, making the inter-protein contact prediction a relatively easier task (e.g., the inter-protein contact maps for homodimers are also symmetric). Besides, comparing with heteromeric PPIs, we may more likely to successfully infer the inter-protein coevolutionary information for homomeric PPIs to assist the contact prediction for two reasons: First, it is straightforward to pair sequences in the MSAs for homomeric PPIs, thus the paired MSA for homomeric PPIs may have higher qualities; second, homomeric PPIs may undergo stronger evolutionary constraints, as homomeric PPIs are generally permanent interactions, but many heteromeric PPIs are transient interactions.

In addition to using the mean precision on each test set to evaluate the performance of each method, the performance comparisons between PLMGraph-Inter and other models on the top 50 predicted contacts for each individual target in HomoPDB and HeteroPDB are shown in **Figure 2** (Separate comparisons are shown in **Figure S1** and **S2**). Specifically, when the experimental (predicted) structures were used as the input, PLMGraph-Inter achieved the best performance for 60% (53%) of the targets in HomoPDB and 58% (51%) of the targets in HeteroPDB. We further group targets in each dataset according to their inter-protein contact densities defined as $\frac{\#\ of\ contacts}{L_A * L_B}$ and the normalized number of the effective sequences (N_{eff}^{norm}) of paired MSAs. We found that all the methods tend to have lowers performances for targets with lower contact densities (**Figure S3**), which is reasonable, since obviously it is more challenging to identify the true contacts when their ratio is lower. We also found when the N_{eff}^{norm} is low ($\log(N_{eff}^{norm})<3$), the prediction performances of all methods tend to improve with N_{eff}^{norm}, but when N_{eff}^{norm} reaches to certain thresholds ($\log(N_{eff}^{norm})>4$), the performances of all the methods tends to fluctuate with N_{eff}^{norm} (**Figure S4**). However, PLMGraph-Inter consistently achieved the best performances in most categories.

Methods	HomoPDB (precision %)					HeteroPDB (precision %)				
	L/5	L/10	50	10	5	L/5	L/10	50	10	5
DeepHomo	43.2 (39.3)	46.7 (42.7)	42.4 (38.8)	48.5 (44.8)	49.9 (46.2)					
GLINTER	42.9 (47.3)	45.0 (50.1)	42.2 (52.1)	46.4 (51.9)	48.5 (53.6)	23.9 (25.1)	24.7 (27.0)	20.9 (21.9)	25.5 (25.8)	26.7 (26.2)
DRN-1D2D_Inter	52.5	55.2	51.3	56.6	57.6	34.9	37.1	32.6	38.1	38.5
DeepHomo2	55.6 (52.4)	58.1 (53.9)	55.0 (51.7)	59.4 (55.7)	61.3 (56.7)					
CDPred	59.4 (54.7)	61.3 (56.2)	58.4 (54.1)	62.4 (57.1)	62.9 (57.7)	30.0 (30.2)	31.0 (31.7)	27.6 (27.3)	32.0 (32.2)	32.1 (32.7)
PLMGraph-Inter	**68.6 (61.8)**	**70.4 (63.6)**	**67.3 (60.9)**	**71.6 (65.0)**	**72.1 (65.25)**	**45.9 (41.9)**	**48.6 (43.6)**	**41.4 (37.8)**	**49.1 (44.1)**	**51.6 (45.0)**

Note: The highest mean precision (%) in each column is highlighted in bold.

Table 1.

The performances of DeepHomo, GLINTER, DRN-1D2D_Inter, CDPred, DeepHomo2 and PLMGraph-Inter on the HomoPDB and HeteroPDB test sets using experimental structures (AlphaFold2 predicted structures)

		HomoPDB		HeteroPDB	
		Count	Precision (Top 50 (%))	Count	Precision (Top 50 (%))
Sequence identity (MMSeqs2)	Original	400	67.3 (60.9)	200	41.4 (37.8)
	40%	341	68.7 (62.5)	160	38.6 (35.6)
	30%	257	64.7 (58.7)	144	38.1 (35.1)
	20%	211	63.2 (56.3)	138	38.5 (35.3)
	10%	211	63.2 (56.3)	138	38.5 (35.3)
Fold similarity (TM-align)	0.9	370	65.2 (58.3)	185	39.7 (35.7)
	0.8	281	61.8 (53.4)	153	38.1 (34.1)
	0.7	179	56.5 (45.8)	126	38.8 (34.6)
	0.6	124	50.4 (39.9)	102	37.4 (34.1)
	0.5	70	49.6 (41.3)	83	36.5 (34.5)

*The results using experimental structures are shown outside the parentheses, and the results using the AlphaFold2 predicted structures are shown in the parentheses.

Table 2.

The performance of PLMGraph-Inter when using different sequence identity and fold similarity thresholds to further remove potential redundancies in HomoPDB and HeteroPDB.

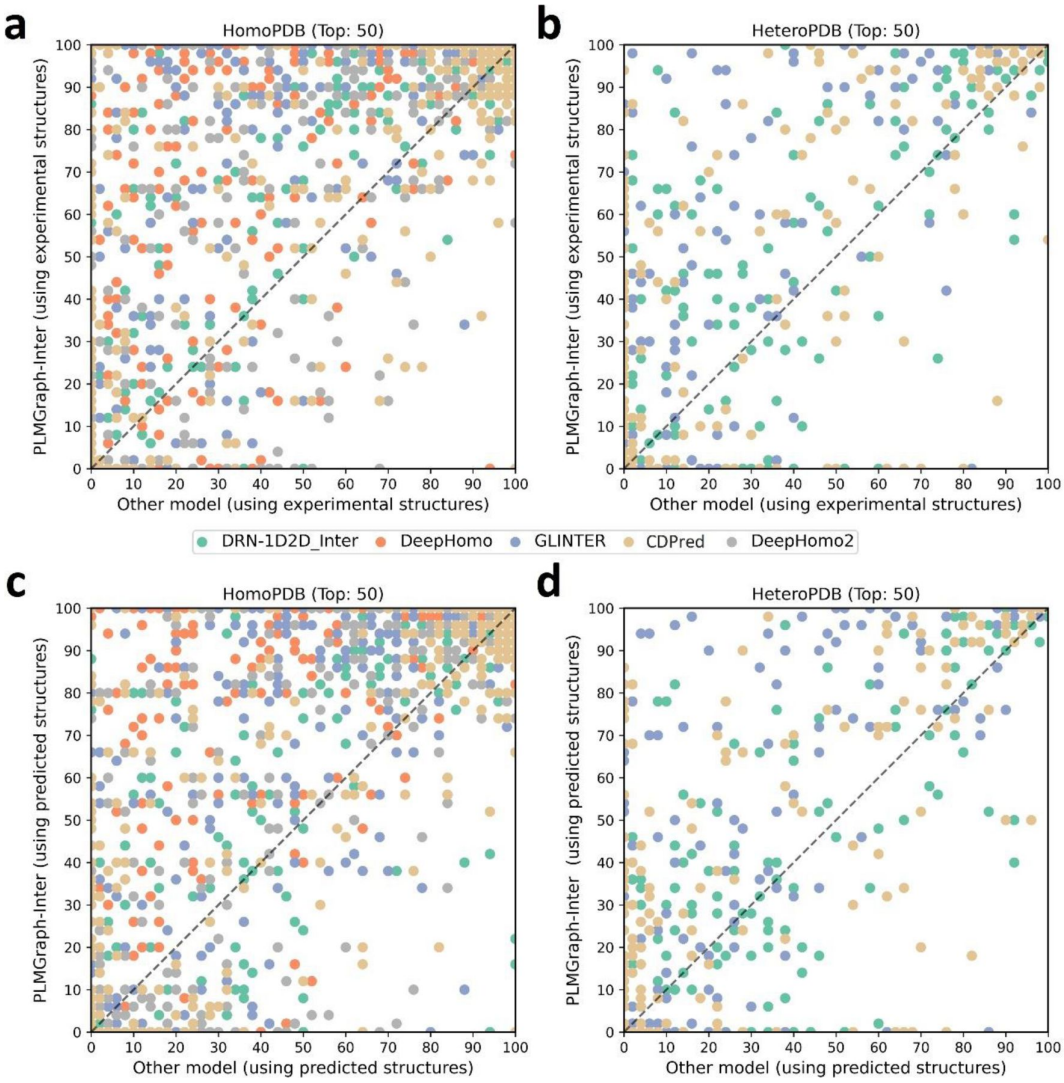


Figure 2.

The performances of PLMGraph-Inter and other methods on the HomoPDB and HeteroPDB test sets. (a)~(b): The head-to-head comparison of the precisions (%) of the top 50 contacts predicted by PLMGraph-Inter and other methods for each target in (a) HomoPDB and (b) HeteroPDB using experimental structures. (c)~(d): The head-to-head comparison of the precisions (%) of the top 50 contacts predicted by PLMGraph-Inter and other methods for each target in (c) HomoPDB and (d) HeteroPDB using AlphaFold2 predicted structures.

Impact of the monomeric structure quality on contact prediction

We further analyzed performance difference of PLMGraph-Inter when using the AlphaFold2 predicted structures and using the experimental structures as the inputs. As it can be seen from **Figures 3a-b**, when the predicted structures were used by PLMGraph-Inter for inter-protein contact prediction, mean precisions of the predicted inter-protein contacts in each metric on both HomoPDB and HeteroPDB test sets decrease by about 5% (also see **Table 1**), indicating qualities of the input structures do have certain impact on the prediction performance.

We further explored the impact of the monomeric structure quality on the inter-protein contact prediction performance of PLMGraph-Inter. Specifically, for a given PPI, the TM-score(Zhang & Skolnick, 2004) was used to evaluate the quality of the predicted structure for each interacting monomer, and the TM-score of the predicted structure with lower quality was used to evaluate the overall monomeric structure prediction quality for the PPI, denoted as "DTM-score". In **Figure 3c**, we show the performance gaps (using the mean precisions of the top 50 predicted contacts as the metric) between applying the predicted structures and applying the experimental structures in the inter-protein contact prediction, in which we grouped targets according to DTM-scores of their monomeric structure prediction, and in **Figure 3d**, we show the performance comparison for each specific target. From **Figure 3c**-**d**, we can clearly see that when the DTM-score is lower, the prediction using the prediction structure tends to have lower accuracy. However, when the DTM-score is greater than or equal to 0.8, there is almost no difference between the applying the predicted structures and applying the experimental structure, which shows the robustness of PLMGraph-Inter to the structure quality.

Ablation study

To explore the contribution of each input component to the performance of PLMGraph-Inter, we conducted ablation study on PLMGraph-Inter. The graph representation from the structure of each interacting proteins is the base feature of PLMGraph-Inter, so we first trained the baseline model using only the geometric graphs as the input feature, denoted as model a. Our previous study in DRN-1D2D_Inter has shown that the single sequence embeddings, the MSA 1D features (including the MSA embeddings and PSSMs) and the 2D pairwise features from the paired MSA play important roles in the model performance. To further explore the importance of these features when integrated with the geometric graphs, we trained model b-d separately (model b: geometric graphs + sequence embeddings, model c: geometric graphs + sequence embeddings + MSA 1D features, model d: geometric graphs + sequence embeddings + MSA 1D features + 2D features). Finally, we included the structure embeddings as additional features to train the model e (model e uses all the input features of PLMGraph-Inter). All the five models were trained using the same protocol as PLMGraph-Inter on the same training and validation partition without cross validation. We further evaluated performances of models a-e models together with PLMGraph-Inter (i.e., model f: model e + cross validation) on HomoPDB and HeteroPDB using experimental structures of interacting monomers respective.

In **Figure 4a**, we show the mean precisions of the top 50 predicted contacts by model a-f on HomoPDB and HeteroPDB respectively. It can be seen from **Figure 4a** that including the sequence embeddings in the geometric graphs has a very good boost to the model performance (model b versus model a), while the additional introduction of MSA 1D features and 2D pairwise features can further improve the model performance (model d versus model c versus model b). DRN-1D2D_Inter also uses the same set of sequence embeddings, MSA 1D features and 2D pairwise features as the input, and our model d shows a significant performance improvement over DRN-1D2D_Inter (single model) (the model trained on the same training and validation partition without cross validation) on both HomoPDB and HeteroPDB (the mean precision improvement: HomoPDB:14%, HeteroPDB:5.6%), indicating that the introduced graph representation is important for the model performance. The head-to-head comparison of model d and DRN-1D2D_Inter (single model) on each specific target in **Figure 4b** further demonstrates the value of

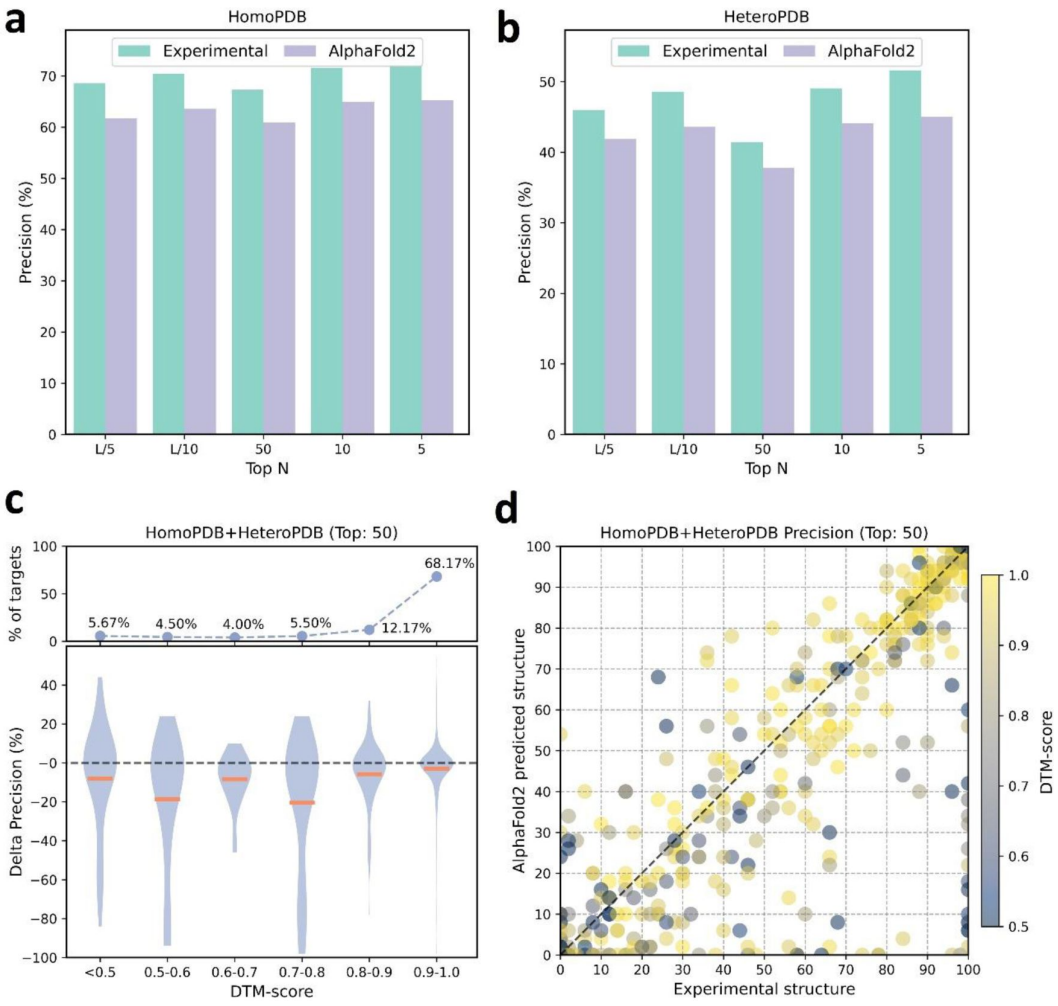


Figure 3.

The performances of PLMGraph-Inter when using experimental and AlphaFold2 predicted structures as the input. (a)~(b): The performance comparison of PLMGraph-Inter when using experimental structures and AlphaFold2 predicted structures as the input on (a) HomoPDB and (b) HeteroPDB. (c) The performance gaps (measured as the difference of the mean precision of the top 50 predicted contacts) of PLMGraph-Inter with the application of AlphaFold2 predicted structures and experimental structures as the input when the PPIs are within different intervals of DTM-score. The upper panel shows the percentage of the total number of PPI's in each interval. (d) The comparison of the precision of top 50 contacts predicted by PLMGraph-Inter for each target when using experimental structures and AlphaFold2 predicted structures as the input.

the graph representation. Besides, the additional introduced structure embeddings from ESM-IF can further improve the mean precisions of the predicted contacts by 3∼4% on both HomoPDB and HeteroPDB (model e versus model d) and the application of the cross validation can also improve the precisions by 1.0 % on HomoPDB and 2.7% on HeteroPDB (model f versus model d) (see **Table S3**).

To demonstrate the efficacy of our proposed graph representation of protein structures, we also trained a model using the structural representation proposed in the work of GVP(Jing, Eismann, Soni, et al., 2021; Jing, Eismann, Suriana, et al., 2021) (denoted as "GVP Graph"), as a control. Our structural representation differs significantly from GVP Graph. For example, we extracted inter-residue distances and orientations between five atoms (C,O,Cα,N and a virtual Cβ) from the structure as the geometric scalar and vector features, in which the vector features are calculated in a local coordinate system. However, GVP Graph only uses the distances and orientations between Cα atoms as the geometric scalar and vector features and the vector features are calculated in a global coordinate system. In addition, after the geometric graph is transformed by the graph encoder module, GVP Graph only uses the scalar features of each node as the node representation, while we concatenate the scalar and vector features of the node as the node representation. In **Figure 4c**, we show the performance comparison between this model and our base model (model a). From **Figure 4c**, we can clear see that our base model significantly outperforms the GVP Graph-based model on both HomoPDB and HeteroPDB, illustrating the high efficacy of our proposed graph representation.

We also explored the performance of PLMGraph-Inter on the HomoPDB and HeteroPDB test sets when using different protocols to further remove potential redundancies between the training and the test sets. Specifically, although the "40% sequence identity" used in our study is a widely used threshold to remove redundancy when evaluating deep learning-based protein-protein interaction and protein complex structure prediction methods(Evans et al., 2022b; Sledzieski et al., 2021), it is worth testing whether PLMGraph-Inter can keep its performance when more stringent threshold is applied. Besides, it is also worth evaluating whether PLMGraph-Inter can keep its performance on targets for which the folds of their interacting monomers are different from the targets in the training set (i.e. non-redundant in main chain structures of interacting monomers). To the best of our knowledge, all the previous studies failed to remove potential redundancies in folds of the interacting monomers when evaluating their methods.

In Table2, we show mean precisions of the contacts (top 50) predicted by PLMGraph-Inter on the HomoPDB and HeteroPDB when various sequence identity thresholds (with MMseq2(Steinegger & Söding, 2018)) and fold similarity thresholds (with TMalign(Zhang & Skolnick, 2005)) were further used in the de-redundancy (see "Further potential redundancies removal between the training and the test" in the Methods section**)**. It can be seen from that table that when using more stringent sequence identity thresholds for de-redundancy, the performance of PLMGraph-Inter on both the HomoPDB and HeteroPDB datasets decreases very little. For example, even when using "10% sequence identity" for de-redundancy, mean precisions of the predicted contacts only decreases by 2∼4%. Whereas when using fold similarities of the interacting monomers for de-redundancy, although the performance of PLMGraph-Inter on HeteroPDB decreases very little (only 3%∼4% when TM-Score 0.5 is used as the threshold), the performance of PLMGraph-Inter on HomoPDB decreases significantly (17%∼19% when TM-Score 0.5 is used as the threshold). One possible reason for the performance decrease on HomoPDB is that the binding mode of the homomeric PPI is largely determined by the fold of its monomer, thus the model does not generalize well on targets whose folds have never been seen during the training.

Evaluation of PLMGraph-Inter on DHTest and DB5.5 test sets

We further evaluated PLMGraph-Inter on DHTest and DB5.5. The DHTest test set was formed by removing PPIs redundant to our training set from the original test set of DeepHomo, which contains 130 homomeric PPIs. The DB5.5 test set was formed by removing PPIs redundant to our

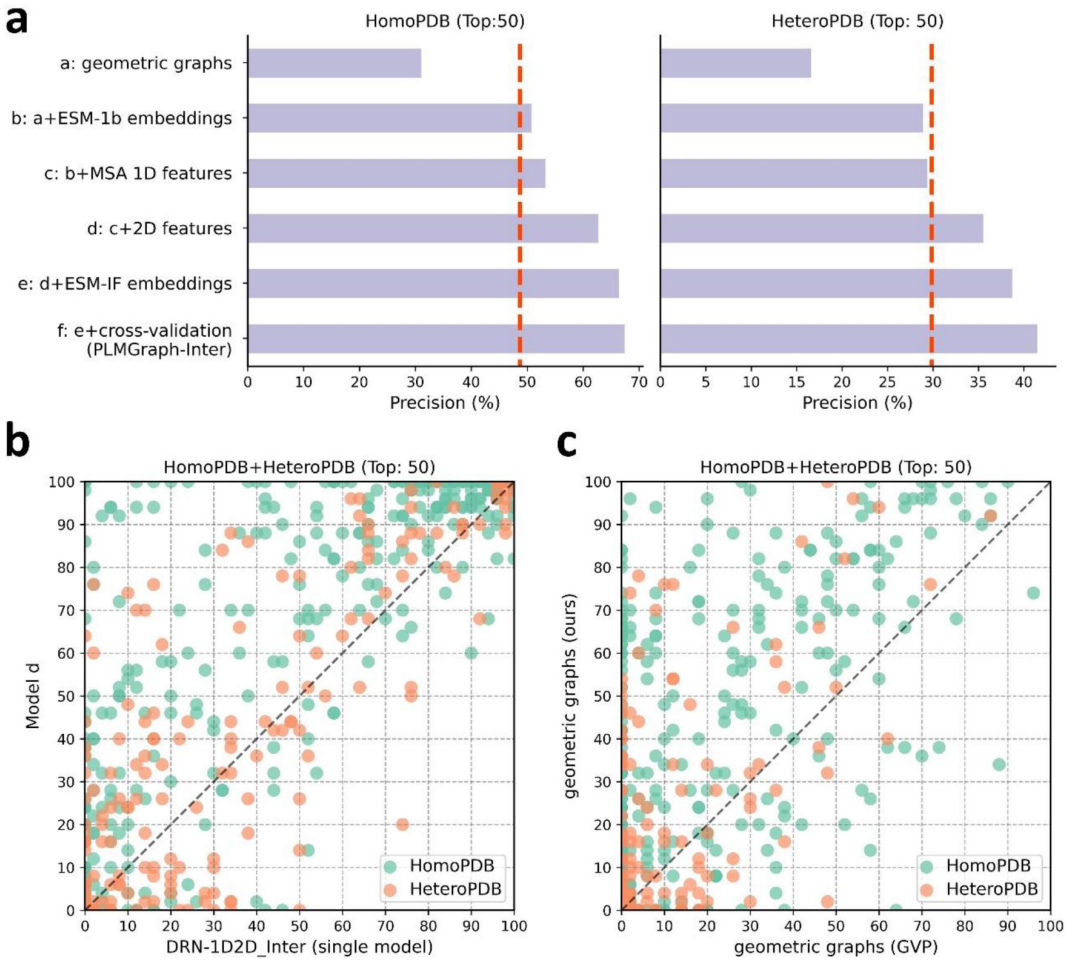


Figure 4.

The ablation study of PLMGraph-Inter on the HomoPDB and HeteroPDB test sets. (a) The mean precisions of the top 50 contacts predicted by different ablation models on the HomoPDB and HeteroPDB test sets. (b) The head-to-head comparisons of mean precisions of the top 50 contacts predicted by model d and DRN-1D2D_Inter (single model) for each target in HomoPDB and HeteroPDB. (c) The head-to-head comparison of mean precisions of the top 50 contacts predicted by the model using our geometric graphs and the GVP geometric graphs.

training dataset from the heterodimers in Protein-protein Docking Benchmark 5.5, which contains 59 heteromeric PPIs. Still, both the experimental structures and the predicted structures (generated using the same protocol as in HomoPDB and HeteroPDB) of the interacting monomers were used respectively in the inter-protein contact prediction. It should be noted that since DHTest and DB5.5 are not de-redundant with the training sets of CDPred and GLINTER, particularly, all PPIs in the DHTest test set are included in the training set of CDPred, thus the performances of the two methods may be overestimated.

As shown in **Figure 5a-b**, when using the experimental structures in the prediction, the mean precisions of the top 50 contacts predicted by PLMGraph-Inter are 71.9% on DHTest and 29.5% on DB5.5 (also see **Table S3**), which are dramatically higher than DeepHomo, GLINTER, DeepHomo2 and DRN-1D2D_Inter (Note: GLINTER encountered errors for 47 targets in DHTest and 3 targets in DB5.5 at run time and did not produce predictions, thus we removed these targets in the evaluation of the performance of GLINTER. The performances of other methods on DHTest and DB5.5 after the removal of these targets are shown in **Table S4**). We can also see that although PLMGraph-Inter achieved significantly better performance than CDPred on DB5.5, its performance on DHTest is quite close to CDPred. However, it should be noted that the performance of CDPred on DHTest might be grossly overestimated since PPIs in DHTest are fully included in the training set of CDPred. The distributions of the precisions of top 50 predicted contacts by different methods on DHTest and DB5.5 are shown in **Figure 5d**-**e**, from which we can clearly see that PLMGraph-Inter can make high-quality predictions for more targets on both DHTest and DB5.5.

When using the predicted structures in the prediction, the mean precisions of top 50 contacts predicted by PLMGraph-Inter show reasonable decrease to 61.1% on DHTest and 23.8% on DB5.5 respectively (see **Figure 5a-b** and **Table S3**). We also analyzed the impact of the monomeric structure quality on the inter-protein contact prediction performance of PLMGraph-Inter. As shown in **Figure 5c**, when the DTM-score is greater than or equal to 0.8, there is almost no difference between applying the predicted structures and the experimental structures, which is consistent with our analysis in HomoPDB and HeteroPDB.

We noticed that the performance of PLMGraph-Inter on the DB5.5 is significantly lower than that on HeteroPDB, and so are the performances of other methods. That the targets in DB5.5 have relatively lower mean contact densities (1.01% versus 1.29%) may partly explain this phenomenon. In **Figure 5f**, we show the variations of the precisions of predicted contacts with the variation of contact density. As can be seen from **Figure 5f**, as the contact density increases, precisions of predicted contacts tend to increase regardless of whether the experimental structures or predicted structures are used in the prediction. 37.29% targets in DB5.5 are with inter-protein contact densities lower than 0.5%, for which precisions of predicted contacts are generally very low, making the overall inter-protein contact prediction performance on DB5.5 relatively low.

Comparison of PLMGraph-Inter with AlphaFold-Multimer

After the development of AlphaFold2, DeepMind also released AlphaFold-Multimer, as an extension of AlphaFold2 for protein complex structure prediction. The inter-protein contacts can also be extracted from the complex structures generated by AlphaFold-Multimer. It is worth making a comparison between the performances of AlphaFold-Multimer and PLMGraph-Inter on inter-protein contact prediction. Therefore, we also employed AlphaFold-Multimer (version 2.2) with its default settings to generate complex structures for all the PPIs in the four datasets which we used to evaluate PLMGraph-Inter. We then selected the 50 inter-protein residue pairs with the shortest heavy atom distances in each generated protein complex structures as the predicted inter-protein contacts. It should be noted that AlphaFold-Multimer used all protein complex structures in Protein Data Bank deposited before 2018-04-30 in the model training, thus these PPIs may have a large overlap with the training set of AlphaFold-Multimer. Therefore, there is no doubt that the performance of AlphaFold-Multimer would be overestimated here. It should also be noted that

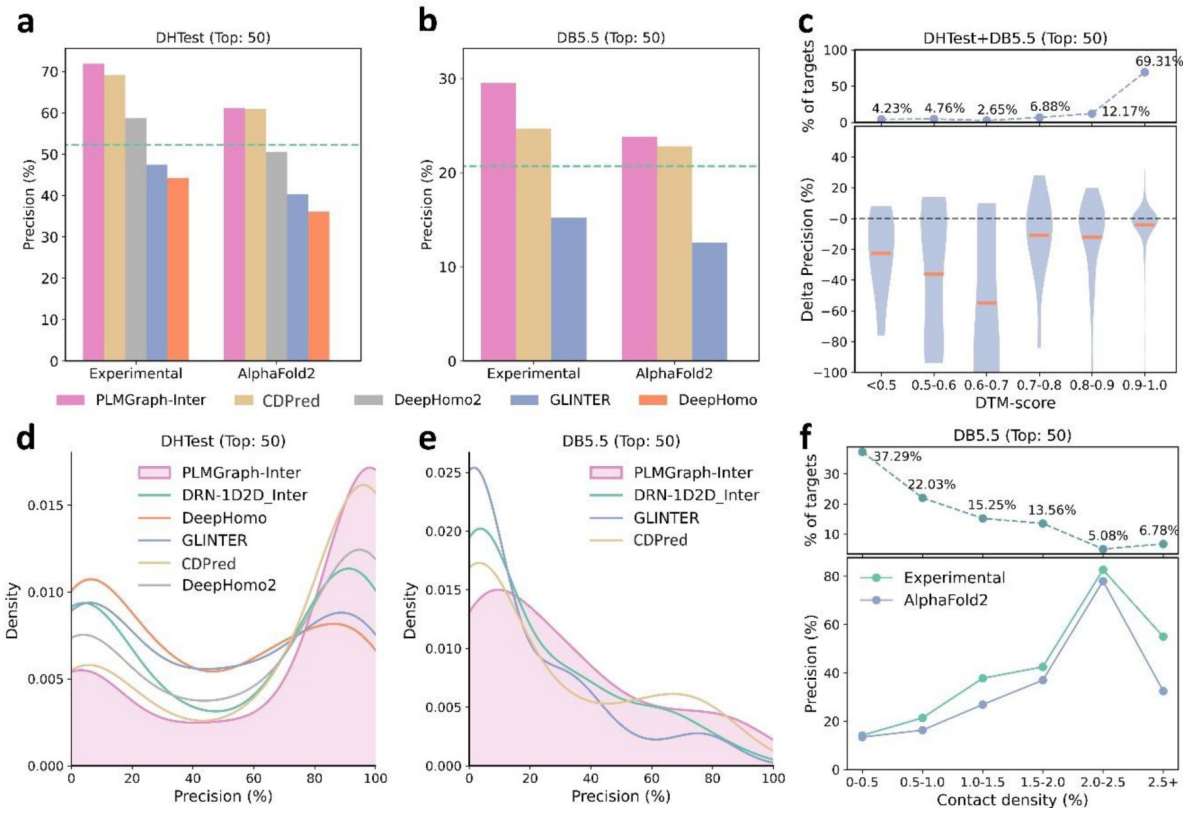


Figure 5.

The performances of PLMGraph-Inter and other methods on the DHTest and DB5.5 test sets. (a)~(b): The mean precisions of top 50 contacts predicted by PLMGraph-Inter, GLINTER, DeepHomo2, CDPred and DeepHomo on (a) DHTest and (b) DB5.5 when using experimental structures and AlphaFold2 predicted structures as the input, where the green lines indicate the performance of DRN-1D2D_Inter. (c) The performance gaps (measured as the difference of the mean precision of the top 50 predicted contacts) of PLMGraph-Inter with the application of AlphaFold2 predicted structures and experimental structures as the input when the PPIs are within different intervals of DTM-score. The upper panel shows the percentage of the total number of PPI's in each interval. (d)~(e): The distributions of precisions of the top 50 contacts predicted by PLMGraph-Inter and other methods for PPIs in (d) DHTest and (e) DB5.5. (f) The mean precisions of the top 50 contacts predicted by PLMGraph-Inter on PPIs within different intervals of contact densities in DB5.5. The upper panel shows the percentage of the total number of PPI's in each interval.

although AlphaFold-Multimer makes the prediction from sequences, it automatically searches templates of the interacting monomers. When we checked our AlphaFold-Multimer runs, we noticed for 99% of the targets (including all the targets in the four datasets: HomoPDB, HeteroPDB, DHTest and DB5.5), at least 20 templates were identified (AlphaFold-Multimer only employed the top 20 templates), and AlphaFold-Multimer employed the native template (i.e. the template which has the same PDB id with the target) for 87.8% of the targets. Besides, AlphaFold-Multimer employs multiple sequence databases including huge metagenomics database(Jumper et al., 2021), but PLMGraph-Inter only employs the UniRef100, thus the comparison is not on the same footing.

In **Figure 6a**, we show the relationship between the quality of the generated protein complex structure (evaluated with DockQ) and the precision of the top 50 inter-protein contacts extracted from the protein complex structure predicted by AlphaFold-Multimer for each PPI in the homomeric PPI (DHTest + HomoPDB) and heteromeric PPI (DB5.5 + HeteroPDB) datasets. As it can be seen from the figure that the precision of the predicted contacts is highly correlated with the quality of the generated structure. Especially when the precision of the contacts is higher than 50%, most of the generated complex structures have at least acceptable qualities (DockQ≥ 0.23), in contrast, almost all the generated complex structures are incorrect (DockQ<0.23) when the precision of the contacts is below 50%. Therefore, 50% can be considered as a critical precision threshold for inter-protein contact prediction (the top 50 contacts).

In **Figure 6b**, we show the comparison of the precisions of top 50 contacts predicted by AlphaFold-Multimer and PLMGraph-Inter for each target when using the experimental monomeric structures as the input for PLMGraph-Inter respectively (also see **Table S5**, and the comparison when using the AlphaFold2 predicted structures is shown in **Figure S5b**-**d**). It can be seen from the figure that although for most of the targets, AlphaFold-Multimer yielded better results, but for a significant number of the targets that AlphaFold-Multimer made poor predictions (precision<50%), the results of PLMGraph-Inter can have certain improvement over the AlphaFold-Multimer predictions.

We further explored the performance of PLMGraph-Inter on the PPIs which AlphaFold-Multimer failed to make correct predictions. Specifically, we denoted a PPI for which the precision of top 50 inter-protein contacts predicted by AlphaFold-Multimer is lower than 50% or the DockQ of protein complex structure predicted by AlphaFold-Multimer is less than 0.23 as “precision-Failed”, “DockQ-Failed”. The “iptm+ptm” metrics output by AlphaFold-Multimer for each target also has certain ability to characterize the quality of the predicted complexes. Our DockQ versus “iptm+ptm” analysis shows that 0.5 can be reasonably chosen as the cutoff of “iptm+ptm” to evaluate whether the prediction of AlphaFold-Multimer is successful or not (see **Figure S5a**), so we denoted a prediction for which the “iptm + ptm” of the prediction is lower than 0.5 as “pTM-Failed”. Then the mean precisions of top 50 contacts predicted by PLMGraph-Inter and AlphaFold-Multimer on the “precision-Failed”, “pTM-Failed” and “DockQ-Failed” sub-test sets from “DHTest+HomoPDB” and “DB5.5+HeteroPDB” are shown in **Figure 6c-d**. From **Figure 6c-d** we can see that the mean precisions of contacts predicted by PLMGraph-Inter are higher than the mean precisions of contacts predicted by AlphaFold-Multimer, further demonstrating that PLMGraph-Inter can complement AlphaFold-Multimer in certain cases.

PLMGraph-Inter can significantly improve protein-protein docking performance

Prior to AlphaFold-Multimer, protein-protein docking is generally used for protein complex structure prediction. HADDOCK(Honorato et al., 2021; van Zundert et al., 2016) is a widely used information-driven protein-protein docking approach to model complex structures of PPIs, which allows us to encode predicted inter-protein contacts as constraints to drive the docking. In this study, we used HADDOCK (version 2.4) to explore the contribution of PLMGraph-Inter to protein complex structure prediction.

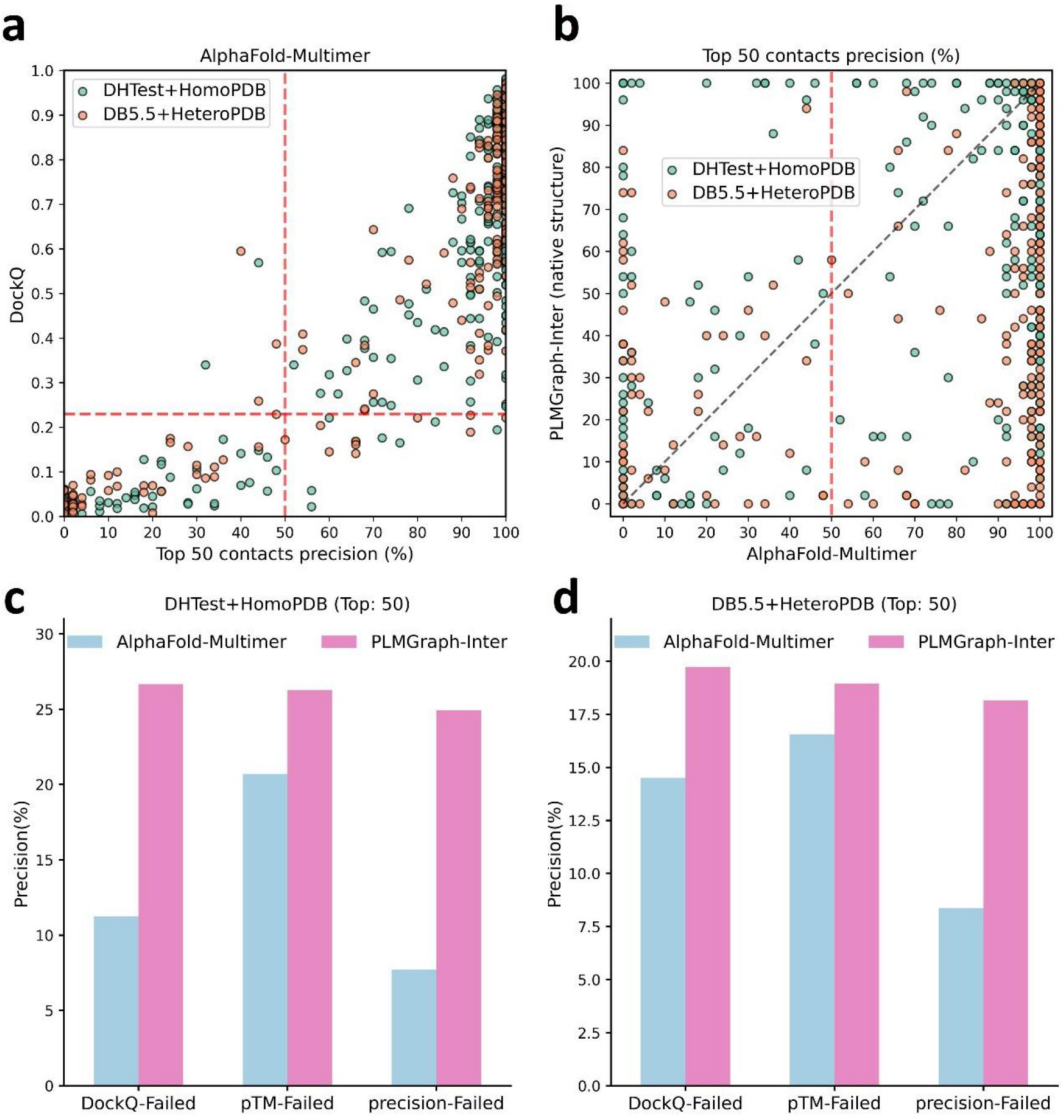


Figure 6.

The comparison of PLMGraph-Inter with AlphaFold-Multimer.

(a) The head-to-head comparison between the qualities of the protein complex structures generated by AlphaFold-Multimer (evaluated with DockQ) and the precision of the top 50 inter-protein contacts extracted from the generated protein complex structures. The red horizontal lines represent the threshold (DockQ=0.23) to determine whether the complex structure prediction is successful or not. (b) The head-to-head comparisons of precisions of the top 50 inter-protein contacts predicted by PLMGraph-Inter and AlphaFold-Multimer for each target in the homomeric PPI and heteromeric PPI datasets. (c)~(d): The mean precisions of top 50 inter-protein contacts predicted by PLMGraph-Inter and AlphaFold-Multimer on the PPI subsets from (c)"DHTest+HomoPDB" and (d) "DB5.5+HeteroPDB" in which the precision of the top 50 inter-protein contacts predicted by AlphaFold-Multimer is lower than 50% or the DockQ of the complex structure predicted by AlphaFold-Multimer is lower than 0.23 or the "iptm + ptm" of the complex structure predicted by AlphaFold-Multimer is lower than 0.5.

We prepared the test set of homomeric PPIs by merging HomoPDB and DHTest and the test set of heteromeric PPIs by merging HeteroPDB and DB5.5, where the monomeric structures generated previously by AlphaFold2 were used as the input to HADDOCK for protein-protein docking, in which the top 50 contacts predicted by PLMGraph-Inter with the application of the predicted monomeric structures were used as the constraints. Since HADDOCK generally cannot model large conformational changes in protein-protein docking, we filtered PPIs in which either of the AlphaFold2 generated interacting monomeric structure has a TM-score lower than 0.8. Finally, the homomeric PPI test set contains 462 targets, denoted as Homodimer, and the heteromeric PPI test set contains 174 targets, denoted as Heterodimer.

For each PPI, we used the top 50 contacts predicted by PLMGraph-Inter and other methods as ambiguous distance restraints between the alpha carbons (CAs) of residues (distance=8Å, lower bound correction=8Å, upper-bound correction=4Å) to drive the protein-protein docking. All other parameters of HADDOCK were set as the default parameters. In each protein-protein docking, HADDOCK output 200 predicted complex structures ranked by the HADDOCK scores. As a control, we also performed protein-protein docking with HADDOCK in ab initio docking mode (center of mass restraints and random ambiguous interaction restraints definition). Besides, for homomeric PPIs, we additionally added the C2 symmetry constraint in both cases.

As shown in **Figure 7a-c**, the success rate of docking on Homodimer and Heterodimer test sets can be significantly improved when using the PLMGraph-Inter predicted inter-protein contacts as restraints. Where on the Homodimer test set, the success rate (DockQ ≥ 0.23) of the top 1 (top 10) prediction of HADDOCK in ab initio docking mode are 15.37% (39.18%), and when predicted by HADDOCK with PLMGraph-Inter predicted contacts, the success rate of the top 1 (top 10) prediction 57.58% (61.04%). On the Heterodimer test set, the success rate of top 1 (top 10) predictions of HADDOCK in ab initio docking mode is only 1.72% (6.32%), and when predicted by HADDOCK with PLMGraph-Inter predicted contacts, the success rate of top 1 (top 10) prediction is 29.89% (37.93%). From **Figure 7a**-**7b** we can also see that integrating PLMGraph-Inter predicted contacts with HADDOCK not only allows for a higher success rate, but also more high-quality models in the docking results.

We further explored the relationship between the precision of top 50 contacts predicted by PLMGraph-Inter and the success rate of the top prediction of HADDOCK with PLMGraph-Inter predicted contacts. It can be clearly seen from **Figure 7d** that the success rate of protein-protein docking increases with the precision of contact prediction. Especially, when the precision of the predicted contacts reaches 50%, the docking success rate of both homologous and heterologous complexes can reach 80%, which is consistent with our finding in AlphaFold-Multimer. Therefore, we think this threshold can be used as a critical criterion for inter-protein contact prediction. It is important to emphasize that for some targets, although precisions of predicted contacts are very high, HADDOCK still failed to produce acceptable models. We manually checked these targets and found many of these targets have at least one chain totally entangled by another chain (e.g., PDB 3DFU in **Figure S6**). We think large structural rearrangements may exist in forming the complex structures, which is difficult to model by traditional protein-protein docking approach.

Finally, we compared the qualities of the complex structures predicted by HADDOCK (with PLMGraph-Inter predicted contacts) and AlphaFold-Multimer. Although for some targets (e.g., PDB 5HPS in **Figure S7a**), the qualities of the structures predicted by HADDOCK were higher than those by AlphaFold-Multimer, for most targets, AlphaFold-Multimer generated higher quality structures (**Figure S7**). Several reasons can account for the performance gap. First, precisions of the PLMGraph-Inter predicted contacts are still not enough, especially for heteromeric PPIs; second, HADDOCK cannot model large structural rearrangements in protein-protein docking, as we can see that for some targets, HADDOCK made poor predictions with high precisions of contact

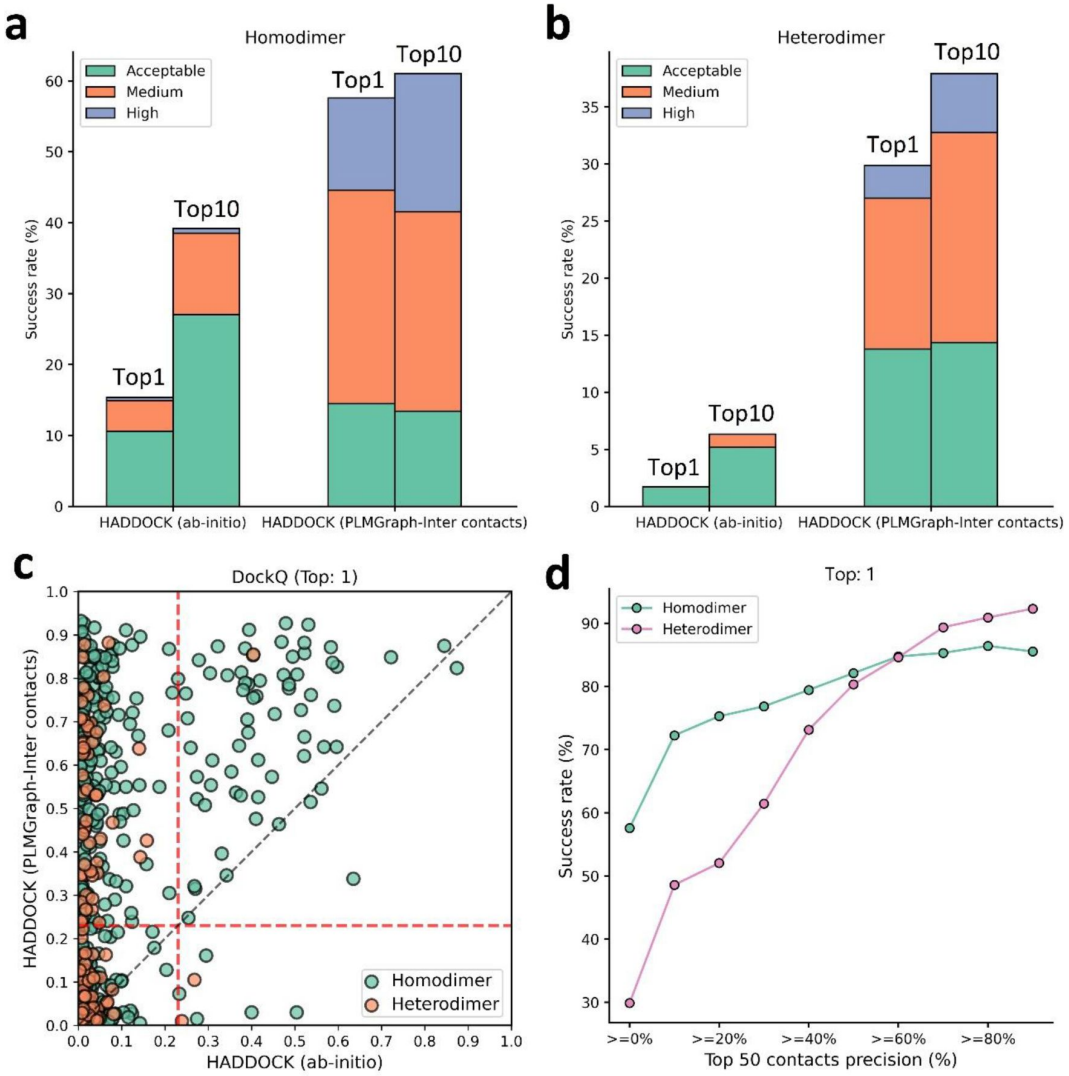


Figure 7.

Protein-protein docking performances on the Homodimer and Heterodimer test sets. (a)~(b) The protein-protein docking performance comparison between HADDOCK with and without (ab-initio) using PLMGraph-Inter predicted contacts as restraints on (a) Homodimer and (b) Heterodimer. The left side of each column shows the performance when the top 1 predicted model for each PPI is considered, and the right side shows the performance when the top 10 predicted models for each PPI are considered. (c) The head-to-head comparison of qualities of the top 1 model predicted by HADDOCK with and without using PLMGraph-Inter predicted contacts as restraints for each target PPI. The red lines represent the threshold (DockQ=0.23) to determine whether the complex structure prediction is successful or not. (d) The success rates (the top 1 model) for protein complex structure prediction when only including targets for which precisions of the predicted contacts are higher than certain thresholds.

constraints (**Figure 7d**); third, it is difficult to provide an objective evaluation of the true performance of AlphaFold-Multimer, since many targets have already been included in the training set of AlphaFold-Multimer.

Discussion

In this study, we proposed a new method to predict inter-protein contacts, denoted as PLMGraph-Inter. PLMGraph-Inter is based on the SE(3) invariant geometric graphs obtained from structures of interacting proteins which are embedded with multiple PLMs. The predicted inter-protein contacts are obtained by successively transforming the PLM embedded geometric graphs with graph encoders and residual networks. Benchmarking results on four test datasets show that PLMGraph-Inter outperforms five state-of-the-art inter-protein contact prediction methods including GLINTER, DeepHomo, CDPred, DeepHomo2 and DRN-1D2D_Inter by large margins, regardless of whether the experimental or predicted monomeric structures are used in building the geometric graphs. The ablation study further shows that the integration of the PLMs with the protein geometric graphs can dramatically improve the model performance, illustrating the efficacy of the PLM embedded geometric graphs in protein representations. The protein representation framework proposed in this work can also be used to develop models for other tasks like protein function prediction, PPI prediction, etc. Very recently, Fang et al. have also shown in their work that the incorporation of protein language models in geometric networks can significantly improve the model performances on a variety of protein-related tasks including protein-protein interface prediction, model quality assessment, protein-protein rigid docking and binding affinity prediction (F. Wu et al., 2023), which further supports this claim. We further show PLMGraph-Inter can complement the result of AlphaFold-Multimer and leveraging the inter-protein contacts predicted by PLMGraph-Inter as constraints in protein-protein docking implemented with HADDOCK can dramatically improve its performance for protein complex structure prediction.

We noticed that although PLMGraph-Inter has achieved remarkable progress in inter-protein contact prediction, there is still room for further improvement, especially for heteromeric PPIs. Using more advanced PLMs, larger training datasets and explicitly integrating physicochemical features of interacting proteins are directions worthy of exploration. Besides, since protein-protein docking approach generally have difficulties in modelling large conformational changes in PPIs, developing new approaches to integrate the predicted inter-protein contacts in the more advanced folding-and-docking framework like AlphaFold-Multimer or directly incorporating an additional structural module for protein complex structure generation in the network architecture can also be the future research directions.

Methods

Training and test datasets

We used the training set and test sets prepared in our previous work DRN-1D2D_Inter(Si & Yan, 2023) to train and evaluate PLMGraph-Inter. More details for the dataset generation can be found in the previous work. Specifically, we first prepared a non-redundant PPI dataset containing 4828 homomeric PPIs and 3134 heteromeric PPIs (with sequence identity 40% as the threshold), and after randomly selecting 400 homomeric PPIs (denoted as HomoPDB) and 200 heteromeric PPIs (denoted as HeteroPDB) as independent test sets, the remaining 7362 homomeric and heteromeric PPIs were used for training and validation.

DHTest and DB5.5 were also prepared in the work of DRN-1D2D_Inter by removing PPIs which are redundant (with sequence identity 40% as the threshold) to our training and validation set from the test set of DeepHomo and Docking Benchmark 5.5. DHTest contains 130 homomeric PPIs, and DB5.5 contains 59 heteromeric PPIs. Therefore, all the test sets used in this study are non-redundant (with sequence identity 40% as the threshold) to the dataset for the model development.

Inter-protein contact definition

For a given PPI, two residues from the two interacting proteins are defined to be in contact if the distance of any two heavy atoms belonging the two residues is smaller than 8 Å.

Preparing the Input features

Geometric graphs from structures of interacting monomers

We first represent the protein as a graph, where each residue is represented as a node, and an edge is defined if the Cα atom distance between two residues is less than 18Å (In our small-scale tests, increasing the cutoff used for defining edges can slightly increase the performance of the model. However, due to GPU memory limitations, we set the cutoff as 18Å). For each node and edge, we use scalars and vectors extracted from the 3D structures as their geometric features.

For each residue, we use its C,O,Cα,N and a virtual Cβ atom coordinates to extract information, the virtual Cβ coordinates are calculated using the following formula(Dauparas et al., 2022): b = Cα - N, c = C -Cα, a = cross(b, c), Cβ = -0.58273431*a + 0.56802827*b - 0.54067466*c + Cα.

To achieve a SE(3) invariant graph representation, as shown in **Figure S3b**, we define a local coordinate system on each residue(Jumper et al., 2021; Pagès et al., 2019). Specifically, for each residue, the unit vector in the Cα − C direction is set as the $\vec{x}$ axis, the unit vector in the Cα-C-N plane and perpendicular $\vec{x}$ to is used as $\vec{y}$, and the z-direction is obtained through the cross product of $\vec{x}$ and $\vec{y}$.

For the ith node, we use the three dihedral angles (φ, ψ, ω) of the corresponding residue as the scalar features of the node (**Figure S8a**), and the unit vectors between the C_i, N_i,O_i, $C\alpha_i$, and $C\beta_i$ atoms of the corresponding residue and the C_{i-1}, N_{i-1},O_{i-1}, $C\alpha_{i-1}$, and $C\beta_{i-1}$ atoms of the forward residue and the C_{i+1}, N_{i+1}, O_{i+1}, $C\alpha_{i+1}$, and $C\beta_{i+1}$ atoms of backward residue as the vector features of the node. In total, for each node, the dimension of the scalar features is 6 (each dihedral angle is encoded with its sine and cosine) and the dimension of the vector features is 50*3.

For the edge between ith node and jth node, we use the distances and directions between the atoms of the two residues as the scalar features and vector features(See **Figure S8b**). The distances between the C_i, N_i,O_i, $C\alpha_i$, and $C\beta_i$ atoms of ith residue and the C_j, N_j,O_j, $C\alpha_j$, and $C\beta_j$ atoms of the jth residue are used as scalar features after encoded with the 16 Gaussian radial basis functions(Jing, Eismann, Suriana, et al., 2021). The position difference between i and j (j-i) is also used as a scalar feature after sinusoidal encoding(Vaswani et al., 2017). The unit vectors between the C_i, N_i,O_i, $C\alpha_i$, and $C\beta_i$ atoms of ith residue and the C_j, N_j,O_j, $C\alpha_j$, and $C\beta_j$ atoms of the jth residue are used as vector features. In total, for each edge, the dimension of the scalar features is 432 and the dimension of the vector features is 25*3.

Embeddings of single sequence, MSA and structure

The single sequence embedding is obtained by feeding the sequence into ESM-1b, and the structure embedding is obtained by feeding the structure into ESM-IF. To obtain the MSA embedding, we first search the Uniref100 protein sequence database for the sequence using

JACKHMMER(Potter et al., 2018) with the parameter (--incT L/2) to obtain the MSA, which is then inputted to hhmake(Steinegger et al., 2019) to get the HMM file, and to the LoadHMM.py script from RaptorX_Contact(Wang et al., 2017) to obtain the PSSM. The number of sequences of MSA is limited to 256 by hhfilter(Steinegger et al., 2019) and then input to ESM-MSA-1b to get the MSA embedding. The dimensions of the sequence embedding, PSSM, MSA embedding and structural embeddings are 1280, 20, 768 and 512 respectively. After adding embeddings to the scalar features of the nodes, the dimension of the scalar features of each node is 2586.

2D feature from paired MSA

For homomeric PPIs, the paired MSA is formed by concatenating two copies of the MSA. For heteromeric PPIs, the paired MSA is formed by pairing the MSAs through the phylogeny-based approach described in (*https://github.com/ChengfeiYan/PPI_MSA-taxonomy_rank*)(Si & Yan, 2022). We input the paired MSA into CCMpred(Seemayer et al., 2014) to get the evolutionary coupling matrix, and into alnstats(Jones et al., 2015) to get mutual information matrix, APC-corrected mutual information matrix and contact potential matrix. The number of sequences of paired MSA is limited to 256 by hhfilter(Steinegger et al., 2019) and then input to ESM-MSA-1b to get the attention maps. In total, the channel of 2D features is 148.

GVP and GVPConv

GVP is a two-track neural network module consisting of a scalar track and a vector track, which can perform SE(3) invariant transformations on scalar features and SE(3) equivariant transformations on vector features. A detailed description can be found in the work of GVP(Jing, Eismann, Soni, et al., 2021; Jing, Eismann, Suriana, et al., 2021).

GVPConv is a message passing based graph neural network, which mainly consists of a message function and a feedforward function. Where the message function contains a sequence of three GVP modules and the feedforward function contains a sequence of two GVP modules. GVPConv is used to transform the node features. Specifically, the input node features are first processed by the message function. We denote the features of node i by $\boldsymbol{h}^i$, the feature of edge (j→i) by $\boldsymbol{h}^{j\to i}$, the set of nodes connected to node i by ε_i, and the three GVP modules of the message function by gm, then the node features processed by the message function can be represented as:

$$\boldsymbol{h}_m^i = \frac{1}{len(\boldsymbol{\varepsilon}_i)}\left(\sum_{j\in\boldsymbol{\varepsilon}_i} gm(concat(\boldsymbol{h}^i,\ \boldsymbol{h}^{j\to i}, \boldsymbol{h}^j))\right) \quad (1)$$

Where $len(\varepsilon_i)$ denotes the number of nodes connected to node i. After sequential normalization (Equation 2) and feedforward function (Equation 3), the features of node i updated by GVPConv Layer are obtained:

$$\boldsymbol{h}_m^i = \mathrm{LayerNorm}(\boldsymbol{h}_m^i\ +\ \mathrm{Dropout}(\boldsymbol{h}_m^i)) \quad (2)$$

$$\boldsymbol{h}_{out}^i = \mathrm{LayerNorm}(\boldsymbol{h}_m^i\ +\ \mathrm{Dropout}(\mathrm{gs}(\boldsymbol{h}_m^i))) \quad (3)$$

Where gs denotes the two GVP modules of the feedforward function, $\boldsymbol{h}^i$ denotes the outputs of GVPConv Layer.

The transforming procedure in the residual network module

We first use a convolution layer with kernel size of 1*1 to reduce the number of channels of the input 2D feature maps from 1044 to 96, which are then transformed successively by 9 dimensional hybrid residual blocks and another convolution layer with kernel size of 1*1 for the channel reduction (from 96 to 1). Finally, we use the sigmoid function to transform the feature map to obtain the predicted inter-protein contact map.

Training protocol

Our training set contains 7362 PPIs, and we used seven-fold cross-validation to train PLMGraph-Inter. Specifically, we randomly divided the training set into seven subsets, and each time, we selected six subsets as the training set and the remaining subset as the validation set. Seven models were trained in total, and the final prediction was the average of the predictions from the seven models. Each model was trained using AdamW optimizer with 0.001 as the initial learning rate, in which the singularity enhanced loss function proposed by in our previous study(Si & Yan, 2021) was used calculate the training and validation loss. During training, if the validation loss did not decrease within 2 epochs, we would decay the learning rate by 0.1. The training stopped after the learning rate decayed twice and the model with the highest top-50 mean precision on the validation dataset was saved as the prediction model.

PLMGraph-Inter was implemented with pytorch (v.1.11) and trained on one NVIDIA TESLA A100 GPU with batch size equaling to 1. Due to memory limitation of GPU, the length of each protein sequence was limited to 400. When a sequence was longer than 400, a fragment with sequence length equaling to 400 was randomly selected in the model training.

Quality assessment of the predicted protein complex structures

We evaluated the models generated by AlphaFold-Multimer and HADDOCK using DockQ(Basu & Wallner, 2016), a score ranging between 0 and 1. Specifically, a model with DockQ<0.23 means that the prediction is incorrect; 0.23≤DockQ<0.49 means the model is an acceptable prediction; 0.49≤DockQ<0.8 corresponds to a medium quality prediction; and 0.8≤ DockQ corresponds to a high quality prediction.

Further potential redundancies removal between the training and the test

Removing potential redundancies using different sequence similarity thresholds

CD-HIT(W. Li et al., 2001) was originally used in removing redundancies between the training and test sets used in this study. Since the lowest sequence identity threshold accepted by CD-HIT is 40%, to use more stringent threshold in the redundancy removal. We further clustered all the monomer sequences from the training set and the test sets (HomoPDB, HeteroPDB) using MMSeq2(Steinegger & Söding, 2018) with different sequence identity thresholds (40%, 30%, 20%, 10%). Under a certain threshold, each sequence is uniquely labeled by the cluster (e.g. cluster 0, cluster 1, ...) to which it belongs, from which each PPI can be marked with a pair of clusters (e.g. cluster 0-cluster 1). The PPIs belonging to the same cluster pair (note: cluster n - cluster m and cluster n-cluster m were considered as the same pair) were considered as redundant with this sequence identity threshold. For each PPI in the test set, if the pair cluster it belonging to contains any PPI belonging to the training set, we remove that PPI from the test set.

Removing potential redundancies using different fold similarity thresholds of interacting monomers

We used TM-align(Zhang & Skolnick, 2005) to evaluate the fold similarities (in TM-scores) between the experimental structures of the interacting monomers in the training set and the test sets (HomoPDB, HeteroPDB). Specifically, for any two targets A-B and A'-B' in the training set and test sets respectively, where A, B, A' and B' represent the interacting monomers. We calculated the MTM-score defined as

$$\mathrm{MTM}-\mathrm{score}=\max\begin{cases}min(TM-score_{AA'},TM-score_{BB'})\\ min(TM-score_{AB'},\ TM-score_{BA'})\end{cases} \quad (4)$$

between the two targets. The MTM-score higher than a certain value means that both the two interacting monomers in the two targets have fold similarity scores (TM-scores) higher than this value. When a threshold is chosen, we remove targets in the test tests if they have MTM-scores higher than this threshold when comparing with any target in the training set. In this study, different thresholds including 0.9, 0.8, 0.7, 0.6, 0.5 were used in the study. 0.5 was chosen as the lowest threshold for protein pairs with TM-score<0.5 are mainly not in the same fold.

Calculating the normalized number of the effective sequences of paired MSA

We define the normalized number of the effective sequences (N_{eff}^{norm}) as follows:

$$N_{eff}^{norm} = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{\sum_{m=1}^{N} I[S_{m,n} \geq 0.8]} \quad (5)$$

Where L is the length of the paired MSA, N is the number of sequences in the paired MSA, $S_{m,n}$ is the sequence identity between the m-th and n-th sequences, and I[] represents the Iverson bracket, which means $I[S_{m,n} \geq 0.8] = 1$ if $S_{m,n} \geq 0.8$ or 0 otherwise.

Data Availability

The PDB accession codes for the all the training and test sets are provided in *https://github.com/ChengfeiYan/PLMGraph-Inter/tree/main/data*. Other data for supporting the finds of this study are available from the corresponding author upon request.

Code Availability

The code for implementing PLMGraph-Inter is provided in *https://github.com/ChengfeiYan/PLMGraph-Inter*.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (32101001) and new faculty startup grant (3004012167) of Huazhong University of Science and Technology. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

Author Contributions

Y.S. and C.Y. designed and performed the experiments. Y.S. and C.Y. wrote the manuscript. C.Y. supervised the work.

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary

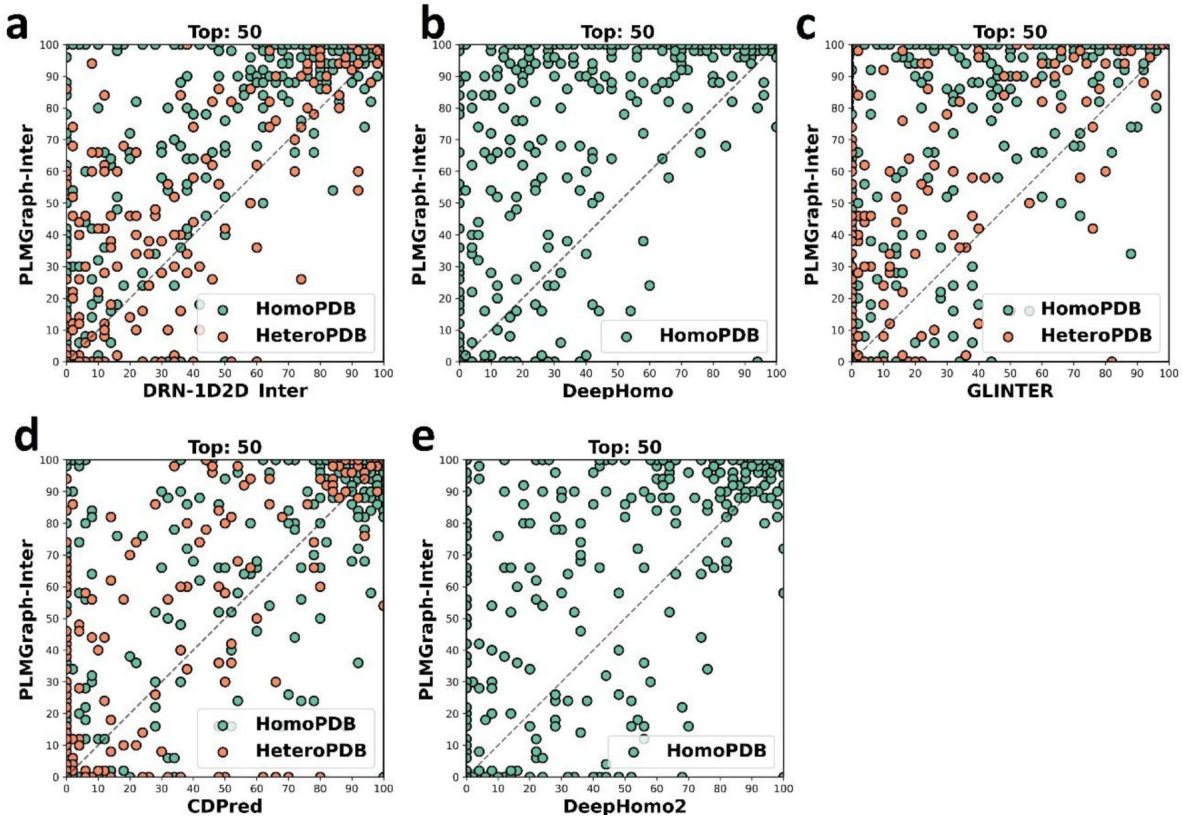


Figure S1.

The head-to-head comparison of the precisions (%) of the top 50 contacts predicted by PLMGraph-Inter and other methods for each target in HomoPDB and HeteroPDB using experimental structures.

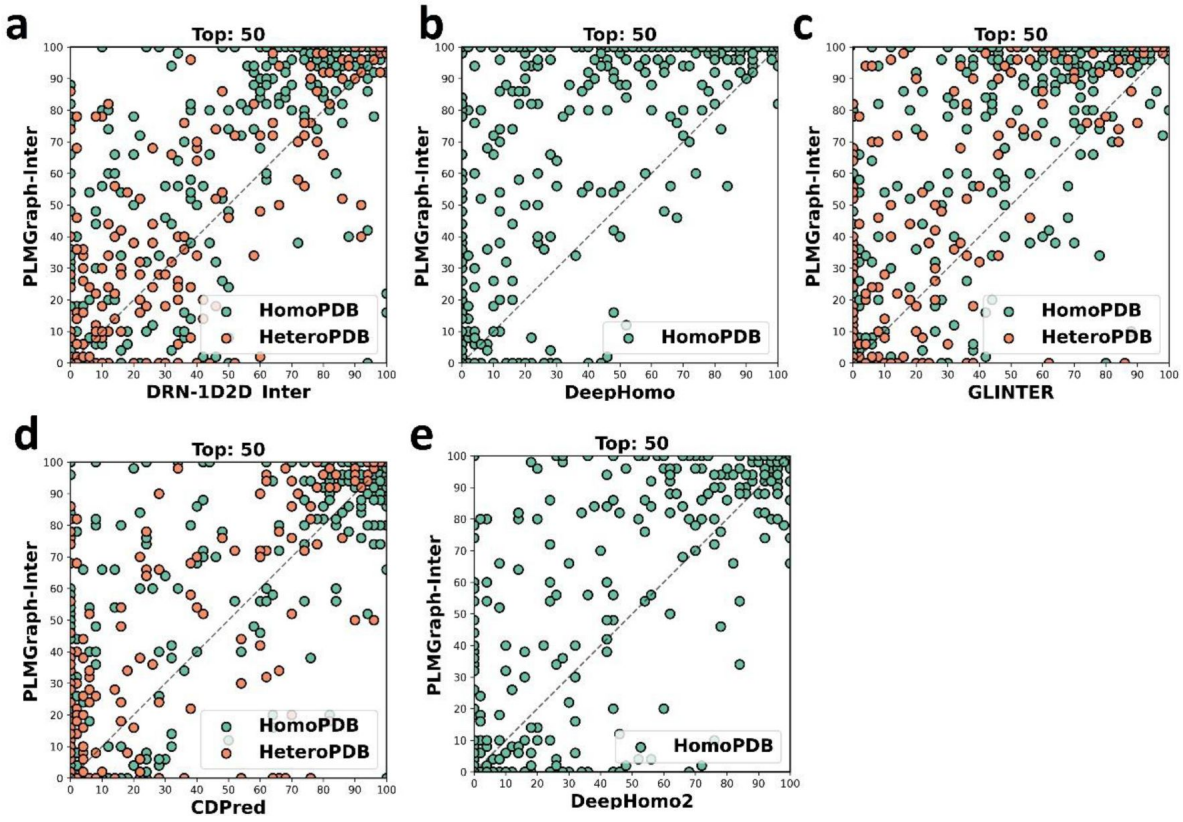


Figure S2.

The head-to-head comparison of the precisions (%) of the top 50 contacts predicted by PLMGraph-Inter and other methods for each target in HomoPDB and HeteroPDB using AlphaFold2 predicted structures.

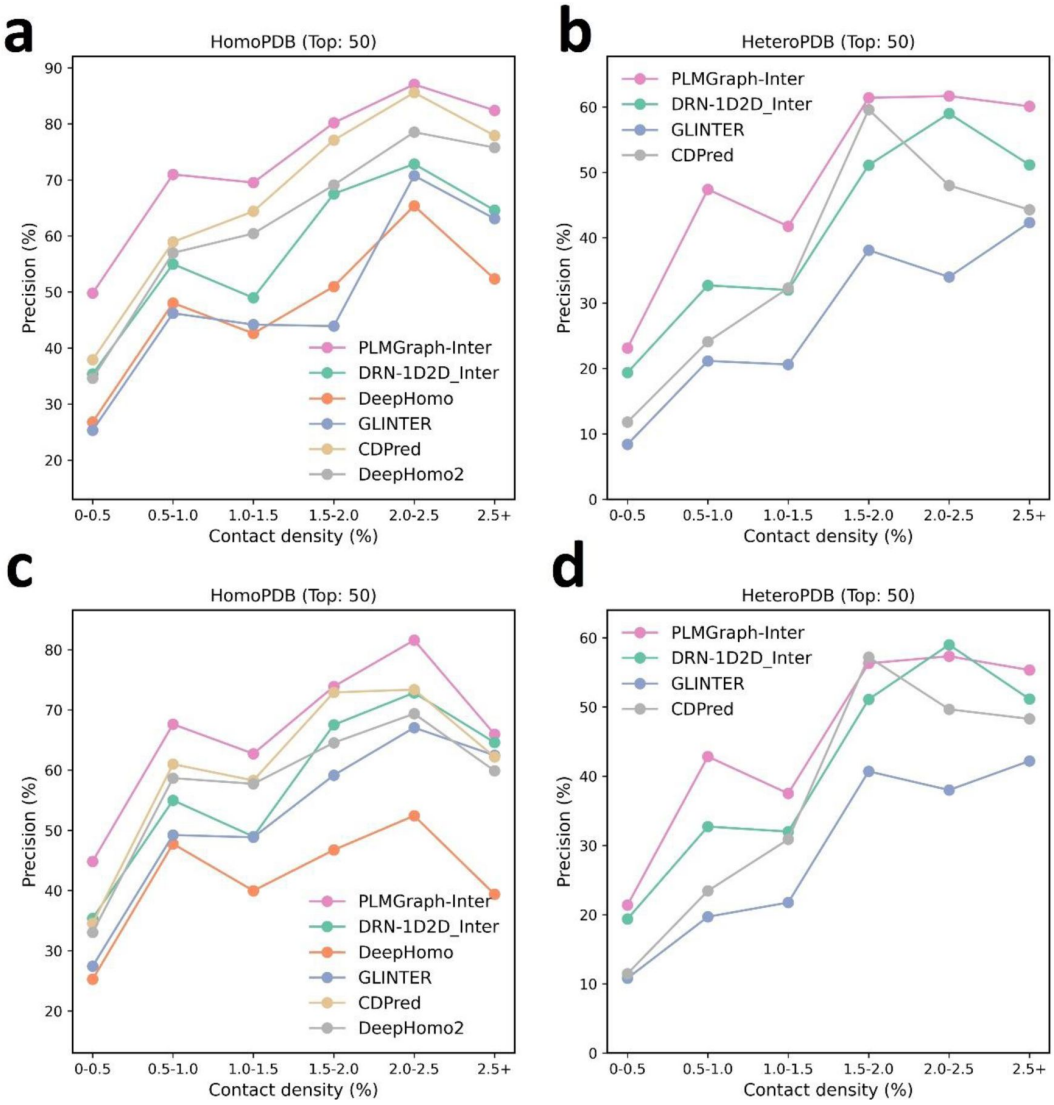


Figure S3.

The mean precision versus contact density for the top 50 contacts predicted by PLMGraph-Inter, GLINTER, DeepHomo, DeepHomo2, CDPred, DRN-1D2D_Inter on the HomoPDB test set (a, c) and HeteroPDB test set (b, d) using experimental structures (first row) and AlphaFold2 predicted structures (second row).

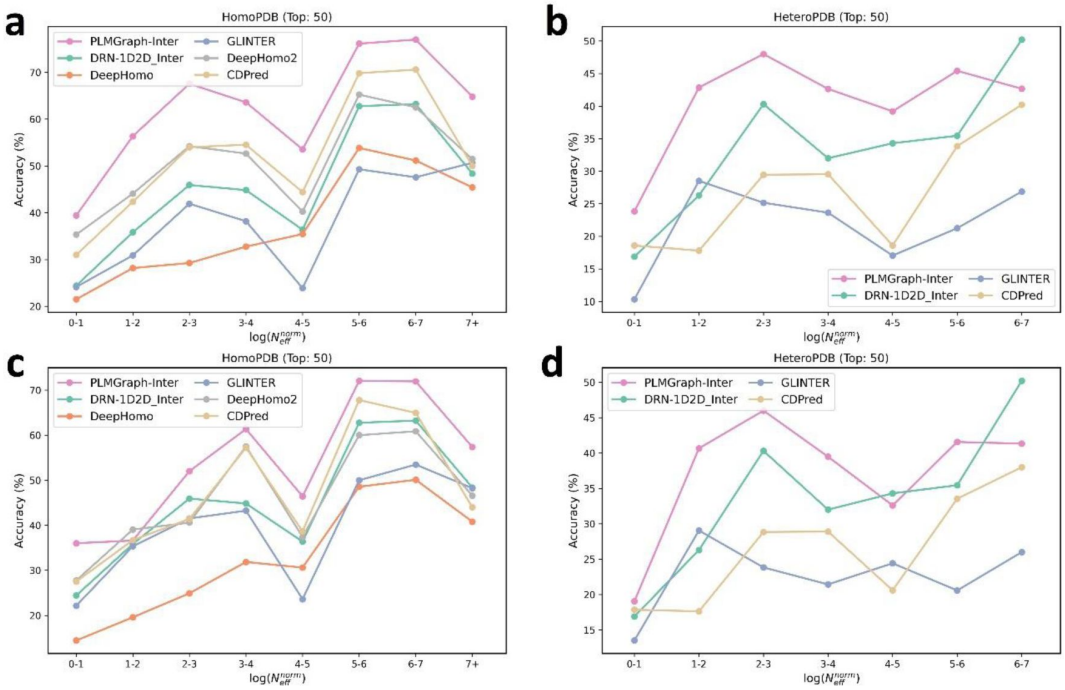


Figure S4.

The mean precision versus log(N_{eff}^{norm}) for the top 50 contacts predicted by PLMGraph-Inter, GLINTER, DeepHomo, DeepHomo2, CDPred, DRN-1D2D_Inter on the HomoPDB test set (a, c) and HeteroPDB test set (b, d) using experimental structures (first row) and AlphaFold2 predicted structures (second row).

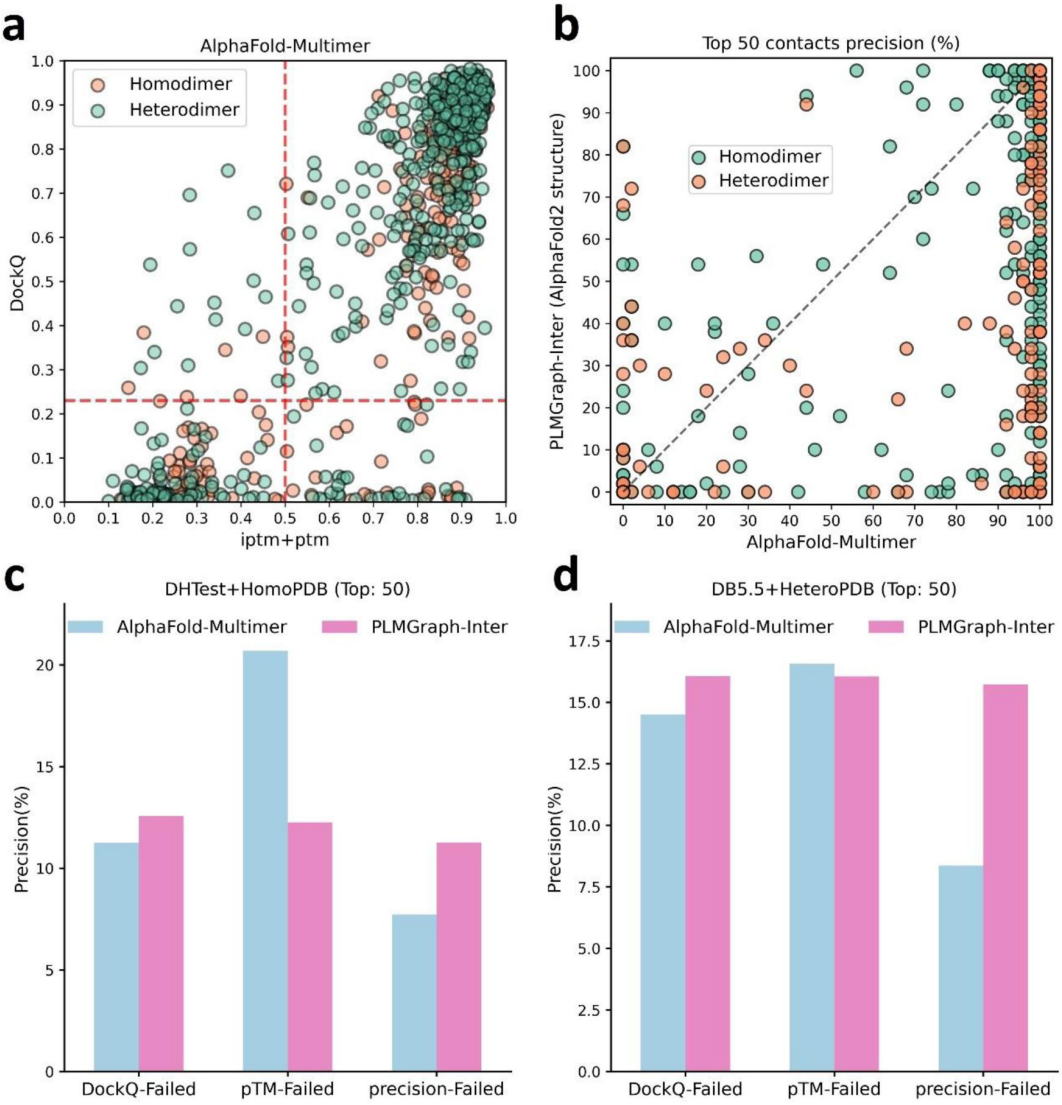


Figure S5.

The comparison of PLMGraph-Inter with AlphaFold-Multimer.

(b) The head-to-head comparisons of precisions of the top 50 inter-protein contacts predicted by PLMGraph-Inter(using AlphaFold2 predicted structures) and AlphaFold-Multimer for each target in the homomeric PPI and heteromeric PPI datasets. (c)~(d): The mean precisions of top 50 inter-protein contacts predicted by PLMGraph-Inter(using AlphaFold2 predicted structures as input) and AlphaFold-Multimer on the PPI subsets from (c) "DHTest+HomoPDB" and (d) "DB5.5+HeteroPDB" in which the precision of the top 50 inter-protein contacts predicted by AlphaFold-Multimer is lower than 50% or the DockQ of the complex structure predicted by AlphaFold-Multimer is lower than 0.23 or the "iptm + ptm" of the complex structure predicted by AlphaFold-Multimer is lower than 0.5.

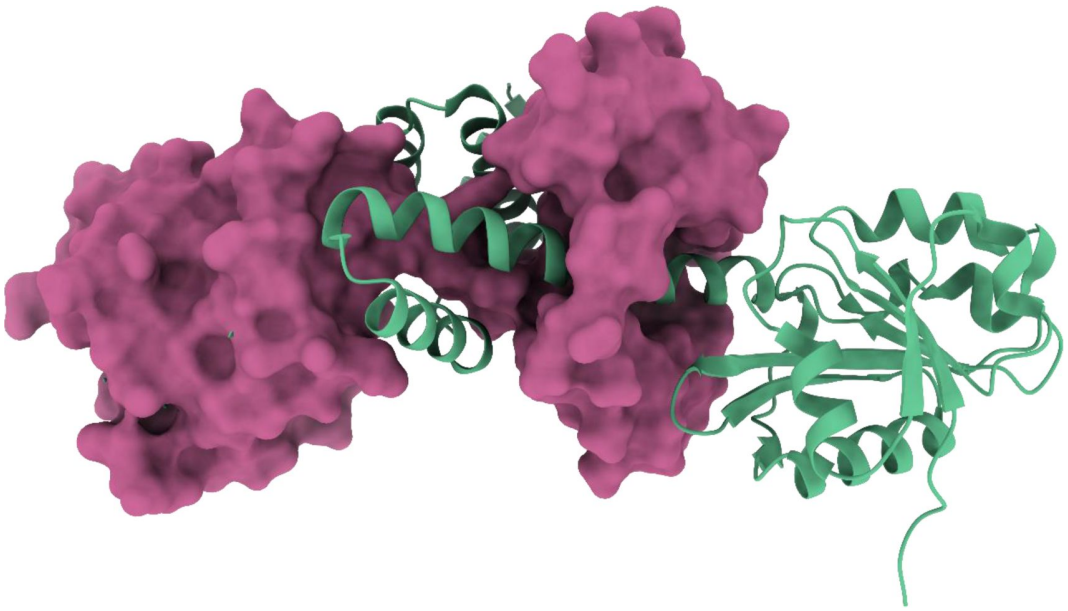


Figure S6.

3D structure of the homodimer (PDB: 3DFU).

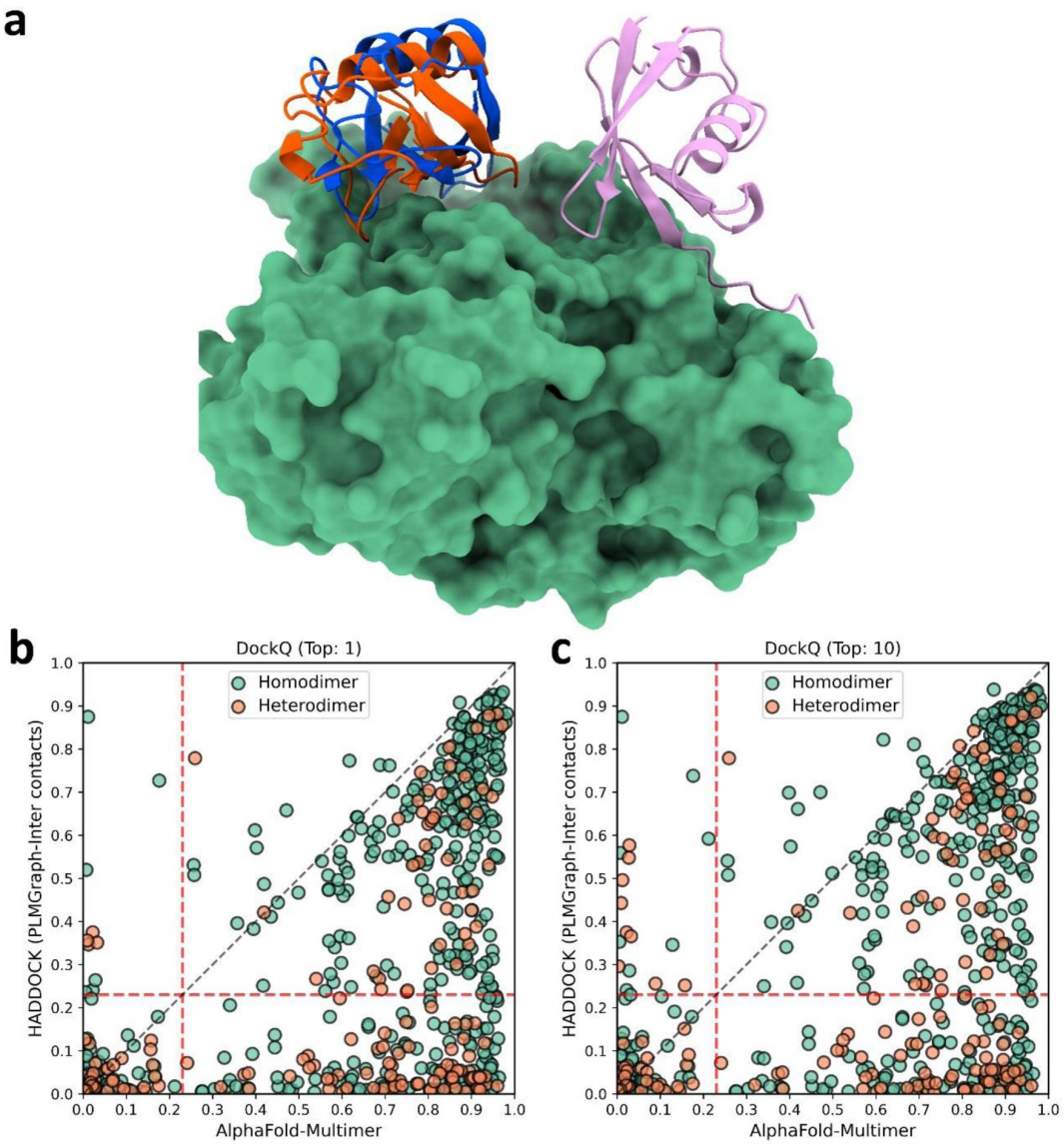


Figure S7.

The comparison of HADDOCK (with PLMGraph-Inter contact constraints) with AlphaFold-Multimer in protein complex structure prediction. (a) The binding configurations predicted by HADDOCK (colored orange, DockQ 0.375), predicted bys AlphaFold-Multimer (colored pink, DockQ 0) and the native binding configuration (colored blue) for the chain B of the protein complex structure in PDB 5HPS. The chain A is shown in the protein surface mode (colored green). (b)~(c): The head-to-head comparison of qualities of the (a) top 1 or (b) top 10 model predicted by HADDOCK with using PLMGraph-Inter predicted contacts as restraints and AlphaFold-Multimer for each target PPI.

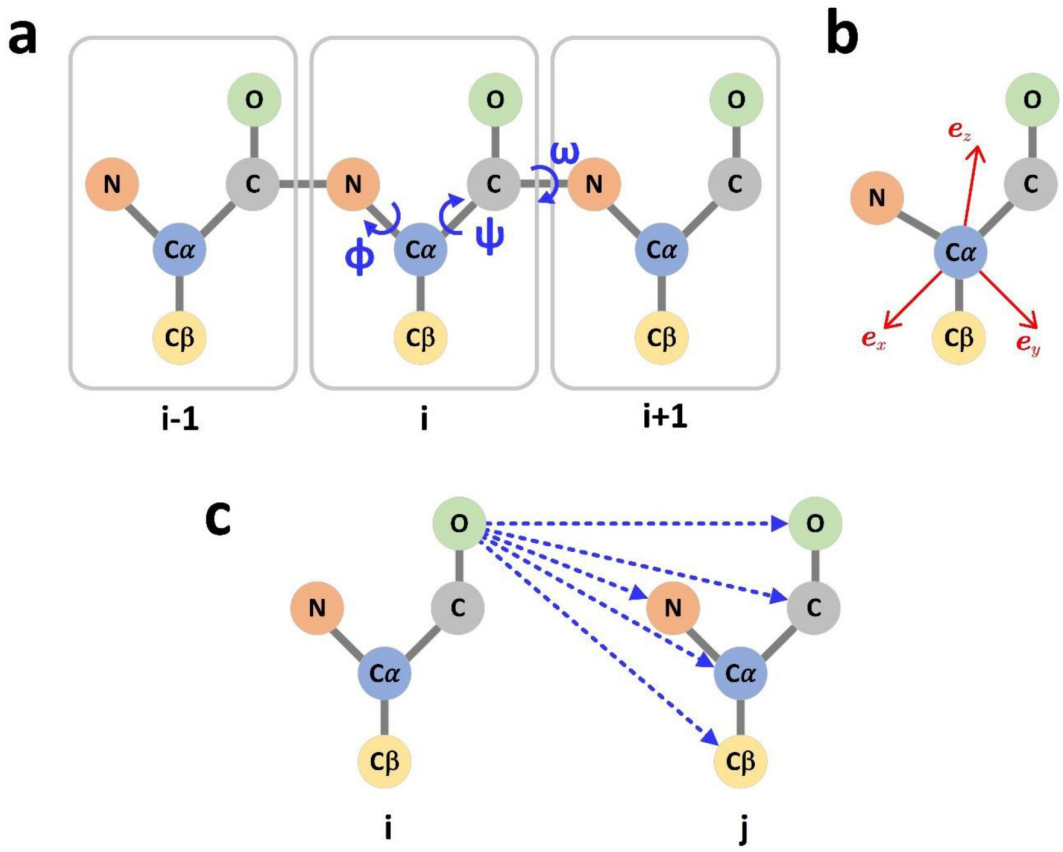


Figure S8.

The graph representation of protein structures. (a) Dihedral angles of the protein backbone. (b) The local coordinate system of each amino acid. (c) The scalar (distances) and vector (directions) of the edge i->j.

Methods	HomoPDB (precision %)					HeteroPDB (precision %)				
	L/5	L/10	50	10	5	L/5	L/10	50	10	5
DeepHomo	40.8 (36.8)	44.0 (39.5)	40.2 (36.1)	45.7 (41.6)	46.6 (42.6)					
GLINTER	42.6 (45.5)	44.7 (48.2)	41.8 (44.2)	46.1 (50.1)	48.2 (51.5)	24.3 (24.6)	25.2 (26.6)	21.3 (21.8)	25.9 (25.3)	27.1 (25.8)
DRN-1D2D_Inter	50.2	52.8	49.1	53.9	54.8	33.9	36.0	31.7	36.5	37.3
DeepHomo2	53.5 (49.8)	56.0 (51.3)	52.9 (49.2)	57.3 (53.4)	59.2 (54.1)					
CDPred	56.3 (51.5)	58.4 (52.9)	55.4 (50.9)	59.7 (53.9)	60.3 (54.2)	30.6 (30.7)	31.5 (32.3)	28.1 (27.7)	32.6 (32.8)	32.4 (33.2)
PLMGraph-Inter	**66.3 (59.1)**	**68.4 (61.0)**	**65.2 (58.4)**	**69.7 (62.4)**	**70.1 (62.5)**	**45.8 (41.9)**	**48.7 (43.5)**	**41.2 (37.7)**	**49.2 (44.3)**	**52.0 (45.2)**

Note: The highest mean precision (%) in each column is highlighted in bold.

Table S1.

The performances of DeepHomo, GLINTER, DRN-1D2D_Inter, DeepHomo2, CDPred and PLMGraph-Inter on HomoPDB and HeteroPDB after the removal of targets which GLINTER failed to make the prediction using experimental structures (AlphaFold2 predicted structures)

Methods	HomoPDB (precision %)					HeteroPDB (precision %)				
	L/5	L/10	50	10	5	L/5	L/10	50	10	5
Model a	31.7	34.2	30.9	35.8	37.4	18.2	19.1	16.5	19.3	19.7
Model b	52.1	54.9	50.7	57.3	58.3	32.7	32.8	28.8	32.9	34.3
Model c	54.5	57.4	52.8	58.9	60.6	32.9	34.1	29.3	34.0	34.3
Model d	64.0	65.7	62.6	67.3	68.1	37.8	39.7	35.4	40.1	41.0
Model e	67.1	69.2	66.3	70.5	71.3	42.2	43.3	38.7	43.9	44.6
Model f	68.6	70.4	67.3	71.6	72.1	45.9	48.6	41.4	49.1	51.6

Table S2.

The performances of different ablation study models on the HomoPDB and HeteroPDB test sets

Methods	DHTest (precision %)					DB5.5 (precision %)				
	L/5	L/10	50	10	5	L/5	L/10	50	10	5
DeepHomo	45.0 (36.5)	47.4 (38.2)	44.2 (36.1)	50.2 (40.4)	50.9 (42.9)					
GLINTER	48.5 (46.7)	50.8 (48.8)	47.4 (45.1)	51.2 (51.0)	52.3 (51.5)	18.8 (16.1)	20.5 (15.0)	15.2 (12.6)	21.9 (15.6)	20.7 (16.9)
DRN-1D2D_Inter	52.1	53.3	52.3	54.9	56.0	24.4	26.9	20.7	25.4	27.5
DeepHomo2	59.5 (50.7)	60.0 (51.7)	58.7 (50.5)	61.3 (52.5)	62.0 (52.6)					
CDPred	69.2 (61.4)	72.2 (62.4)	69.1 (61.0)	73.7 (63.3)	74.8 **(64.1)**	27.7 (24.7)	29.3 (26.6)	24.7 (22.8)	30.6 (27.4)	30.2 (28.1)
PLMGraph-Inter	**72.0 (61.7)**	**73.3 (62.6)**	**71.9 (61.1)**	**74.7 (63.4)**	**75.4** (63.5)	**33.6 (28.6)**	**36.4 (31.4)**	**29.5 (23.8)**	**36.9 (30.3)**	**40.0 (31.5)**

Note: The highest mean precision (%) in each column is highlighted in bold.

Table S3.

The performances of DeepHomo, GLINTER, DRN-1D2D_Inter, DeepHomo2, CDPred and PLMGraph-Inter on the DHTest and DB5.5 test sets using experimental structures (AlphaFold2 predicted structures)

Methods	DHTest (precision %)					DB5.5 (precision %)				
	L/5	L/10	50	10	5	L/5	L/10	50	10	5
DeepHomo	41.6 (32.6)	43.9 (33.7)	41.3 (32.9)	47.6 (35.8)	48.6 (37.8)					
GLINTER	48.5 (41.6)	50.8 (43.4)	47.4 (40.4)	51.2 (46.1)	52.3 (46.3)	18.8 (15.5)	20.5 (14.0)	15.2 (12.6)	21.9 (15.4)	20.7 (16.4)
DRN-1D2D_Inter	45.5	46.9	45.5	48.9	49.9	23.4	25.3	19.9	24.6	25.7
DeepHomo2	56.3 (48.2)	57.1 (49.7)	56.0 (48.4)	58.9 (50.1)	59.2 (49.4)					
CDPred	**65.4 (56.9)**	**67.9 (58.1)**	**65.4 (56.9)**	**69.7 (59.3)**	**70.8 (59.7)**	27.8 (24.3)	29.2 (25.9)	24.4 (22.6)	30.7 (27.1)	30.3 (27.9)
PLMGraph-Inter	64.7 (55.8)	66.2 (57.0)	65.3 (55.8)	68.3 (57.5)	69.2 (57.3)	**32.4 (27.9)**	**34.4 (30.4)**	**28.5 (23.4)**	**35.5 (30.0)**	**38.6 (30.7)**

Note: The highest mean precision (%) in each column is highlighted in bold.

Table S4.

The performances of DeepHomo, GLINTER, DRN-1D2D_Inter, DeepHomo2, CDPred and PLMGraph-Inter on DHTest and DB5.5 after the removal of targets which GLINTER failed to make the prediction using experimental structures (AlphaFold2 predicted structures)

Methods	Precision@Top 50 (%)	
	Homodimer	Heterodimer
AlphaFold-Multimer	76.5	67.8
PLMGraph-Inter (experimental structure as inputs)	68.7	38.6
PLMGraph-Inter (AlphaFold2 predicted structure as inputs)	61.2	34.6

Table S5.

The performances of AlphaFold-Multimer and PLMGraph-Inter on the homodimer and heterodimer test sets

References

1. Alberts B (1998) **The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists** *Cell* **92**:291–294 https://doi.org/10.1016/S0092-8674(00)80922-8

2. Basu S., Wallner B (2016) **DockQ: A Quality Measure for Protein-Protein Docking Models** *PLOS ONE* **11** https://doi.org/10.1371/journal.pone.0161879

3. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E (2000) **The Protein Data Bank** *Nucleic Acids Research* **28**:235–242 https://doi.org/10.1093/nar/28.1.235

4. Bonvin A. M (2006) **Flexible protein–protein docking** *Current Opinion in Structural Biology* **16**:194–200 https://doi.org/10.1016/j.sbi.2006.02.002

5. Cong Q., Anishchenko I., Ovchinnikov S., Baker D (2019) **Protein interaction networks revealed by proteome coevolution** *Science* **365**:185–189 https://doi.org/10.1126/science.aaw6718

6. Dauparas J. *et al.* (2022) **Robust deep learning–based protein sequence design using ProteinMPNN** *Science* **0** https://doi.org/10.1126/science.add2187

7. Dominguez C., Boelens R., Bonvin A. M. J. J (2003) **HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information** *Journal of the American Chemical Society* **125**:1731–1737 https://doi.org/10.1021/ja026939x

8. Evans R. *et al.* (2022) **Protein complex prediction with AlphaFold-Multimer** *bioRxiv* https://doi.org/10.1101/2021.10.04.463034

9. Evans R. *et al.* (2022) **Protein complex prediction with AlphaFold-Multimer** *bioRxiv* https://doi.org/10.1101/2021.10.04.463034

10. Goodsell D. S., Olson A. J (2000) **Structural Symmetry and Protein Function** *Annual Review of Biophysics and Biomolecular Structure* **29**:105–153 https://doi.org/10.1146/annurev.biophys.29.1.105

11. Green A. G., Elhabashy H., Brock K. P., Maddamsetti R., Kohlbacher O., Marks D. S (2021) **Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences** *Nature Communications* **12** https://doi.org/10.1038/s41467-021-21636-z

12. Guo Z., Liu J., Skolnick J., Cheng J (2022) **Prediction of inter-chain distance maps of protein complexes with 2D attention-based deep neural networks** *Nature Communications* **13** https://doi.org/10.1038/s41467-022-34600-2

13. Hanson J., Paliwal K., Litfin T., Yang Y., Zhou Y (2018) **Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks** *Bioinformatics* **34**:4039–4045 https://doi.org/10.1093/bioinformatics/bty481

14. Honorato R. V., Koukos P. I., Jiménez-García B., Tsaregorodtsev A., Verlato M., Giachetti A., Rosato A., Bonvin A. M. J. J (2021) **Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem** *Frontiers in Molecular Biosciences* **8** https://doi.org/10.3389/fmolb.2021.729513

15. Hopf T. A., Schärfe C. P. I., Rodrigues J. P. G. L. M., Green A. G., Kohlbacher O., Sander C., Bonvin A. M. J. J., Marks D. S (2014) **Sequence co-evolution gives 3D contacts and structures of protein complexes** *eLife* **3** https://doi.org/10.7554/eLife.03430

16. Hsu C., Verkuil R., Liu J., Lin Z., Hie B., Sercu T., Lerer A., Rives A (2022) **Learning inverse folding from millions of predicted structures** *bioRxiv* https://doi.org/10.1101/2022.04.10.487779

17. Hsu C., Verkuil R., Liu J., Lin Z., Hie B., Sercu T., Lerer A., Rives A (2022) **Learning inverse folding from millions of predicted structures** *bioRxiv* https://doi.org/10.1101/2022.04.10.487779

18. Jing B., Eismann S., Soni P. N., Dror R. O. (2021) **Equivariant Graph Neural Networks for 3D Macromolecular Structure** *arXiv*

19. Jing B., Eismann S., Suriana P., Townshend R. J. L., Dror R. O (2021) **Learning from protein structure with geometric vector perceptrons** *arXiv*

20. Jones D. T., Singh T., Kosciolek T., Tetchner S (2015) **MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics (Oxford** *England* **31**:999–1006 https://doi.org/10.1093/bioinformatics/btu791

21. Ju F., Zhu J., Shao B., Kong L., Liu T. Y., Zheng W. M., Bu D (2021) **CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction** *Nature Communications* **12** https://doi.org/10.1038/S41467-021-22869-8

22. Jumper J. *et al.* (2021) **Highly accurate protein structure prediction with AlphaFold** *Nature* **596** https://doi.org/10.1038/s41586-021-03819-2

23. Li H., Huang S.-Y (2021) **Protein–protein docking with interface residue restraints\ast** *Chinese Physics B* **30** https://doi.org/10.1088/1674-1056/abc14e

24. Li W., Jaroszewski L., Godzik A (2001) **Clustering of highly homologous sequences to reduce the size of large protein databases** *Bioinformatics* **17**:282–283 https://doi.org/10.1093/bioinformatics/17.3.282

25. Li Y., Hu J., Zhang C., Yu D. J (2019) **ResPRE: High-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks** *Bioinformatics* **35**:4647–4655 https://doi.org/10.1093/bioinformatics/btz291

26. Lin P., Yan Y., Huang S.-Y (2023) **DeepHomo2.0: Improved protein–protein contact prediction of homodimers by transformer-enhanced deep learning** *Briefings in Bioinformatics* **24** https://doi.org/10.1093/bib/bbac499

27. Martino E., Chiarugi S., Margheriti F., Garau G (2021) **Mapping, Structure and Modulation of PPI** *Frontiers in Chemistry* **9** https://doi.org/10.3389/fchem.2021.718405

28. Ovchinnikov S., Kamisetty H., Baker D (2014) **Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information** *eLife* **3** https://doi.org/10.7554/eLife.02030

29. Pagès G., Charmettant B., Grudinin S (2019) **Protein model quality assessment using 3D oriented convolutional neural networks** *Bioinformatics* **35**:3313–3319 https://doi.org/10.1093/bioinformatics/btz122

30. Potter S. C., Luciani A., Eddy S. R., Park Y., Lopez R., Finn R. D (2018) **HMMER web server: 2018 update** *Nucleic Acids Research* **46**:W200–W204 https://doi.org/10.1093/nar/gky448

31. Rao R., Liu J., Verkuil R., Meier J., Canny J. F., Abbeel P., Sercu T., Rives A (2021) **MSA Transformer** *bioRxiv*

32. Rao R. M., Liu J., Verkuil R., Meier J., Canny J., Abbeel P., Sercu T., Rives A. (2021) **MSA Transformer** :8844–8856

33. Rao R. M., Liu J., Verkuil R., Meier J., Canny J., Abbeel P., Sercu T., Rives A. (2021) **MSA Transformer** :8844–8856

34. Rives A. *et al.* (2021) **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences** *Proceedings of the National Academy of Sciences of the United States of America* **118**:1–46 https://doi.org/10.1073/pnas.2016239118

35. Roy R. S., Quadir F., Soltanikazemi E., Cheng J (2022) **A deep dilated convolutional residual network for predicting interchain contacts of protein homodimers** *Bioinformatics* **38**:1904–1910 https://doi.org/10.1093/bioinformatics/btac063

36. Seemayer S., Gruber M., Söding J (2014) **CCMpred—Fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics (Oxford** *England* **30**:3128–3130 https://doi.org/10.1093/bioinformatics/btu500

37. Si Y., Yan C (2021) **Improved protein contact prediction using dimensional hybrid residual networks and singularity enhanced loss function** *Briefings in Bioinformatics* **22** https://doi.org/10.1093/bib/bbab341

38. Si Y., Yan C (2022) **Protein complex structure prediction powered by multiple sequence alignments of interologs from multiple taxonomic ranks and AlphaFold2** *Briefings in Bioinformatics* **23** https://doi.org/10.1093/bib/bbac208

39. Si Y., Yan C (2023) **Improved inter-protein contact prediction using dimensional hybrid residual networks and protein language models. Briefings in Bioinformatics** *bbad* **39** https://doi.org/10.1093/bib/bbad039

40. Sledzieski S., Singh R., Cowen L., Berger B (2021) **D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions** *Cell Systems* **12**:969–982 https://doi.org/10.1016/j.cels.2021.08.010

41. Spirin V., Mirny L. A (2003) **Protein complexes and functional modules in molecular networks** *Proceedings of the National Academy of Sciences* **100**:12123–12128 https://doi.org/10.1073/pnas.2032324100

42. Steinegger M., Meier M., Mirdita M., Vöhringer H., Haunsberger S. J., Söding J (2019) **HH-suite3 for fast remote homology detection and deep protein annotation** *BMC Bioinformatics* **20** https://doi.org/10.1186/s12859-019-3019-7

43. Steinegger M., Söding J (2018) **Clustering huge protein sequence sets in linear time** *Nature Communications* **9** https://doi.org/10.1038/s41467-018-04964-5

44. Sun D., Liu S., Gong X (2020) **Review of multimer protein–protein interaction complex topology and structure prediction\ast** *Chinese Physics B* **29** https://doi.org/10.1088/1674-1056/abb659

45. Uguzzoni G., John Lovis S., Oteri F., Schug A., Szurmant H., Weigt M (2017) **Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis** *Proceedings of the National Academy of Sciences* **114**:E2662–E2671 https://doi.org/10.1073/pnas.1615068114

46. van Zundert G. C. P., Rodrigues J. P. G. L. M., Trellet M., Schmitz C., Kastritis P. L., Karaca E., Melquiond A. S. J., van Dijk M., de Vries S. J., Bonvin A. M. J. J. (2016) **The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes** *Journal of Molecular Biology* **428**:720–725 https://doi.org/10.1016/j.jmb.2015.09.014

47. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I (2017) **Attention is All you Need** *Advances in Neural Information Processing Systems* **30**

48. Wang S., Sun S., Li Z., Zhang R., Xu J (2017) **Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model** *PLOS Computational Biology* **13** https://doi.org/10.1371/journal.pcbi.1005324

49. Weigt M., White R. A., Szurmant H., Hoch J. A., Hwa T (2009) **Identification of direct residue contacts in protein–protein interaction by message passing** *Proceedings of the National Academy of Sciences* **106**:67–72 https://doi.org/10.1073/pnas.0805923106

50. Wu F., Wu L., Radev D., Xu J., Li S. Z (2023) **Integration of pre-trained protein language models into geometric deep learning networks** *Communications Biology* **6** https://doi.org/10.1038/s42003-023-05133-1

51. Wu T., Huang H., Li J., Wang W., Gong X (2022) **Inter-chain contact map prediction for protein complex based on graph attention network and triangular multiplication update** *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM* :2143–2148 https://doi.org/10.1109/BIBM55620.2022.9995360

52. Xie Z., Xu J (2022) **Deep graph learning of inter-protein contacts** *Bioinformatics* **38**:947–953 https://doi.org/10.1093/bioinformatics/btab761

53. Yan Y., Huang S. Y (2021) **Accurate prediction of inter-protein residue-residue contacts for homo-oligomeric protein complexes** *Briefings in Bioinformatics* **22**:1–13 https://doi.org/10.1093/bib/bbab038

54. Zeng H., Wang S., Zhou T., Zhao F., Li X., Wu Q., Xu J (2018) **ComplexContact: A web server for inter-protein contact prediction using deep learning** *Nucleic Acids Research* **46**:W432–W437 https://doi.org/10.1093/nar/gky420

55. Zhang Y., Skolnick J (2004) **Scoring function for automated assessment of protein structure template quality. Proteins: Structure** *Function, and Bioinformatics* **57**:702–710 https://doi.org/10.1002/prot.20264

56. Zhang Y., Skolnick J (2005) **TM-align: A protein structure alignment algorithm based on the TM-score** *Nucleic Acids Research* **33**:2302–2309 https://doi.org/10.1093/nar/gki524

Article and author information

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

Chengfei Yan
School of Physics, Huazhong University of Science and Technology, China
For correspondence: chengfeiyan@hust.edu.cn
ORCID iD: 0000-0002-2010-6668

Copyright

© 2023, Yunda Si & Chengfei Yan

This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Editors

Reviewing Editor
Anne-Florence Bitbol
Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland

Senior Editor
Aleksandra Walczak
École Normale Supérieure - PSL, Paris, France

Reviewer #1 (Public Review):

Summary:

Given knowledge of the amino acid sequence and of some version of the 3D structure of two monomers that are expected to form a complex, the authors investigate whether it is possible to accurately predict which residues will be in contact in the 3D structure of the expected complex. To this effect, they train a deep learning model which takes as inputs the geometric structures of the individual monomers, per-residue features (PSSMs) extracted from MSAs for each monomer, and rich representations of the amino acid sequences computed with the pre-trained protein language models ESM-1b, MSA Transformer, and ESM-IF. Predicting inter-protein contacts in complexes is an important problem. Multimer variants of AlphaFold, such as AlphaFold-Multimer, are the current state of the art for full protein complex structure prediction, and if the three-dimensional structure of a complex can be accurately predicted then the inter-protein contacts can also be accurately determined. By contrast, the method presented here seeks state-of-the-art performance among models that have been trained end-to-end for inter-protein contact prediction.

Strengths:

The paper is carefully written and the method is very well detailed. The model works both for homodimers and heterodimers. The ablation studies convincingly demonstrate that the chosen model architecture is appropriate for the task. Various comparisons suggest that PLMGraph-Inter performs substantially better, given the same input, than DeepHomo, GLINTER, CDPred, DeepHomo2, and DRN-1D2D_Inter.
The authors control for some degree of redundancy between their training and test sets, both using sequence and structural similarity criteria. This is more careful than can be said of most works in the field of PPI prediction.
As a byproduct of the analysis, a potentially useful heuristic criterion for acceptable contact prediction quality is found by the authors: namely, to have at least 50% precision in the prediction of the top 50 contacts.

Weaknesses:

The authors check for performance drops when the test set is restricted to pairs of interacting proteins such that the chain pair is not similar *as a pair* (in sequence or structure) to a pair present in the training set. A more challenging test would be to restrict the test set to pairs of interacting proteins such that *none* of the chains are separately similar to monomers present in the training set. In the case of structural similarity (TM-scores), this would amount to replacing the two "min"s with "max"s in Eq. (4). In the case of sequence similarity, one would simply require that no monomer in the test set is in any MMSeqs2 cluster observed in the training set. This may be an important check to make, because a protein may interact with several partners, and/or may use the same sites for several distinct interactions, contributing to residual data leakage in the test set.

The training set of AFM with v2 weights has a global cutoff of 30 April 2018, while that of PLMGraph-Inter has a cutoff of March 7 2022. So there may be structures in the test set for PLMGraph-Inter that are not in the training set of AFM with v2 weights (released between May 2018 and March 2022). The "Benchmark 2" dataset from the AFM paper may have a few additional structures not in the training or test set for PLMGraph-Inter. I realize there may be only few structures that are in neither training set, but still think that showing the comparison between PLMGraph-Inter and AFM there would be important, even if no statistically significant conclusions can be drawn.

Finally, the inclusion of AFM confidence scores is very good. A user would likely trust AFM predictions when the confidence score is high, but look for alternative predictions when it is low. The authors' analysis (Figure 6, panels c and d) seems to suggest that, in the case of heterodimers, when AFM has low confidence, PLMGraph-Inter improves precision by (only) about 3% on average. By comparison, the reported gains in the "DockQ-failed" and "precision-failed" bins are based on knowledge of the ground truth final structure, and thus are not actionable in a real use-case.

https://doi.org/10.7554/eLife.92184.2.sa1

Reviewer #2 (Public Review):

This work introduces PLMGraph-Inter, a new deep learning approach for predicting inter-protein contacts, which is crucial for understanding protein-protein interactions. Despite advancements in this field, especially driven by AlphaFold, prediction accuracy and efficiency in terms of computational cost still remains an area for improvement. PLMGraph-Inter utilizes invariant geometric graphs to integrate the features from multiple protein language models into the structural information of each subunit. When compared against

other inter-protein contact prediction methods, PLMGraph-Inter shows better performance which indicates that utilizing both sequence embeddings and structural embeddings is important to achieve high-accuracy predictions with relatively smaller computational costs for the model training.

https://doi.org/10.7554/eLife.92184.2.sa0

Author Response

The following is the authors' response to the current reviews.

Overall Response

We thank the reviewers for reviewing our manuscript, recognizing the significance of our study, and offering valuable suggestions. Based on the reviewer's comments and the updated eLife assessment, we would like to chose the current version of our manuscript as the Version of Record of our manuscript.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *Given knowledge of the amino acid sequence and of some version of the 3D structure of two monomers that are expected to form a complex, the authors investigate whether it is possible to accurately predict which residues will be in contact in the 3D structure of the expected complex. To this effect, they train a deep learning model which takes as inputs the geometric structures of the individual monomers, per-residue features (PSSMs) extracted from MSAs for each monomer, and rich representations of the amino acid sequences computed with the pre-trained protein language models ESM-1b, MSA Transformer, and ESM-IF. Predicting inter-protein contacts in complexes is an important problem. Multimer variants of AlphaFold, such as AlphaFold-Multimer, are the current state of the art for full protein complex structure prediction, and if the three-dimensional structure of a complex can be accurately predicted then the inter-protein contacts can also be accurately determined. By contrast, the method presented here seeks state-of-the-art performance among models that have been trained end-to-end for inter-protein contact prediction.*
>
> *Strengths:*
>
> *The paper is carefully written and the method is very well detailed. The model works both for homodimers and heterodimers. The ablation studies convincingly demonstrate that the chosen model architecture is appropriate for the task. Various comparisons suggest that PLMGraph-Inter performs substantially better, given the same input, than DeepHomo, GLINTER, CDPred, DeepHomo2, and DRN-1D2D_Inter.*
>
> *The authors control for some degree of redundancy between their training and test sets, both using sequence and structural similarity criteria. This is more careful than can be said of most works in the field of PPI prediction.*
>
> *As a byproduct of the analysis, a potentially useful heuristic criterion for acceptable contact prediction quality is found by the authors: namely, to have at least 50% precision in the prediction of the top 50 contacts.*

We thank the reviewer for recognizing the strengths of our work!

> *Weaknesses:*
>
> *The authors check for performance drops when the test set is restricted to pairs of interacting proteins such that the chain pair is not similar* as a pair *(in sequence or structure) to a pair present in the training set. A more challenging test would be to restrict the test set to pairs of interacting proteins such that* none *of the chains are separately similar to monomers present in the training set. In the case of structural similarity (TM-scores), this would amount to replacing the two "min"s with "max"s in Eq. (4). In the case of sequence similarity, one would simply require that no monomer in the test set is in any MMSeqs2 cluster observed in the training set. This may be an important check to make, because a protein may interact with several partners, and/or may use the same sites for several distinct interactions, contributing to residual data leakage in the test set.*

We thank the reviewer for the suggestion! In the case of protein-protein prediction (“0D prediction”) or protein-protein interfacial residue prediction(“1D prediction”), we think making none of the chains in the test set separately similar to monomers in the training set is necessary, as the reviewer pointed out that a protein may interact with several partners, and may even use the same sites for the interactions. Since the task of this study is predicting the inter-protein residue-residue contacts (“2D prediction”), even though a protein uses the same site to interact with different partners, as long as the interacting partners are different, the inter-protein contact maps would be different. Therefore, we don’t think that in our task, making this restriction to the test set is necessary.

> *The training set of AFM with v2 weights has a global cutoff of 30 April 2018, while that of PLMGraph-Inter has a cutoff of March 7 2022. So there may be structures in the test set for PLMGraph-Inter that are not in the training set of AFM with v2 weights (released between May 2018 and March 2022). The "Benchmark 2" dataset from the AFM paper may have a few additional structures not in the training or test set for PLMGraph-Inter. I realize there may be only few structures that are in neither training set, but still think that showing the comparison between PLMGraph-Inter and AFM there would be important, even if no statistically significant conclusions can be drawn.*

We thank the reviewer for the suggestion! It is not enough to only use the date cutoff to remove the redundancy, since similar structures can be deposited in the PDB in different dates. Because AFM does not release the PDB codes of its training set, it is difficult for us to totally remove the redundancy. Therefore, we think no rigorous conclusion can be drawn by including these comparisons in the manuscript. Besides, the main point of this study is to demonstrate that the integration of multiple protein language models using protein geometric graphs can dramatically improve the model performance for inter-protein contact prediction, which can provide some important enlightenments for the future development of more powerful protein complex structure prediction methods beyond AFM, rather than providing a tool which can beat AFM at this moment. We think including too many stuffs in the comparison with AFM may distract the readers. Therefore, we choose to not include these comparisons in the manuscript.

> *Finally, the inclusion of AFM confidence scores is very good. A user would likely trust AFM predictions when the confidence score is high, but look for alternative predictions when it is low. The authors' analysis (Figure 6, panels c and d) seems to suggest that, in the case of heterodimers, when AFM has low confidence, PLMGraph-Inter improves precision by (only) about 3% on average. By comparison, the reported gains in the "DockQ-failed" and "precision-failed" bins are based on knowledge of the ground truth final structure, and thus are not actionable in a real use-case.*

We agree with the reviewer that more studies are needed for providing a model which can well complement or even beat AFM. The main point of this study is to demonstrate that the integration of multiple protein language models using protein geometric graphs can dramatically improve the model performance for inter-protein contact prediction, which can provide some important enlightenments for the future development of more powerful protein complex structure prediction methods beyond AFM.

> ***Reviewer #2 (Public Review):***
>
> *This work introduces PLMGraph-Inter, a new deep learning approach for predicting inter-protein contacts, which is crucial for understanding proteinprotein interactions. Despite advancements in this field, especially driven by AlphaFold, prediction accuracy and efficiency in terms of computational cost still remains an area for improvement. PLMGraph-Inter utilizes invariant geometric graphs to integrate the features from multiple protein language models into the structural information of each subunit. When compared against other inter-protein contact prediction methods, PLMGraph-Inter shows better performance which indicates that utilizing both sequence embeddings and structural embeddings is important to achieve high-accuracy predictions with relatively smaller computational costs for the model training.*

We thank the reviewer for recognizing the strengths of our work!

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> - *I recommend renaming the section "Further potential redundancies removal between the training and the test" to "Further potential redundancies removal between the training and the test sets"*

Changed.

> - *In lines 768-769, the sentence seems to end prematurely in "to use more stringent threshold in the redundancy removal"*

Corrected.

> - *In Eq. (4), line 789, there are many instances of dashes that look like minus signs, creating some confusion.*

Corrected.

> - *I think I may have mixed up figure references in my first review. When I said (Recommendations to the authors): "p. 22, line 2: from the figure, I would have guessed "greater than or equal to 0.7", not 0.8", I think I was referring to what is now lines 423-424, referring to what is now Figure 5c. The point stands there, I think.*

Corrected.

> - *A couple of new grammatical mishaps have been introduced in the revision. These could be rectified.*

We carefully rechecked our revisions, and corrected the grammatical issues we found.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *Most of my concerns were resolved through the revision. I have only one suggestion for the main figure.*
>
> *The current scatter plots in Figure 2 are hard to understand as too many different methods are abstracted into a single plot with multiple colors. I would suggest comparing their performances using box plot or violin plot for the figure 2.*

We thank the reviewer for the suggestion! In the revision, we tried violin plot, but it does not look good since too many different methods are included in the plot. Besides, we chose the scatter plot as it can provide much more details. We also provided the individual head-to-head scatter plots as supplementary figures, we think which can also be helpful for the readers to capture the information of the figures.

The following is the authors' response to the original reviews.

Overall Response

We would like to thank the reviewers for reviewing our manuscript, recognizing the significance of our study, and offering valuable suggestions. We have carefully revised the manuscript to address all the concerns and suggestions raised by the reviewers.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *Given knowledge of the amino acid sequence and of some version of the 3D structure of two monomers that are expected to form a complex, the authors investigate whether it is possible to accurately predict which residues will be in contact in the 3D structure of the expected complex. To this effect, they train a deep learning model that takes as inputs the geometric structures of the individual monomers, per-residue features (PSSMs) extracted from MSAs for each monomer, and rich representations of the amino acid sequences computed with the pre-trained protein language models ESM-1b, MSA Transformer, and ESM-IF. Predicting inter-protein contacts in complexes is an important problem. Multimer variants of AlphaFold, such as AlphaFold-Multimer, are the current state of the art for full protein complex structure prediction, and if the three-dimensional structure of a complex can be accurately predicted then the inter-protein contacts can also be accurately determined. By contrast, the method presented here seeks state-of-the-art performance among models that have been trained end-to-end for inter-protein contact prediction.*
>
> *Strengths:*
>
> *The paper is carefully written and the method is very well detailed. The model works both for homodimers and heterodimers. The ablation studies convincingly demonstrate that the chosen model architecture is appropriate for the task. Various comparisons suggest that PLMGraph-Inter performs substantially better, given the same input than DeepHomo, GLINTER, CDPred, DeepHomo2, and DRN-1D2D_Inter. As a byproduct of the analysis, a potentially useful heuristic criterion for acceptable contact prediction quality is found by the authors: namely, to have at least 50% precision in the prediction of the top 50 contacts.*

We thank the reviewer for recognizing the strengths of our work!

> *Weaknesses:*
>
> *My biggest issue with this work is the evaluations made using* bound *monomer structures as inputs, coming from the very complexes to be predicted. Conformational changes in protein-protein association are the key element of the binding mechanism and are challenging to predict. While the GLINTER paper (Xie & Xu, 2022) is guilty of the same sin, the authors of CDPred (Guo et al., 2022) correctly only report test results obtained using predicted unbound tertiary structures as inputs to their model. Test results using experimental monomer structures in bound states can hide important limitations in the model, and thus say very little about the realistic use cases in which only the unbound structures (experimental or predicted) are available. I therefore strongly suggest reducing the importance given to the results obtained using bound structures and emphasizing instead those obtained using predicted monomer structures as inputs.*

We thank the reviewer for the suggestion! In the revision, to emphasize the performance of PLMGraph-Inter using the predicted monomer structures, we moved the evaluation results based on the predicted monomer from the supplementary to the main text (see the new Table 1 and Figure 2 in the revised manuscript) and re-organized the two subsections “Evaluation of PLMGraph-Inter on HomoPDB and HeteroPDB test sets” and “Impact of the monomeric structure quality on contact prediction” in the main text.

> *In particular, the most relevant comparison with AlphaFold-Multimer (AFM) is given in Figure S2,* not *Figure 6. Unfortunately, it substantially shrinks the proportion of structures for which AFM fails while PLMGraph-Inter performs decently. Still, it would be interesting to investigate why this occurs. One possibility would be that the predicted monomer structures are of bad quality there, and PLMGraph-Inter may be able to rely on a signal from its language model features instead. Finally, AFM multimer confidence values ("iptm + ptm") should be provided, especially in the cases in which AFM struggles.*

We thank the reviewer for the suggestion! It is worth noting that AFM automatically searches monomer templates in the prediction, and when we checked our AFM runs, we found that 99% of the targets in our study (including all the targets in the four datasets: HomoPDB, HeteroPDB, DHTest and DB5.5) at least 20 templates were identified (AFM employed the top 20 templates in the prediction), and 87.8% of the targets employed the native templates (line 455-462 in page 25 in the subsection of “Comparison of PLMGraph-Inter with AlphaFold-Multimer”). Therefore, we think Figure 6 not Figure S5 (the original Figure S2) shows a fairer comparison. Besides, it is also worth noting the targets used in this study would have a large overlap with the training set of AlphaFold-Multimer, since AFM used all protein complex structures in PDB deposited before 2018-04-30 in the model training, which would further cause the overestimation of the performance of AFM (line 450-455 in page 24-25 in the subsection of “Comparison of PLMGraph-Inter with AlphaFold-Multimer”).

To mimic the performance of AlphaFold2 in real practice and produce predicted monomeric structures with more diverse qualities, we only used the MSA searched from Uniref100 protein sequence database as the input to AlphaFold2 and set to not use the template (line 203~210 in page 12 in the subsection of “Evaluation of PLMGraph-Inter on HomoPDB and HeteroPDB test sets”). Since some of the predicted monomer structures are of bad quality, it is reasonable that the performance of PLMGraph-Inter drops when the predicted monomeric structures are used in the prediction. We provided a detailed analysis of the impact of the monomeric structure quality on the prediction performance in the subsection “Impact of the monomeric structure quality on contact prediction” in the main text.

We provided the analysis of the AFM multimer confidence values ("iptm + ptm") in the revision (Figure 6, Figure S5 and line 495-501 in page 27 in the subsection of "Comparison of PLMGraph-Inter with AlphaFold-Multimer").

> *Besides, in cases where* any *experimental structures - bound or unbound - are available and given to PLMGraph-Inter as inputs, they should also be provided to AlphaFold-Multimer (AFM) as templates. Withholding these from AFM only makes the comparison artificially unfair. Hence, a new test should be run using AFM templates, and a new version of Figure 6 should be produced. Additionally, AFM's mean precision, at least for top-50 contact prediction, should be reported so it can be compared with PLMGraph-Inter's.*

We thank the reviewers for the suggestion, and we are sorry for the confusion! In the AFM runs to predict protein complex structures, we used the default setting of AFM which automatically searches monomer templates in the prediction. When we checked our AFM runs, we found that 99% of the targets in our study (including all the targets in the four datasets: HomoPDB, HeteroPDB, DHTest and DB5.5) employed at least 20 templates in their predictions (AFM only used the top 20 templates), and 87.8% of the targets employed the native template. We further clarified this in the revision (line 455462 in page 25 in the subsection of "Comparison of PLMGraph-Inter with AlphaFoldMultimer"). We also included the mean precisions of AFM (top-50 contact prediction) in the revision (Table S5 and line 483-484 in page 26 in the subsection of "Comparison of PLMGraph-Inter with AlphaFold-Multimer").

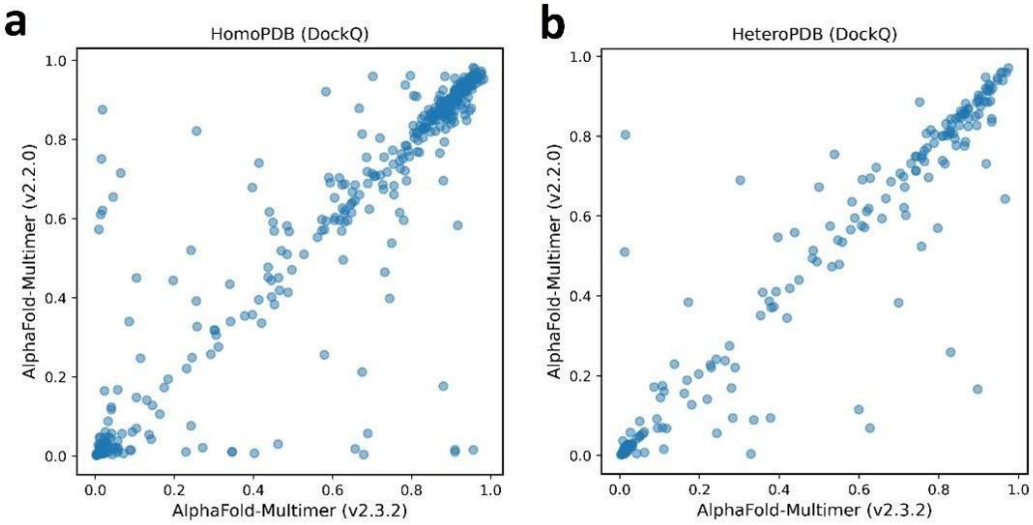
> *It's a shame that many of the structures used in the comparison with AFM are actually in the AFM v2 training set. If there are any outside the AFM v2 training set and, ideally, not sequence- or structure-homologous to anything in the AFM v2 training set, they should be discussed and reported on separately. In addition, why not test on structures from the "Benchmark 2" or "Recent-PDB-Multimers" datasets used in the AFM paper?*

We thank the reviewer for the suggestion! The biggest challenge to objectively evaluate AFM is that as far as we known, AFM does not release the PDB ids of its training set and the "Recent-PDB-Multimers" dataset. "Benchmark 2" only includes 17 heterodimer proteins, and the number would be further decreased after removing targets redundant to our training set. We think it is difficult to draw conclusions from such a small number of targets.

> *It is also worth noting that the AFM v2 weights have now been outdated for a while, and better v3 weights now exist, with a training cutoff of 2021-09-30.*

Author response image 1.

The head-to-head comparison of qualities of complex predicted by AlphaFold-Multimer (2.2.0) and AlphaFold-Multimer (2.3.2) for each target PPI.



We thank the reviewer for reminding the new version of AFM. The only difference between AFM V3 and V2 is the cutoff date of the training set. During the revision, we also tested the new version of AFM on the datasets of HomoPDB and HeteroPDB, but we found the performance difference between the two versions of AFM is actually very little (see the figure above, not shown in the main text). One reason might be that some targets in HomoPDB and HeteroPDB are redundant with the training sets of the two version of AFM. Since our test sets would have more overlaps with the training set of AFM V3, we keep using the AFM V2 weights in this study.

> *Another weakness in the evaluation framework: because PLMGraph-Inter uses structural inputs, it is not sufficient to make its test set non-redundant in sequence to its training set. It must also be non-redundant in structure. The Benchmark 2 dataset mentioned above is an example of a test set constructed by removing structures with homologous templates in the AF2 training set. Something similar should be done here.*

We thank the reviewer for the suggestion! In the revision, we explored the performance of PLMGraph-Inter when using different thresholds of fold similarity scores of interacting monomers to further remove potential redundancies between the training and test sets (i.e. redundancy in structure) (line 353-386 in page 19-21 in the subsection “Ablation study”; line 762-797 in page 41-43 in the subsection “Further potential redundancies removal between the training and the test”). We found that for heteromeric PPIs (targets in HeteroPDB), the further removal of potential redundancy in structure has little impact on the model performance (~3%, when TM-score 0.5 is used as the threshold). However, for homomeric PPIs (targets in HomoPDB), the further removal of potential redundancy in structure significantly reduce the model performance (~18%, when TM-score 0.5 is used as the threshold) (see Table 2). One possible reason for this phenomenon is that the binding mode of the homomeric PPI is largely determined by the fold of its monomer, thus the does not generalize well on targets whose folds have never been seen during the training.

Whether the deep learning model can generalize well on targets with novel folds is a very interesting and important question. We thank the reviewer for pointing out this! However, to the best of our knowledge, this question has rarely been addressed by previous studies including AFM. For example, the Benchmark 2 dataset is prepared by ClusPro TBM (bioRxiv 2021.09.07.459290; Proteins 2020, 88:1082-1090) which uses a sequence-based approach (HHsearch) to identify templates not structure-based. Therefore, we don’t think this dataset is non-redundant in structure.

> *Finally, the performance of DRN-1D2D for top-50 precision reported in Table 1 suggests to me that, in an ablation study, language model features alone would yield better performance than geometric features alone. So, I am puzzled why model "a" in the ablation is a "geometry-only" model and not a "LM-only" one.*

Using the protein geometric graph to integrate multiple protein language models is the main idea of PLMGraph-Inter. Comparing with our previous work (DRN-1D2D_Inter), we consider the building of the geometric graph as one major contribution of this work. To emphasize the efficacy of this geometric graph, we chose to use the “geometry-only” model as the base model.

> ***Reviewer #1 (Recommendations For The Authors):***
>
> *Some sections of the paper use technical terminology which limits accessibility to a broad audience. An obvious example is in the section "Results > Overview of PLMGraph-Inter > The residual network module": the average eLife reader is not a machine learning expert and might not be familiar with a "convolution with kernel size of 1 * 1". In general, the "Overview of PLMGraph-Inter" is a bit heavy with technical details, and I suggest moving many of these to Methods. This overview section can still be there but it should be shorter and written using less technical language.*

We thank the reviewer for the suggestion! We moved some technical details to the Methods section in the revision (line 184-185 in page 11; line 729-735 in page 39).

> *List of typos and minor issues (page number according to merged PDF):*
>
> - *p. 3. line -3: remove "to"*

Corrected (line 36, page 3)

> - *p. 5, line 7: "GINTER" should be "GLINTER"*

Corrected (line 64, page 5)

> - *p. 6, line -4: "Given structures" -> "Given the structures"*

Corrected (line 95, page 6)

> - *p. 6, line -2: "with which encoded"... ?*

We rephrased this sentence in revision. (line 97, page 6)

> - *p. 9, line 1: "principal" -> "principle"*

Corrected (line 142, page 9)

> - *p. 13, line 1: "has" -> "but have"*

Corrected (line 231, page 13)

- *p. 14, lines 6-7: "As can be seen from the figure that the predicted" -> "As can be seen from the figure, the predicted"*

We rephrased this paragraph, and the sentence was deleted in the revision (line 257-259 in page 15).

- *p. 18, line 1: the "five models" are presumably models a-e? If so, say "of models a-e"*

Corrected (line 310, page 17)

- *p. 22, line 2: from the figure, I would have guessed "greater than or equal to 0.7", not 0.8*

Based the Figure 3C, we think 0.8 is a more appropriate cutoff, since the precision drops significantly when the DTM-score is within 0.7~0.8.

- *p. 23, lines 2-3: "worth to making" -> "worth making"*

Corrected (line 443, page 24)

- *p. 24, line -5: "predict" -> "predicted"*

Corrected (line 484, page 26)

- *p 28, line -5: Please clarify what you mean by "We doubt": are you saying that you don't think these rearrangements exist in nature? If not, then reword.*

Corrected (line 566, page 30)

- *Figure 2, panel c, "DCPred" in the legend should be "CDPred"*

Corrected

- *Figures 3 and 5: Please improve the y-axis title in panel C. "Percent" of what?*

We changed the “Percent” to “% of targets” in the revision.

We thank the reviewer for carefully reading our manuscript!

Reviewer #2 (Public Review):

This work introduces PLMGraph-Inter, a new deep-learning approach for predicting inter-protein contacts, which is crucial for understanding proteinprotein interactions. Despite advancements in this field, especially driven by AlphaFold, prediction accuracy and efficiency in terms of computational cost) still remains an area for improvement. PLMGraph-Inter utilizes invariant geometric graphs to integrate the features from multiple protein language models into the structural information of each subunit. When compared against other inter-protein contact prediction methods, PLMGraph-Inter shows better performance which indicates that utilizing both sequence embeddings and

> *structural embeddings is important to achieve high-accuracy predictions with relatively smaller computational costs for the model training.*
>
> *The conclusions of this paper are mostly well supported by data, but test examples should be revisited with a more strict sequence identity cutoff to avoid any potential information leakage from the training data. The main figures should be improved to make them easier to understand.*

We thank the reviewer for recognizing the significance of our work! We have carefully revised the manuscript to address the reviewer's concerns.

> *(1) The sequence identity cutoff to remove redundancies between training and test set was set to 40%, which is a bit high to remove test examples having homology to training examples. For example, CDPred uses a sequence identity cutoff of 30% to strictly remove redundancies between training and test set examples. To make their results more solid, the authors should have curated test examples with lower sequence identity cutoffs, or have provided the performance changes against sequence identities to the closest training examples.*

We thank the reviewer for the valuable suggestion! The "40 sequence identity" is a widely used threshold to remove redundancy when evaluating deep-learning based protein-protein interaction and protein complex structure prediction methods, thus we also chose this threshold in our study (bioRxiv 2021.10.04.463034, Cell Syst. 2021 Oct 20;12(10):969-982.e6). In the revision, we explored whether PLMGraph-inter can keep its performance when more stringent thresholds (30%,20%,10%) is applied (line 353386 in page 20-21 in the subsection of "Ablation study" and line 762-780 in page 40 in the subsection of "Further potential redundancies removal between the training and the test"). The result shows that even when using "10% sequence identity" as the threshold, mean precisions of the predicted contacts only decreases by ~3% (Table 2).

> *(2) Figures with head-to-head comparison scatter plots are hard to understand as scatter plots because too many different methods are abstracted into a single plot with multiple colors. It would be better to provide individual head-tohead scatter plots as supplementary figures, not in the main figure.*

We thank the reviewer for the suggestion! We will include the individual head-to-head scatter plots as supplementary figures in the revision (Figure S1 and Figure S2 in the supplementary).

> *(3) The authors claim that PLMGraph-Inter is complementary to AlphaFoldmultimer as it shows better precision for the cases where AlphaFold-multimer fails. To strengthen the point, the qualities of predicted complex structures via protein-protein docking with predicted contacts as restraints should have been compared to those of AlphaFold-multimer structures.*

We thank the reviewer for the suggestion! We included this comparison in the revision (Figure S7).

> *(4) It would be interesting to further analyze whether there is a difference in prediction performance depending on the depth of multiple sequence alignment or the type of complex (antigen-antibody, enzyme-substrates, single species PPI, multiple species PPI, etc).*

We thank the reviewer for the suggestion! We analyzed the relationship between the prediction performance and the depth of MSA in the revision (Figure S4 and Line 253264 in page 15 in the subsection of “Evaluation of PLMGraph-Inter on HomoPDB and HeteroPDB test sets” and line 798-806 in page 42 in the subsection of “Calculating the normalized number of the effective sequences of paired MSA”).

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *I have the following suggestions in addition to the public review.*
>
> *(1) Overall, the manuscript is well-written; however, I recommend a careful review for minor grammar corrections to polish the final text.*

We carefully checked the manuscript and corrected all the grammar issues and typos we found in the revision.

> *(2) It would be better to indicate that single sequence embeddings, MSA embeddings, and structure embeddings are ESM-1b, ESM-MSA & PSSM, and ESM-IF when they are first mentioned in the manuscript e.g. single sequence embeddings from ESM-1b, MSA embeddings from ESM-MSA and PSSM, and structural embeddings from ESM-IF.*

We revised the manuscript according to the reviewer’s suggestion (line 86-88 in page 6; line 99-101 in page 7).

> *(3) I don't think "outer concatenation" is commonly used. Please specify whether it's outer sum, outer product, or horizontal & vertical tiling followed by concatenation.*

It is horizontal & vertical tiling followed by concatenation. We clarified this in the revision (line 129-130 in page 8).

> *(4) 10th sentence on the page where the Results section starts, please briefly mention what are the other 2D pairwise features.*

We clarified this in the revision (line 131-132 in page 8).

> *(5) In the result section, it states edges are defined based on Ca distances, but in the method section, it says edges are determined based on heavy atom distances. Please correct one of them.*

It should be Ca distances. We are sorry for the carelessness, and we corrected this in the revision (line 646 in page 35).

> *(6) For the sentence, "Where ESM-1b and ESM-MSA-1b are pretrained PLMs learned from large datasets of sequences and MSAs respectively without label supervision,", I'd suggest replacing "without label supervision" with "with masked language modeling tasks" for clarity.*

We revised the manuscript according to the reviewer’s suggestion (line 150-151 in page 9).

> *(7) It would be better to briefly explain what is the dimensional hybrid residual block when it first mentioned.*

We explained the dimensional hybrid residue block when it first mentioned in the revision (line 107 in page 7).

> *(8) Please include error bars for the bar plots and standard deviations for the tables.*

We thank the reviewer for the suggestion! Our understanding is the error bars and standard deviations are very informative for data which follow gaussian-like distributions, but our data (precisions of the predicted contacts) are obviously not this type. Most previous studies in protein contact prediction and inter-protein contact prediction also did not include these in their plots or tables. In our case, including these elements requires a dramatic change of the styles of our figures and tables, but we would like to not change our figures and tables too much in the revision.

> *(9) Please indicate whether the chain break is considered to generate attention map features from ESM-MSA-1b. If it's considered, please specify how.*

The paired sequences were directly concatenated without using any letter to connect them, which means we did not consider chain break in generating the attention maps from ESM-MSA-1b.