Integrating structural homology with deep learning to achieve highly accurate protein-protein interface prediction for the human interactome

Dapeng Xiong[1,2,3#], Mateo Torres[1,2,3#], Diana Murray[4], Le Li[1,2,3], Aniket C. Naravane[4], Robert Fragoza[1,2,3], Barry Honig[4,5,6,7*], Haiyuan Yu[1,2,3*]

Affiliations:

[1]Department of Computational Biology, Cornell University, Ithaca, NY 14853, USA

[2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA

[3]Center for Innovative Proteomics, Cornell University, Ithaca, NY 14853, USA

[4]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032, USA

[5]Department of Biochemistry and Molecular Biophysics, Columbia University Irving Medical Center, New York, NY 10032

[6]Department of Medicine, Columbia University, New York, NY 10032

[7]Zuckerman Institute, Columbia University, New York, NY 10027

[#]These authors contributed equally to this work

*Corresponding authors: bh6@cumc.columbia.edu (B.H.) and haiyuan.yu@cornell.edu (H.Y.)

Abstract

A significant portion of disease-causing mutations occur at protein-protein interfaces however, the number of structurally resolved multi-protein complexes is extremely small. Here we present a computational pipeline, PIONEER2.0, that integrates 3D structural similarity with geometric deep learning to accurately predict protein binding partner-specific interfacial residues for all experimentally observed human binary protein-protein interactions. We estimate that AlphaFold3 fails to produce high-quality structural models for about half of the human interactome; for these challenging cases, PIONEER2.0 significantly outperforms AlphaFold3 in predicting their interface residues, making PIONEER2.0 an excellent alternative and complementary tool in real-world applications. We further systematically validated PIONEER2.0 predictions experimentally by generating 1,866 mutations and testing their impact on 5,010 mutation-interaction pairs, confirming PIONEER-predicted interfaces are comparable in accuracy as experimentally determined interfaces using PDB co-complex structures. We then used PIONEER2.0 to create a comprehensive multiscale structurally informed human interactome encompassing all 352,124 experimentally determined binary human protein interactions in the literature. We find that PIONEER2.0-predicted interfaces are instrumental in prioritizing disease-associated mutations and thus provide insight into their underlying molecular mechanisms. Overall, our PIONEER2.0 framework offers researchers a valuable tool at an unprecedented scale for studying disease etiology and advancing personalized medicine.

Introduction

Most proteins function through interactions with other proteins to form complexes that carry out a myriad of biological tasks (*1, 2*). In the past decade, significant effort has been invested into building a comprehensive interactome network for human proteins (*3-9*). Many proteins are pleiotropic and carry out diverse functions through interactions with different proteins using distinct interaction interfaces (*10*). Also, mutations on the same protein can affect different interactions to cause clinically distinct diseases, depending on the interface where the mutation occurs (*11*). Therefore, it is of great importance to accurately determine the specific interface that mediates each interaction. However, only a small fraction (3.2%) of known human protein interactions have experimentally resolved surfaces (Supplementary Fig. 1).

Over the last years, a number of computational methods have demonstrated exceptional capabilities in representation learning, particularly for processing 3D protein structures, leading to much improved performance for protein-protein interface residue prediction (*12-15*). Among these is our recently developed ensemble deep learning pipeline, PIONEER1.0 (*15*), designed to generate partner-specific interaction interface residue predictions for experimentally determined protein-protein interactions. In contrast, PrePPI (*16*) uses structural similarity to predict whether two proteins interact by using an efficient scoring function which focuses on interfacial residues (*17*). The complementary nature of these two approaches underlines the potential of an integrated framework that combines both strategies. Of note, AlphaFold-based methods (*18-20*) have made significant advances in predicting both single and multimeric protein structural models, providing crucial functional insights for proteins. However, these methods are not yet scalable to model entire interactomes. More importantly, in real-world applications, it has been shown that AlphaFold-based methods can generate high-quality structural models for only a small fraction of protein-protein interactions without known structures (*21*).

Here, we present PIONEER2.0, an equivariant geometric deep learning pipeline that integrates a comprehensive set of biophysical, evolutionary, structural, and sequence features, together with homologous interface information from structural homologs detected by PrePPI, to generate precise partner-specific protein-protein interface residue predictions with high accuracy. Extensive benchmark tests show that PIONEER2.0 is

significantly better than PIONEER1.0 and other similar algorithms (*12-14*). Even compared with the latest AlphaFold3 (*20*) method, PIONEER2.0 achieves comparable overall performance and significantly outperforms it for cass where AF3 produced low confidence structures. We estimate below that these correspond to nearly half of all human protein-protein interactions without known structures. Together with the available atomic resolution co-crystal structures, we established a comprehensive multiscale structurally informed human protein-protein interactome consisting of 352,124 experimentally-detected binary interactions. We then validated PIONEER2.0 predictions through large-scale mutagenesis experiments. Furthermore, we found significant enrichment of disease mutations at specific PIONEER2.0-predicted interaction interfaces, indicating that PIONEER2.0 reveals crucial structural information in delineating disease mechanisms. Our work provides researchers with a powerful tool to accurately predict partner-specific protein-protein interfaces, thereby helping accelerate structural and functional understanding of proteins and complexes.

Results

An equivariant geometric architecture integrating atomic and residue levels of information

PIONEER2.0 uses a different design than PIONEER1,0 and, specifically, is based on an equivariant geometric deep learning architecture designed to incorporate a comprehensive set of features for predicting protein-protein interfaces. Using features from both interacting partners, PIONEER2.0 effectively combines biophysical, evolutionary, structural, and sequence information for accurate characterization of interfaces. Importantly, PIONEER2.0 integrates structural information from 3D spatially neighboring residues.

The architecture is based on the concept of extracting meaningful characteristics from both the target and partner proteins, and then combining these representations to classify whether each residue is part of the interface for the specific interaction. To achieve this, PIONEER2.0 implements two representation learning branches for sequence and 3D structural information, respectively (Fig. 1A). The sequence branch

learns sequence embeddings through a transformer encoder, and a CNN encoder. The transformer is appropriate for capturing long-range information, while the CNN is used for local information. Each residue is annotated with biophysical properties, conservation, co-evolution, solvent-accessible surface area (SASA), secondary structure, residue depth, and docking features for an in-depth feature characterization. Compared with our previous models (e.g., ECLAIR (*22*) and PIONEER1.0 (*15*)), a key new set of features are derived from the interface information of homologous co-complex structures in PDB (*17*) detected by PrePPI (*16*). As shown in Fig. 1B, for a given query of two interacting proteins, PrePPI searches the PDB database with a fast 3D structure alignment algorithm to identify template complexes whose chains are structurally similar (i.e., structural homologs) to the query proteins. PrePPI then scores the structure-based sequence alignment in the PDB template interface based on interface coverage and the interface propensity of the query residues (see Methods). Eight different PrePPI scores are assigned to each aligned residue of the query proteins as features in the sequence branch (Fig. 1A). It should be noted that the structure alignment by PrePPI is based solely on protein topology facilitating the detection of remote homologs, which is further enhanced by considering both the full-length query protein and its constituent domains, ensuring great coverage and utility of these PrePPI-derived features.

In parallel to the sequence branch, the structural branch builds a geometric representation of each protein's 3D structure at both atomic and residue levels. Specifically, atoms and residues are represented in an equivariant manner to preserve geometric properties under rotation and translation. A key step in this process is the integration of information from spatially neighboring atoms and residues—those that are physically close in 3D space, even if they are distant in sequence—allowing the model to capture local environment and long-range dependencies among atoms and among residues. These representations are embedded using a Gaussian mixture model (*13*), which captures local spatial distributions by modeling the positions of atoms and residues as a weighted combination of Gaussian components. This enables the model to learn meaningful spatial patterns in the protein 3D structures. Atomic level features are further refined through self-attention pooling and concatenated with residue-level embeddings to form comprehensive structural representations. Every node is annotated with the same

features as those in the sequence branch. Additionally, in the structure branch, each edge is annotated with Euclidean distance, angle, and coevolution information. Embeddings are finally processed using a graph transformer encoder.

In both sequence and structural branches, target and partner proteins are represented by a feature matrix, and the partner is further processed by self-attention and concatenated with the target protein. Finally, the embeddings from both sequence and structural branches are integrated into a dense layer, culminating in the final prediction.

Accurate prediction of protein-protein interfaces enabled by structural homologs

To comprehensively evaluate the performance of PIONEER2.0, we carefully constructed large labeled datasets with interface information from experimentally determined co-complex structures in PDB (see Methods) for model training, validation, and testing. Compared with PIONEER1.0 labeled sets, we especially prioritize the instances where the same protein interacts with multiple interaction partners using distinct interfaces in our labeled dataset to build a model that better predicts partner-specific interfaces. Furthermore, we ensured that no homologous complexes appear across labeled sets so as to maximize the generalization of our pipeline and to avoid data leakage. In total, the labeled sets contain 4,649, 776 and 802 protein-protein interactions for model training, validation and testing, respectively.

As described above, one key addition in PIONEER2.0 is binding interface information from structural homologous PDB complexes detected by PrePPI. To mitigate data leakage, for each interaction in the labeled sets, we removed PDB complexes with high sequence homology to both query proteins from the pool of template complexes interrogated by PrePPI. As shown in Fig. 1C, predictive performance on the test set directly using the aligned interface information from the top structural homologs detected by PrePPI is already quite good. Comprehensively incorporating all other features and aggregating neighboring information at both sequence and 3D structural levels in the full PIONEER2.0 pipeline further significantly improve the performance (Fig. 1C).

Since we have previously shown that PIONEER1.0 (*15*) outperforms recent state-of-the-art interface prediction methods (such as PeSTo (*14*), ScanNet (*13*), MaSIF-Site

(*12*), and others), we first compared the performance of PIONEER2.0 with PIONEER1.0 and found that PIONEER2.0 has significantly improved its predictive performance (Fig. 1D; AUROC increased from 0.693 to 0.854). AlphaFold2 (*18*) was a breakthrough development in structural modeling for individual proteins. Many AlphaFold-based methods (*21, 23*) have been developed to model protein-protein interactions. AlphaFold3 (*20*) is the latest model with state-of-the-art performance for modeling not only protein-protein interactions, but also protein-DNA and -small-molecule interactions. We further evaluated PIONEER2.0 by comparing it with the AlphaFold3, using the contact probabilities provided by AlphaFold3 on a per-residue basis (Fig. 1D), and found that PIONEER2.0 achieves comparable overall performance (AUROC = 0.854, 0.856 for PIONEER2.0 and AlphaFold3, respectively).

PIONEER2.0 significantly outperforms AlphaFold3 in interface residue prediction for nearly half of the human interactome

There are hundreds of thousands of experimentally determined binary protein-protein interactions (i.e., the human interactome; see Methods) for human proteins already published in the literature, only ~3.2% of which have structural models in PDB (Supplementary Fig. 1). It is thus of great interest to predict the binding interface information for the ~96.8% of the human interactome without any known structures (*24*).

Since AlphaFold3 has become the *de-facto* method for computationally predicted structural models of protein interactions, we carefully analyzed its performance on interactions of known and unknown structures. We generated AlphaFold3 models for the test set of PDB complexes (Fig. 1D) and for 1,024 randomly chosen protein-protein interactions in the literature of unknown structures. It is critical to make sure that these 1,204 interactions are of high quality; they are chosen from our high-quality literature-curated binary interaction dataset from low-throughput experiments (“HQ-LC-Binary”) in our HINT database (*24*), which integrates information from commonly used databases, including BioGRID (*25*), DIP (*26*), IntAct (*27*), MINT (*28*), iRefWeb (*29*), HPRD (*30*), PDB (*17*), and MIPS (*31*) by examining the evidence codes and extent of literature curation for each interaction (see Methods). The HQ-LC-Binary dataset in HINT contains only interactions with at least two separate low-throughput publications with evidence codes

confirming direct, physical binding of the interactions. Therefore, the HINT-HQ-LC-Binary dataset contains protein-protein interactions of very high quality (*32-34*).

We used the AlphaFold3 ranking score (*20*) as the confidence metric for their models. A higher ranking score indicates a greater likelihood that the complex structure is accurately predicted; ranking scores less than 0 were truncated to 0 for this analysis. The ranking score distributions in Fig. 2A, B illustrate that, overall, AlphaFold3 models for interactions of unknown structures are of lower confidence than models for interactions with PDB structures: 48% of models for interactions of unknown structures have ranking scores ≤ 0.5 whereas only 23% of models for PDB interactions score that poorly. This suggests that complementary approaches, such as PIONEER2.0, will be extremely valuable in providing additional confidence metrics for AlphaFold3 models; and, more importantly, in making novel predictions for PPI interfaces that are not structurally resolved.

Because 96.8% of the human interactome does not have known structures in PDB (Sup Fig. 1), to more fairly evaluate PIONEER2.0's performance, we took a bootstrap resampling strategy to generate the same distribution of AlphaFold3 ranking scores as the 1,024 interactions without known structures by taking random samples from the test set. Such a resampling strategy is necessary because we can only evaluate the quality of AlphaFold3 predicted structural models by comparing them to available experimental structures in PDB. As shown in Fig. 2C, in this scenario, PIONEER2.0 (AUROC = 0.81) significantly outperforms AlphaFold3 (AUROC = 0.77) in real-world applications on interactions without known structures. Furthermore, we focused on the more challenging interactions whose AlphaFold3 ranking scores are below 0.5, which accounts for nearly half (48.05%) of all human interactions without known structures (Fig. 2B). Our analysis reveals that PIONEER2.0 significantly outperforms AlphaFold3 in interface predictions for these challenging interactions, as evaluated by F1, Precision, Recall, and MCC scores (Fig. 2D).

We then conducted a detailed performance comparison between PIONEER2.0 and AlphaFold3 across different AlphaFold3 ranking score bins without applying the bootstrap resampling strategy, aiming to further demonstrate the strong capability of PIONEER2.0 for protein-protein interface prediction. The results reveal that PIONEER2

consistently and significantly outperforms AlphaFold3 in all bins where the ranking score is below 0.5 (Fig. 3 and Supplementary Fig. 2), and achieves comparable performance in the bins with higher AlphaFold3 ranking scores (Supplementary Fig. 2).

One key distinction between PIONEER2.0 and AlphaFold3 is that PIONEER2.0 focuses exclusively on predicting interface residues, while AlphaFold3 is designed to predict co-complex structures, which is a much more difficult task. Therefore, even minor changes in the relative positions of the two proteins in a structural model can have a significant impact on the accuracy of interface residue predictions by AlphaFold3. Here, we show two examples (Figs. 3E, F) where AlphaFold3 builds good models for the monomers but places them in different orientations relative to the PDB complexes, leading to misidentification of interface residues even when the contact probabilities may be high. Interface residues predicted by PIONEER2.0 are labeled as green (correct prediction) and red (incorrect) spheres. It is interesting to note that even the technically incorrect predictions (red spheres) by PIONEER2.0 are in the binding interface area, highlighting the robust performance of PIONEER2.0, even for cases when AlphaFold3 struggles.

These comparisons confirm PIONEER2.0's robust predictive performance, particularly for interactions where AlphaFold3-generated models are less reliable. Notably, these low-scoring bins represent a substantial portion of the human interactome without known structures, which are precisely the challenging cases where accurate interface prediction is most needed. In particular, 48.05% of all human interactions without known structures are expected to have AlphaFold3 models with ranking score below 0.5; for these interactions with low-quality AlphaFold3 models (46.51% of the entire human interactome), PIONEER2.0 will provide better quality interface predictions. Finally, the ability of PIONEER2.0 to maintain strong performance across all interactions with both low and high confidence AlphaFold3 models underscores its generalizability and practical applicability.

Construction and validation of the structurally informed human interactome

We compiled a comprehensive set of experimentally validated protein-protein interactions for human from our HINT database, which integrates information from eight commonly used databases as described above. We carefully examine the evidence codes of each interaction to select only those with experimental evidence to be direct binding, because otherwise, the concept of predicting interface residues does not even apply. This resulted in 352,124 binary human protein-protein interactions, 340,860 of which have no available structures in PDB. We then used the fully optimized PIONEER2.0 to predict interfaces for all 340,860 interactions. Since we made the partner-specific interface prediction for every residue in a protein pair, we made predictions for more than 397 million residue-protein interaction pairs. By combining PIONEER2.0 interface predictions with 11,264 interactions with experimental structures in PDB, we generated a comprehensive 3D structural human interactome (Fig. 4A), in which all 352,124 interactions have partner-specific interface information at the residue level, together with atomic resolution 3D models for those appear in PDB.

We conducted a comprehensive evaluation of the quality of our predicted interface residues and their biological implications by performing large-scale mutagenesis experiments. These experiments measured the fraction of disrupted interactions caused by mutations to our predicted interfacial and non-interfacial residues. We also measured the effects of mutations to known interfacial and non-interfacial residues in PDB co-complex structures. Using our Clone-seq pipeline (*35*), we generated 1,866 mutations across 895 proteins and assessed their impact on 5,010 mutation interaction pairs via a high-throughput yeast-two-hybrid (Y2H) assay. As shown in Fig. 4B, our results demonstrate that mutations of PIONEER2.0-predicted interfacial and non-interfacial residues disrupt protein-protein interactions at a rate very similar to that observed in PDB complexes, with both showing significantly higher disruption rates for interfacial versus non-interfacial residues. These large-scale experiments confirm the high quality of our interface predictions and validate the overall effectiveness of our PIONEER2.0 pipeline. Given that interaction disruption is crucial for understanding the molecular mechanisms of disease mutations (*35, 36*), our findings suggest that PIONEER2.0-predicted interfaces can play a key role in prioritizing disease-associated variants and forming experimentally testable hypotheses of functional mechanisms.

We examined the distribution of population genetic variants and found that their enrichment in PIONEER2.0-predicted interfaces and non-interfaces closely aligns with the distribution observed in known interfaces and non-interfaces, respectively (Fig. 4C). The results also reveal a depletion of common, non-deleterious variants in both known and predicted interfaces, indicating that PIONEER2.0 effectively predicts functionally significant interface residues. We also measured the enrichment of known disease-associated mutations from the Human Gene Mutation Database (*37*) (HGMD) in PIONEER2.0-predicted interface residues, and found it to be significantly higher than that of predicted non-interface residues, similar to enrichment of disease mutations in known interfaces of protein co-complexes in PDB (Fig. 4D). Overall, our results demonstrate a strong relationship between PIONEER2.0 prediction scores and key residues essential for protein functions.

Discussion

In this study, we introduce PIONEER2.0, an equivariant geometric deep learning framework designed for highly accurate protein-protein interface residue prediction at the whole interactome scale by integrating interface information from structural homologs with a comprehensive set of biophysiochemical, evolutionary, and structural features. Our analysis reveals that AlphaFold-based methods tend to build lower quality structural models for nearly half of the human interactome of unknown structures compared to those that appear in PDB; for these challenging cases, PIONEER2.0 outperforms AlphaFold3 in interface residue prediction. Our results thus establish PIONEER2.0 as a powerful tool for large-scale interface mapping across entire interactomes.

We applied PIONEER2.0 to resolve the interface information for all 340,860 human protein-protein interactions without experimental structures in the literature, which, together with 11,264 interactions with PDB co-complex structures, yields a comprehensive structurally informed human interactome. Using large-scale mutagenesis Y2H experiments, we extensively tested our predictions and found that PIONEER2.0-predicted interface residues are validated at the same rate as PDB interface residues. Furthermore, our analyses show that PIONEER2.0-predicted interface residues have significant enrichment of rare alleles (i.e., functionally important sites) and known disease

mutations, to the same degrees as those of PDB interfaces and non-interfaces, confirming the strong biological and functional relevance of PIONEER2.0 predictions. In summary, our PIONEER2.0-generated structurally informed human interactome will be an invaluable resource for the study of disease mechanisms and development of personalized treatments.

It is important to recognize that PIONEER2.0 is focused on identifying interface residues for each interaction, while AlphaFold3 is designed to predict the full co-complex structure of interacting proteins, which is a substantially more challenging task. Although it is important to know the binding interface of an interaction (as PIONEER2.0 does), having a structural model of the whole co-complex provides even more insight. By integrating AlphaFold3 models with PIONEER2.0 predictions, we can better understand functional implications of a specific protein-protein interaction or its disruptions by mutations, and better design potential therapeutics to target this interaction. In fact, as described above, many of the AlphaFold3 errors likely involve inter-protein orientations rather than the structures of individual monomers. Thus, for low-confidence models, PIONEER2.0 can be used to assess the quality of the models, and in addition, can potentially be used as basis for algorithms that further refine or correct protein orientations.

While PIONEER2.0's performance is significantly enhanced by the inclusion of structural homolog information and the equivariant geometry architecture, we expect that its capability can be further boosted by implementing even more advanced architectures that will be developed in the future, and by incorporating novel features, such as sequence language models (e.g., ESM2 (*38*)), and structural language models (e.g., Saprot (*39*)). With the ongoing progress in sequencing technologies and large-scale genome/exome sequencing projects, such as those from TCGA (*40*) and precision medicine initiatives, we envision that PIONEER2.0's structurally informed interactome will bridge the gap between genomic data and structural proteomics.

Methods

Dataset construction

In our HINT database, we compiled a comprehensive set of experimentally validated binary protein-protein interactions for human, by integrating information from commonly used databases, including BioGRID (*25*), DIP (*26*), IntAct (*27*), MINT (*28*), iRefWeb (*29*), HPRD (*30*), PDB (*17*), and MIPS (*31*). We carefully examine the evidence codes of each interaction to select only those with experimental evidence to be direct binding. This resulted in 352,124 binary human protein-protein interactions, 340,860 of which have no experimental structures; while the other 11,264 have experimentally determined co-complex structures in PDB. Among these interactions, we carefully selected a subset (named "HINT-HQ-LC-Binary") of 44,115 high-quality literature-curated protein-protein interactions that are required to be validate by two separate publications using low-throughput experimental methods.

To construct our labeled sets, partner-specific interface residues were identified from the known co-crystal structures available in PDB. SIFTS (*41*) was then used to map the UniProt indices to the PDB indices. The interface residues are determined as residues that are surface residues (≥15% exposed surface) with its relative SASA decreasing by ≥1Å^2 in the complex. Surface areas were calculated using NACCESS (*42*). We processed all available structures in PDB for each interaction, and considered a residue to belong to the interface if it matches our criterion in at least one of the processed co-crystal structures. Our dataset includes only interactions for which at least 30% of the UniProt residues are covered by co-crystal structures. In addition, to ensure the robustness and generalizability of our models, as well as a fair performance evaluation, we prevented any homologous interactions from being shared between any two datasets that are from the training, validation, and testing splits. Specially, we defined a pair of interactions to be homologous if both proteins in one interaction are homologs to both proteins in the other interaction. We performed three iterations of PSI-BLAST (*43*) at an E-value cutoff of 0.001 to determine if two proteins are homologous. This resulted in 4,649, 776, and 802 interactions, corresponding to 2,441,383, 357,276, 326,868 labeled residues for sufficient model training, validation, and testing, respectively (Supplementary Data).

Graph representation

For a given protein, we created a graph for each residue to enhance structural representations. Each residue is represented by its Cα atom, and the graph includes the 16 closest neighbors to the target residue as nodes with their interactions as edges. Each node and edge are represented by a comprehensive set of features. In specific, in addition to the features used in our previous pipeline, PIONEER and ECLAIR (*22*), which describe each residue in terms of biophysical residue properties, conservation, coevolution, relative SASA, and secondary structure, we have now incorporated three new feature groups for the node representation. These include residue depth, docking metrics, and PrePPI metrics, offering a more comprehensive feature characterization.

We used MSMS (*44*) to calculate the residue depth, which describes the average distance of the atoms of a residue from the solvent accessible surface. EquiDock (*45*) was used for the rigid body protein-protein docking, after which the interface residues were calculated using the above pipeline. PrePPI (*16*) is a pipeline for proteome-wide prediction of known and novel PPIs; PrePPI models and results are available for examination and download from our website. In specific, pairs of proteins, represented by models from the AlphaFold Structure Database (*46*), are screened for structural neighbors against a large pool of structures from the PDB database with a fast structure alignment method (*47*). For a given pair of query proteins, a suitable interaction template contains structural neighbors of both queries, and the structure-based sequence alignment is examined. The model for the putative PPI is calibrated against the PDB template complex interface, providing several scores that are incorporated in a Bayesian framework to provide a structural modeling likelihood ratio that the query proteins interact The PrePPI feature group includes 8 scores, which are structural modeling likelihood ratio for the overall model, structural similarity between the query protein models and their respective template chain, the number of residue pairs in the template, the number of interacting residue pairs in the template that are preserved in the model, the fraction of interacting residue pairs in the template that are preserved in the model, the number of predicted interfacial residues in the query models that align to interfacial residues in their respective template chains, and the number of interacting residue pairs in the template

that are preserved in the model and are predicted to be interfacial residues in the query protein with the program PredUs (*48*), respectively.

For the edge, the feature group includes the Euclidean distance, angle and the coevolution between the interacted node pair.

Model building

To ensure the in-depth capture of structural information, we incorporated the equivariant geometric information at both atomic and residue levels before the processing of the graph Transformer in the structure branch of PIONEER2.0 architecture. The geometric representation is referred to the literature (*13*) to make sure the equivariant properties under rotation and translation. The geometric representations are embedded through Gaussian mixture model at both atomic and residue levels, respectively, and are subsequently concatenated via the self-attention pooling at the amino acid scale. The concatenation is then integrated with the node features along with the edge features for the subsequent graph Transformer embedding.

We compiled a set of representative protein structures from PDB and AlphaFold database for each protein. To enhance the practical applicability of our method more and prevent the memorization of known interfaces, we trained the model using the single protein structures that are not derived from co-crystal or homologous co-complex structures. PDB structures were given the highest priority, while AlphaFold structures were used as a secondary option. The structures are then ranked based on UniProt residue coverage as determined by SIFTS, while ensuring that homologous PDB structures of interacting protein pairs were excluded. For each target protein, residue-level predictions were made using the first corresponding structure that contains the structural information of that residue. For the partner protein, we selected only the structure with the highest UniProt coverage. The PIONEER2.0 framework was implemented using PyTorch (https://pytorch.org), with the graph Transformer built on the Deep Graph Library (https://www.dgl.ai). The model was trained using binary cross-entropy loss and the Adam optimizer. To accommodate the variable length inputs, we trained the model in a mini-batch mode, processing a single protein pair per batch.

Mutagenesis validation experiments

We conducted mutagenesis experiments by introducing random human population variants from the gnomAD database (*49*) into predicted interfaces, known interfaces and non-interfaces. Variants located in predicted interfaces were selected based on PIONEER2.0 results, while those in known interfaces and non-interfaces from co-crystal structures in PDB served as positive and negative controls, respectively. All selected mutations were introduced using our Clone-seq pipeline (*35*). In total, we generated 1,866 mutations across 895 proteins and assessed their effects on 5,010 mutation interaction pairs—either disrupting or maintaining the interactions—using our high-throughput Y2H assay.

Y2H assay

We conducted the Y2H assays using the pipeline as previously described (*15, 22*). We used Gateway LR reactions to transfer all WT/mutant clones into our Y2H pDEST-AD and pDEST-DB vectors. All DB-X and AD-Y plasmids were transformed into the Y2H strains MATα Y8930 and MATa Y8800, respectively. Subsequently, each DB-X MATα transformant (WT and mutants) was individually mated with its corresponding AD-Y MATa transformant (WT and mutants) using automated 96-well procedures, including inoculation of AD-Y and DB-X yeast cultures, mating on YEPD media (incubated overnight at 30 °C) and replica plating onto selective Synthetic Complete media lacking histidine, leucine and tryptophan and supplemented with 1 mM 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT), SC-Leu-His+3AT plates containing 1 mg L^{-1} cycloheximide (SC-Leu-His+3AT+CHX), SC-Leu-Trp-Adenine (Ade) plates and SC-Leu-Ade+CHX plates to test for CHX-sensitive expression of the LYS2::GAL1–HIS3 and GAL2–ADE2 reporter genes. We identified spontaneous auto-activators (*50*) by growth on plates containing cycloheximide. These plates were incubated overnight at 30 °C and subjected to replica cleaning the following day. The plates were then incubated for an additional three days, after which positive colonies were scored as those growing on SC-Leu-Trp-His+3AT and/or on SC-Leu-Trp-Ade, but not on SC-Leu-His+3AT+CHX or on SC-Leu-Ade+CHX. An interaction was considered disrupted by a mutation if it resulted in a consistent reduction of at least 50% in growth across both reporter genes, relative to the Y2H

phenotypes of the corresponding WT allele, as benchmarked by two-fold serial dilution experiments. All Y2H assays were independently repeated three times.

References

1. A. L. Barabasi, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56-68 (2011).
2. M. Vidal, M. E. Cusick, A. L. Barabasi, Interactome networks and human disease. *Cell* **144**, 986-998 (2011).
3. K. Luck *et al.*, A reference map of the human binary protein interactome. *Nature* **580**, 402-408 (2020).
4. E. L. Huttlin *et al.*, Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-509 (2017).
5. T. Rolland *et al.*, A proteome-scale map of the human interactome network. *Cell* **159**, 1212-1226 (2014).
6. K. Venkatesan *et al.*, An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83-90 (2009).
7. U. Stelzl *et al.*, A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-968 (2005).
8. J. F. Rual *et al.*, Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).
9. Q. C. Zhang *et al.*, Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* **490**, 556-560 (2012).
10. T. Pawson, P. Nash, Protein-protein interactions define specificity in signal transduction. *Genes Dev.* **14**, 1027-1047 (2000).
11. X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin, H. Yu, Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159-164 (2012).
12. P. Gainza *et al.*, Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184-192 (2020).

13. J. Tubiana, D. Schneidman-Duhovny, H. J. Wolfson, ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods* **19**, 730-739 (2022).
14. L. F. Krapp, L. A. Abriata, F. Cortes Rodriguez, M. Dal Peraro, PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat. Commun.* **14**, 2175 (2023).
15. D. Xiong *et al.*, A structurally informed human protein–protein interactome reveals proteome-wide perturbations caused by disease mutations. *Nat. Biotechnol.*, (2024).
16. D. Petrey, H. Zhao, S. J. Trudeau, D. Murray, B. Honig, PrePPI: A Structure Informed Proteome-wide Database of Protein–Protein Interactions. *J. Mol. Biol.* **435**, 168052 (2023).
17. Stephen K. Burley *et al.*, Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Res.* **53**, D564-D574 (2025).
18. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
19. R. Evans *et al.*, Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034 (2022).
20. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
21. D. F. Burke *et al.*, Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* **30**, 216-225 (2023).
22. M. J. Meyer *et al.*, Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods* **15**, 107-114 (2018).
23. M. Gao, D. Nakajima An, J. M. Parks, J. Skolnick, AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022).
24. J. Das, H. Yu, HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).

25. R. Oughtred *et al.*, The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529-D541 (2019).
26. L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, D. Eisenberg, The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449-D451 (2004).
27. S. Orchard *et al.*, The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358-D363 (2014).
28. L. Licata *et al.*, MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857-D861 (2012).
29. B. Turner *et al.*, iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* **2010**, baq023 (2010).
30. T. S. Keshava Prasad *et al.*, Human Protein Reference Database—2009 update. *Nucleic Acids Res.* **37**, D767-D772 (2009).
31. H. W. Mewes *et al.*, MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.* **39**, D220-D224 (2011).
32. M. E. Cusick *et al.*, Literature-curated protein interaction datasets. *Nat. Methods* **6**, 39-46 (2009).
33. M. Taşan, G. Musso, T. Hao, M. Vidal, C. A. MacRae, F. P. Roth, Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nat. Methods* **12**, 154-159 (2015).
34. H. Zhao, D. Petrey, D. Murray, B. Honig, ZEPPI: Proteome-scale sequence-based evaluation of protein–protein interaction models. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2400260121 (2024).
35. X. Wei *et al.*, A Massively Parallel Pipeline to Clone DNA Variants and Examine Molecular Phenotypes of Human Disease Mutations. *PLoS Genet.* **10**, e1004819 (2014).
36. N. Sahni *et al.*, Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell* **161**, 647-660 (2015).
37. P. D. Stenson *et al.*, The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic

diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665-677 (2017).

38. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
39. J. Su, C. Han, Y. Zhou, J. Shan, X. Zhou, F. Yuan, paper presented at the The Twelfth International Conference on Learning Representations, 2024.
40. K. Chang *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113-1120 (2013).
41. S. Velankar *et al.*, SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41**, D483-D489 (2013).
42. B. Lee, F. M. Richards, The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400 (1971).
43. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
44. M. F. Sanner, A. J. Olson, J.-C. Spehner, Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **38**, 305-320 (1996).
45. O.-E. Ganea *et al.*, paper presented at the The Tenth International Conference on Learning Representations, 2022.
46. M. Varadi *et al.*, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439-D444 (2022).
47. J. I. Garzón, L. Deng, D. Murray, S. Shapira, D. Petrey, B. Honig, A computational interactome and functional annotation for the human proteome. *eLife* **5**, e18715 (2016).
48. Q. C. Zhang, L. Deng, M. Fisher, J. Guan, B. Honig, D. Petrey, PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res.* **39**, W283-W287 (2011).
49. S. Gudmundsson *et al.*, Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **43**, 1012-1030 (2022).
50. A. J. M. Walhout, M. Vidal, High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping. *Methods* **24**, 297-306 (2001).

Figure legends

Fig. 1. Overview architecture of PIONEER2.0. (**A**) The sequence information of both target and partner proteins is encoded through the Transformer and CNN encoders for the extraction of long-range and short-range dependencies among residues, respectively. The embeddings are then concatenated through self-attention pooling. The structural information of both target and partner proteins is encoded on both residue- and atom-levels, respectively. The residue- and atom- level embeddings are then concatenated through self-attention pooling and are further encoded by the graph Transformer encoders for target and partner, respectively. The resulting embeddings of both target and partner proteins are concatenated through self-attention pooling. Finally, the sequence and structural embeddings are concatenated and are input to the dense layer for the interface residue prediction of the target proteins. (**B**) Overview of the PrePPI pipeline which provides features for the sequence branch. (**C**) Comparison of PrePPI and PIONEER2.0 performance at recovering the PDB test set. (**D**) Comparison of receiver operating characteristic (ROC) curves of PIONEER2.0 and AlphaFold3 for evaluation on the PDB test set.

Fig. 2. PIONEER2.0 and AlphaFold3 interface residue predictions for protein interactions of known and unknown structures. (**A**) Distribution of AlphaFold3 ranking scores on interactions with and without PDB structures. (**B**) Cumulative distribution of AlphaFold3 ranking scores on interactions with and without PDB structures. (**C**) Comparison of ROC curves for PIONEER2.0 and AlphaFold3 using the bootstrapped samples from the test set. AUROC = 0.5 for a random prediction. (**D**) Comparison of PIONEER2.0 and AlphaFold3 on interactions with AlphaFold3 ranking score below 0.5, using the bootstrapped samples. The vertical axis denotes the metric value. The "***" indicates that the *P*-value is less than 0.001.

Fig. 3. PIONEER2.0 outperforms AlphaFold3 for interface residue prediction across various AlphaFold3 model confidence levels. (**A-D**) Comparison of PIONEER2.0 and AlphaFold3 on interactions with AlphaFold3 ranking scores below 0.2 (**A**), 0.3 (**B**), 0.4 (**C**)

and 0.5 (**D**), respectively. The "*" and "***" indicate that the *P*-value is less than 0.05 and 0.001, respectively. In the bar plots, the vertical axis denotes the metric value. (**E-F**) The AlphaFold3 predictions on the interactions between the proteins encoded by rplI and rpmB (PDB: 5WIT, chains 1I and 11) (**E**) and by casE and casB (PDB: 4TVX, chains A and J) (**F**) demonstrate that AlphaFold3 fails to correctly position the two proteins relative to each other, leading to inaccurate interface predictions. The light blue and blue represent the true co-complex structures, and the yellow and orange represent the AlphaFold3 predictions. The green and red spheres correspond to correct and incorrect PIONEER2.0 interface predictions, respectively.

Fig. 4. PIONEER2.0 provides high-quality interface residues for the structurally informed human protein-protein interactome. (**A**) The interactome includes interface residues calculated from experimentally determined PDB structures, and PIONEER2.0 predictions of interface residues for the remaining unresolved interactions. (**B**) Fraction of interactions disrupted by random population variants in PIONEER2.0-predicted and known interface residues. The error bar denotes standard error for the binomial distribution. Significance was determined by the two-sided *z*-test. (**C**) Enrichment of population variants in PIONEER-predicted and known interfaces. The error bar denotes standard error for the log odds ratio. (**D**) Enrichment of disease-associated mutations in PIONEER-predicted and known interfaces. The error bar denotes standard error for the log odds ratio. Significance was determined by the two-sided *z*-test.

Supplementary Fig. 1. The proportions of the human protein-protein interactome with known (PDB) and unknown structures.

Supplementary Fig. 2. Comparison of PIONEER2.0 and AlphaFold3 across various AlphaFold3 model confidence levels. (**A-I**) Comparison of PIONEER2.0 and AlphaFold3 on interactions with AlphaFold3 ranking scores within 0.0-0.2 (**A**), 0.2-0.3 (**B**), 0.3-0.4 (**C**), 0.4-0.5 (**D**), 0.5-0.6 (**E**), 0.6-0.7 (**F**), 0.7-0.8 (**G**), 0.8-0.9 (**H**), 0.9-1.5 (**I**),

respectively. The "*", "**" and "***" indicate that the *P*-values are less than 0.05, 0.01, and 0.001, respectively. "ns" indicates not significant.

Acknowledgments

Funding

National Institute of General Medical Sciences grant R35GM139585 (B.H.)

National Institute of General Medical Sciences grant R01GM124559 (H.Y.)

National Institute of General Medical Sciences grant RM1GM139738 (H.Y.)

Simons Foundation Autism Research Initiative grant SFARI-893926 (H.Y.)

Author contributions

H.Y. conceived and oversaw all aspects of the study. D.X. and H.Y. developed the model. D.M., A.C.N., and B.H. contributed the PrePPI results. R.F. conducted the Y2H experiments. D.X., M.T., D.M., L.L., B.H. and H.Y. performed the analyses. H.Y. wrote the manuscript with inputs from D.X., M.T., D.M., and B.H. All authors discussed the results and reviewed the manuscript.

Completing interests

The authors declare no completing interest.

Data and materials availability

The main data supporting the results in this study are available within the supplementary materials.

Supplementary Materials

Figs. S1-S2

Data

Figure 1

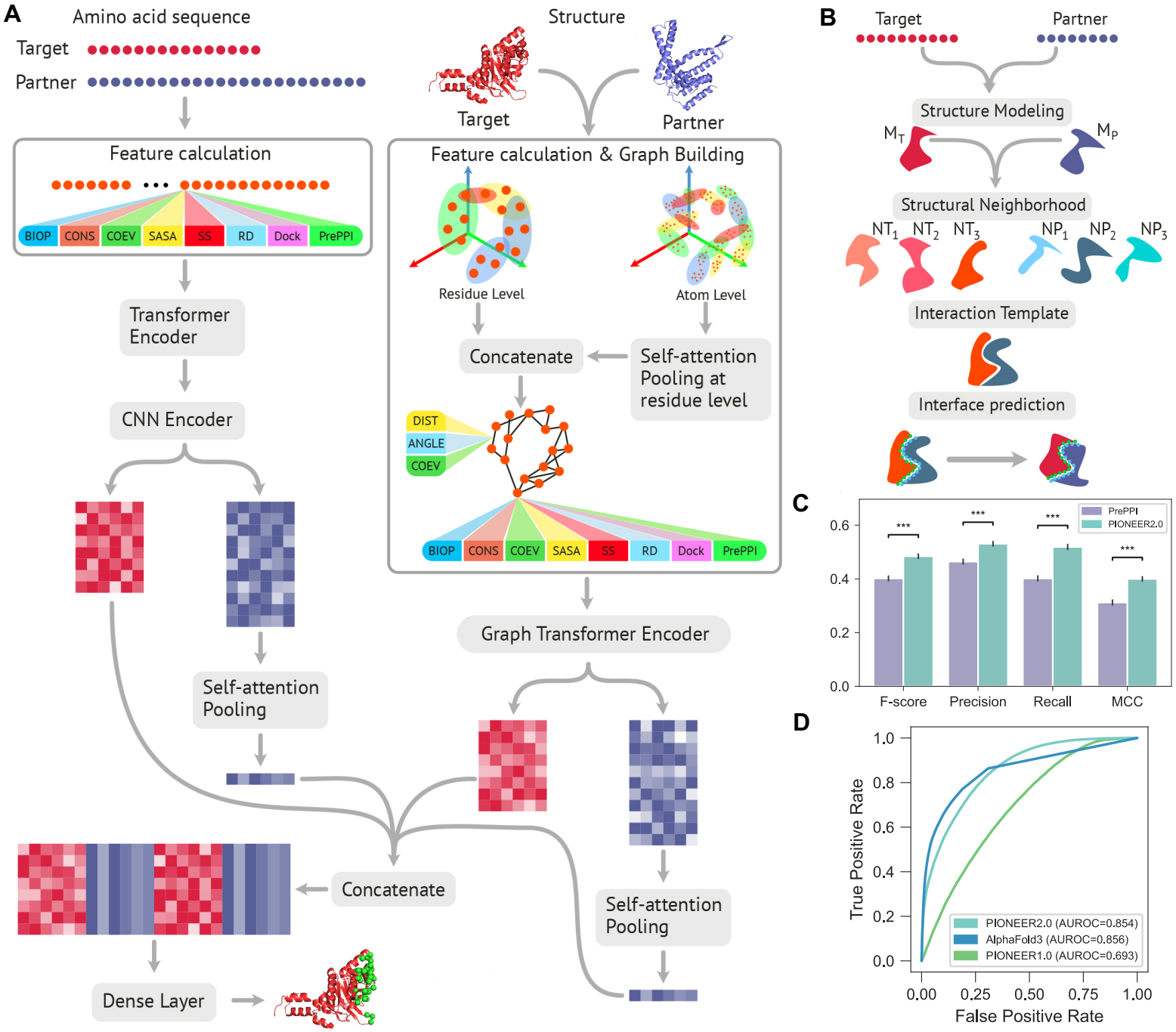


Figure 2

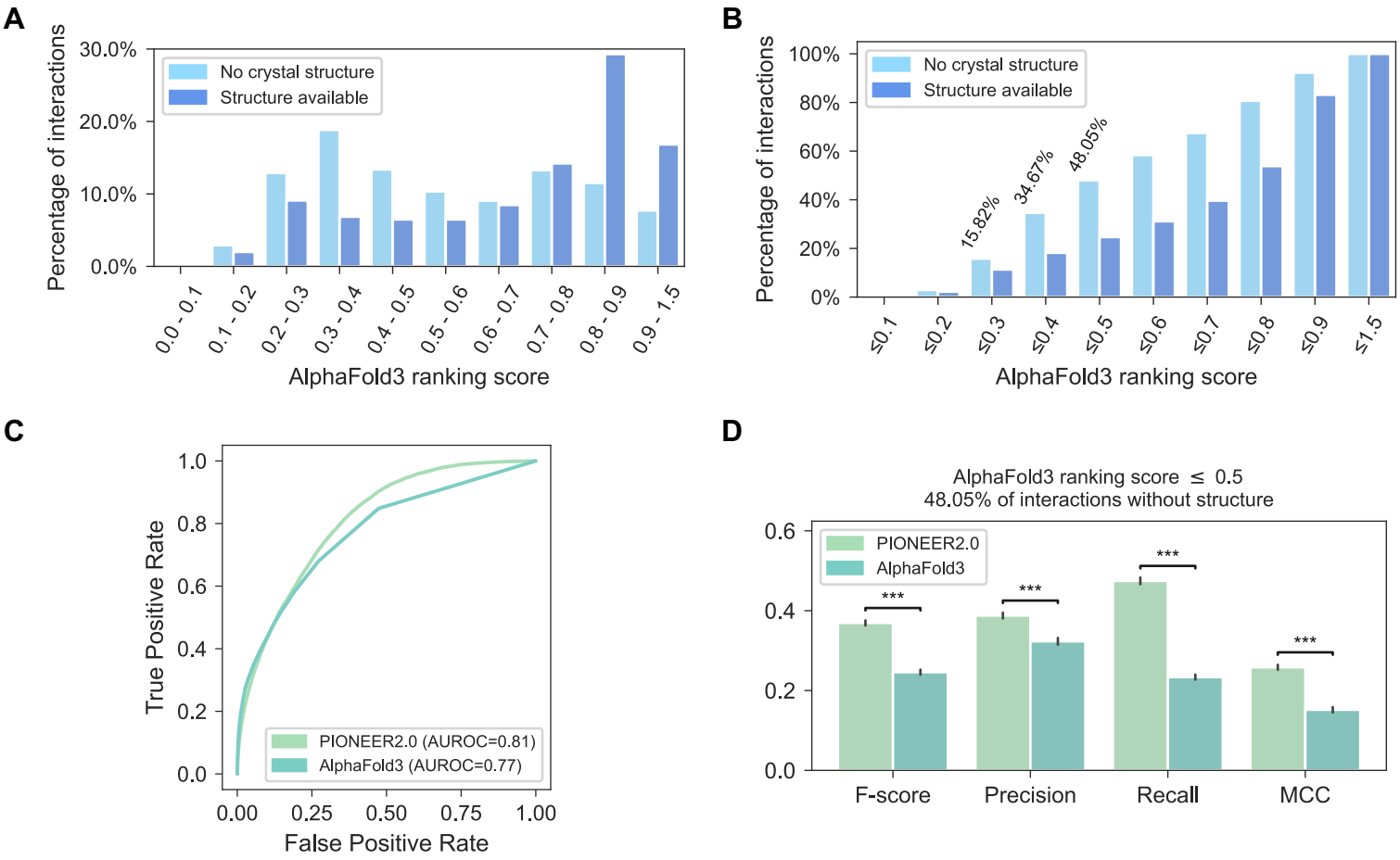
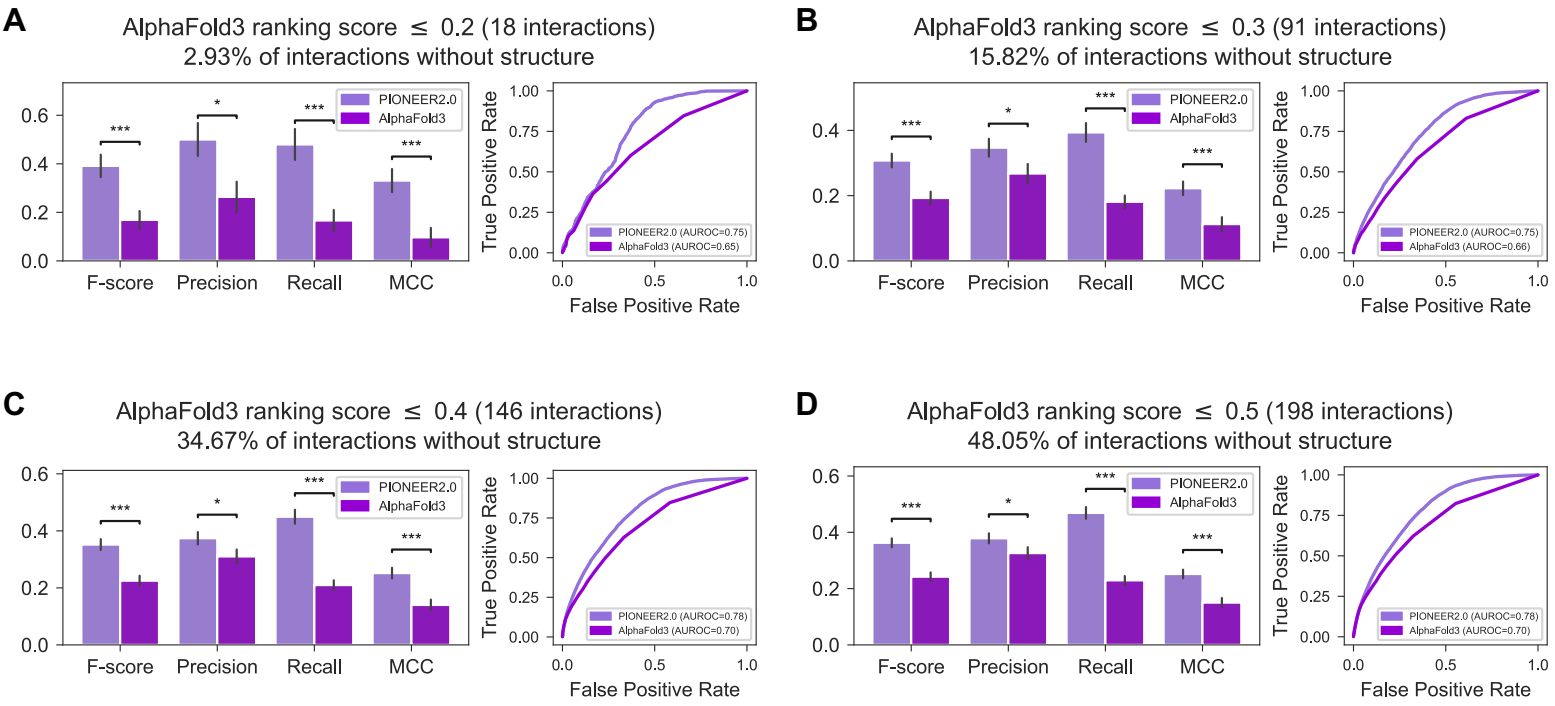


Figure 3



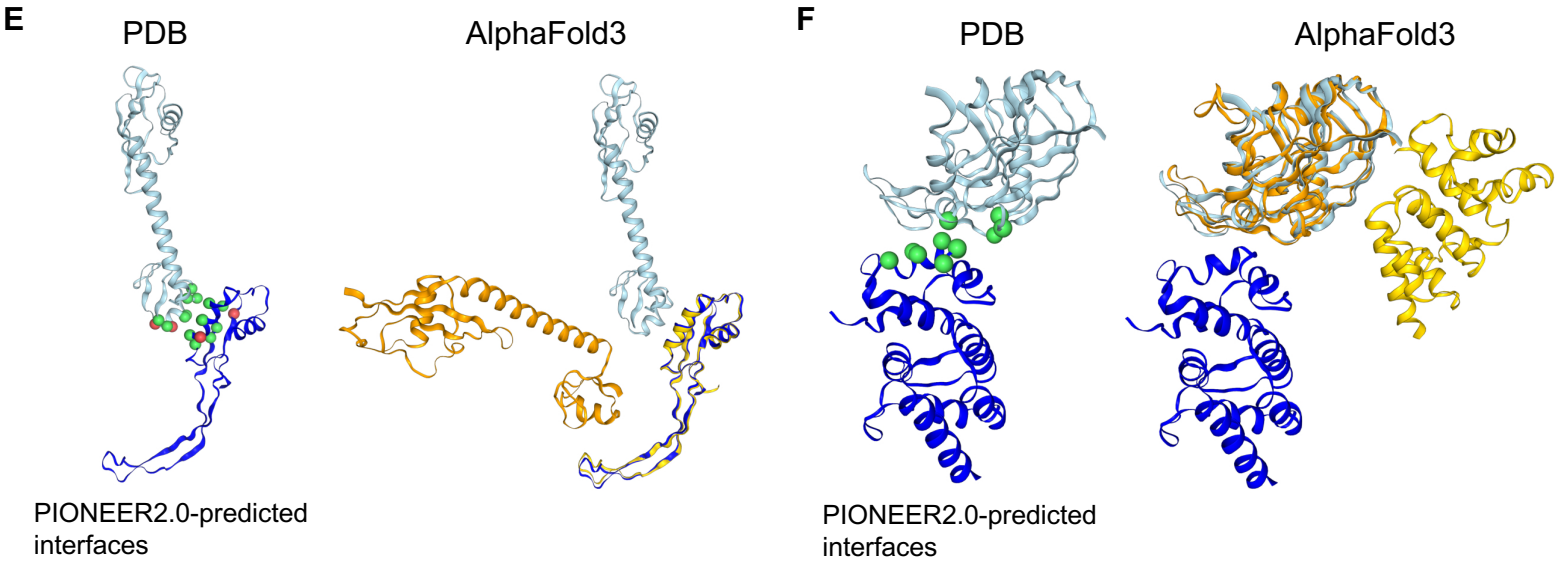


Figure 4