Topology-Driven Negative Sampling Enhances Generalizability in Protein-Protein Interaction Prediction

Ayan Chatterjee[1,2,*], Babak Ravandi[1,2,3,†,*], Naomi H. Philip[1], Mario Abdelmessih[1], William R. Mowrey[1], Piero Ricchiuto[1], Yupu Liang[1], Wei Ding[1], Juan C. Mobarec[4], Tina Eliassi-Rad[2,5,6,7]

[1]Alexion AstraZeneca Rare Disease, Boston, MA, USA
[2]Network Science Institute, Northeastern University, Boston, MA, USA
[3]Department of Physics, Northeastern University, Boston, MA, USA
[4]Mechanistic and Structural Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
[5]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[6]Santa Fe Institute, Santa Fe, NM, USA
[7]The Institute for Experiential AI, Northeastern University, Boston, MA, USA

[†]Corresponding author e-mail: Babak.Ravandi@Alexion.com

ABSTRACT

Unraveling the human interactome to understand biological processes and uncover disease-specific patterns hinges on accurate protein-protein interaction (PPI) predictions. However, challenges persist in machine learning (ML) models due to a scarcity of quality hard negative samples, shortcut learning, and limited generalizability to novel proteins. Here, we introduce ComPPlete (Completing the Protein-Protein Interaction Network), a novel ML pipeline utilizing PPI network topology for strategic sampling of protein-protein non-interactions (PPNIs) by leveraging higher-order network characteristics that capture the inherent complementarity-driven mechanisms of PPIs. Integrating unsupervised pre-training in protein representation learning with topological PPNI samples, ComPPlete improves PPI prediction generalizability and interpretability, particularly in identifying potential binding sites locations on amino acid sequences. ComPPlete strengthens the prioritization of screening assays, facilitates the transferability of ML predictions across protein families and homodimers. ComPPlete establishes the foundation for a fundamental negative sampling methodology in graph machine learning by integrating insights from network topology.

[*]Ayan Chatterjee and Babak Ravandi equally contributed.

INTRODUCTION

Proteins play a central role in essential biological processes, including catalyzing reactions, transporting molecules, responding to pathogens in the immune system, and facilitating cell-to-cell signal transduction [1, 2]. Moreover, crucial cellular processes vital for our health, such as DNA replication, transcription, translation, and transmembrane signal transduction, depend on specific functional proteins [3]. These fundamental biological activities are regulated through protein complexes, typically governed by protein-protein interactions (PPIs) [4, 5]. In humans, deviations from typical patterns of PPIs and protein complexes can either cause or indicate a disease state [6]. Numerous computational methods have been developed to uncover the etiology of diseases form PPIs [7]. However, the incompleteness of the human interactome [8, 9] hinders the understanding of pathogenic and physiologic mechanisms that trigger the onset and progression of diseases [10], and hence in the development of novel therapeutic strategies [11].

Experimental PPI databases such as BioPlex [12], STRING [13], APID [14], BioGRID [15], CoFrac [16], CORUM [17], HuRI [18], HINT [19], and HIPPIE [20], to name a few, capture human PPIs observed via Affinity Purification - Mass Spectrometry (AP-MS) [21] and Yeast-to-hybrid (Y2H) assays [22]. However, none of the PPI databases report the failed experiments, i.e., the non-interactions, creating a scarcity of high-quality protein-protein non-interactions a.k.a. hard negatives [23] in training ML models (see Figure 1A). Negatome [24] stands out as the sole database that has endeavored to tackle this concern; however, it encompasses merely 2,424 interactions linked to human proteins [25] and is limited by specific environmental constraints. As a result, it falls short of capturing the essential interaction mechanisms required for training machine learning models to prioritize large-scale screening. Zhao et al. [26] provided triple-layer validated [27] negatives, yet only 58 pairs are provided for the human proteins reported in UniProt [25]. Furthermore, these databases suffer from high rates of false negatives and false positives [28–30], which combined with the limitations of the traditional ML-based negative sampling methods significantly bias the PPI predictions [31]. For example, some authors propose generating high-quality non-interactions by considering pairs of proteins with distinct cellular localization [32], presumably hindering their participation in biologically relevant interactions [33, 34]. This method samples protein pairs as negative samples from the compartment pairs that do not share any protein (see Figure S1). Alternatively, other authors adopt a simpler approach, randomly selecting protein-protein non-interaction (PPNI) pairs from the entire set of protein pairs that are not known to interact [35–38]. However, both negative sampling scenarios produce a skewed distribution of negative examples resulting in overly optimistic estimates of classifier accuracy [33, 39, 40]. Furthermore, Bardes et al. [41] have shown that the necessity of

negative sampling while training semi-supervised models for predicting similarity between two entities can be overcome by regularization. Nevertheless, as the similarity of proteins does not necessarily indicate interaction, the need for negative sampling persists in PPI prediction. The complementary nature of the PPI network [42] demands hard non-interaction samples for unbiased prediction and improved generalizability. In a recent paper, Li et al. [43] have proposed a heuristic-based hard-negative generation method for link prediction. However, this method lacks biological and chemical rationality in creating protein-protein non-interactions and impedes the interpretability of the predictions. Finally, the complex interdependence of the in-vitro screening experiments on the environmental factors makes it difficult to obtain hard negatives solely driven by the molecular properties of the proteins [44]. Understanding the evolutionary reasoning for protein interactions and non-interactions is a complex task. However, recent research has unveiled the complementarity-driven mechanisms [42] driving protein-protein interactions, which include an electrostatic charge or Coulombian complementarity [45], hydrophobic mismatch [46], conformational complementarity [47], hydrogen bond complementarity [48], etc. (see Figure 1C). Therefore, delineating interactions and non-interactions based on the complementary nature of the PPI network not only facilitates the generation of hard negative samples for training PPI prediction models but also enhances the mechanistic interpretability of these predictions.

Historically, the scarcity of hard negatives and the pronounced imbalance between edges and non-edges during the training of machine learning models for link prediction tasks have led to the practice of obtaining negative samples through random sampling from the complementary graph [49]. Nevertheless, it is noteworthy that random negative samples can introduce substantial bias into the separation hyperplane learned by the classifier, as illustrated in Figure 1B. Selecting an appropriate set of hard negatives is essential for acquiring the relevant hyperplane and enhancing generalizability in machine learning [23].

On a separate note, observation bias [50, 51] and shortcut learning [52, 53] are prevailing in the ML-based PPI prediction models due to selection and laboratory biases in the PPI databases [54]. The excellent transductive (see Section S1, Figures S2 and S3) cross-validation performances of the majority of the state-of-the-art (SOTA) PPI prediction models [55, 56], falsely exaggerated by topological shortcuts [51,53,57], are often misleading and not biologically interpretable. Multiple recent papers have focused on the poor performance of state-of-the-art link prediction models on low-degree nodes [43,50,58]. GraphPatcher [58] proposes a novel test-time augmentation method to improve link prediction performance for the low-degree nodes. However, the approach is infeasible for unseen nodes, making the prediction task for never-

before-seen nodes of greater difficulty [50], which maps to the scenario of predicting interactions between novel proteins. Therefore, inductive tests have garnered significant attention recently for their capability to reveal the authentic predictive efficacy of machine learning models (see Figure S2). This allows for the assessment of a model's performance on novel entities, ultimately contributing to enhanced generalizability and interpretability [53, 59]. Recent studies [59, 60] emphasize the importance of going beyond the PPI network topology in machine learning and advocate for the incorporation of inductive tests in constructing models with biological utility and interpretability. Yet, intricate calibration of these models via fine-tuning [61] and embedding regularization limit their generalizability within particular databases and families of proteins [59].

Finally, using the appropriate performance metric is essential for evaluating the true performance of an ML model [62, 63]. Global performance metrics like precision, recall, the area under the receiver-operating characteristics (AUROC), the area under the precision-recall curve (AUPRC), and accuracy fail to capture the local performance in the PPI network [64]. Hence, performance metrics like Hits@TopK [65] and mean reciprocal rank (MRR) [66], which incorporates local evaluation [67], are necessary for assessing an ML model for novel, unseen proteins in inductive tests. Furthermore, the evaluation of negative predictions is often overlooked. However, the ability of an ML model to distinguish the positives from the negatives largely determines its predictive power [68]. Yet, a systematic approach is absent in evaluating PPI prediction models.

The **contributions** of this work are as follows:

1. We introduce ComPPlete (short for Completing the Protein-Protein Interaction Network), an ML pipeline that bridges the existing gaps in the PPI prediction methodology by
   (a) including a method for obtaining high-quality protein-protein non-interaction (PPNI) pairs, which utilizes the PPI network topology;
   (b) connecting PPNI pairs to complementary generation mechanisms through hyperbolic embedding space;
   (c) improving inductive link prediction performance in PPI using topological negatives; and
   (d) enabling interpretability and transferability of PPI prediction across protein families and homodimers.
2. We propose local performance metrics to evaluate both interaction and non-interaction predictions, quantifying the distinguishing power of any PPI prediction model.

3. We predict and validate interactions between understudied and difficult-to-purify G protein-coupled receptors (GPCRs).

RESULTS

Topological PPNI

Negative sampling is an inextricable part of binary classification task in machine learning [69]. The lack of high-quality hard negatives in biological data is a major hindrance in developing practical ML models. The majority of the biological databases have a sufficient number of positive examples, but a considerably low number of negative samples due to experimental and observational biases [51], which is a major cause of both overall class imbalance and node-wise class imbalance (annotation imbalance) [53]. Random negative sampling is the most frequently used approach to mitigate this problem. Yet, random negative samples are consequential in closed systems [70] and are prone to degree bias [53, 58]. In this section, we propose a novel method of creating hard PPNIs by leveraging the PPI network topology [71], which is underpinned by the complementary nature of the interactome [42]. We propose topological PPNI (see Figure 2A) that combines entropy-based network null models [72] and higher-order PPI network properties [71] in prioritizing hard PPNI samples. Topological PPNI methodology consists of two parts: **(a)** a traditional unipartite configuration model, and **(b)** contrastive-L3 (CL3) filtering.

We use the traditional unipartite configuration model (see Methods), which takes as input only the degree sequence of the PPI network and achieves a commendable performance in transductive link prediction (AUROC: 0.87 ± 0.0002, AUPRC: 0.87 ± 0.0002, Hits@Top100:* 0.98 ± 0.016). The output of the configuration model dictates the interaction probability between two proteins from a topological standpoint. We select the bottom-N (N=10 million) predictions from this configuration model on all human protein pairs and consider them as less probable for binding, a.k.a, potential hard negative samples. Next, we use a complementarity-driven topological property of the PPI network to derive the hard negatives. Unlike the formation of triadic closure in social networks [73], proteins tend to form quadratic closures or even-length cycles [71]. Two friends in a social network have multiple common friends and hence, many paths of length 2 (L2). On the other hand, two interacting proteins have many paths of length 3 or L3 between them, which is related to the evolutionary aspects and the complementary nature of PPIs [42, 71]. We propose the Contrastive-L3 (CL3) hypothesis (see Figure 2A), which states that two non-interacting proteins would not constitute any odd-length (L=3) path in the

*Precision at top K predictions with K=100

PPI. We now use CL3 hypothesis to filter high-quality hard negatives from the output of the configuration model. We filter the output of the configuration model mentioned above to fetch the protein pairs that induce no L3 path in the PPI and obtain 3,063,605 hard negatives (see Figure 2A). Note that, ∼18,000 human proteins in different PPI databases can computationally create ∼156 million potential pairs. This is a computationally infeasible space for computing all L3 paths. Therefore, adding the non-interaction prediction from the configuration model significantly reduces the L3 computation space and provides a regime consisting of proteins lacking sufficient annotations, constituting the potential space for sampling hard PPNIs.

Geometric and Evolutionary Relevance of PPNI

Recent studies have demonstrated the ability of hyperbolic embeddings to capture evolutionary patterns driven by complementarity in PPIs [74]. Figure 2 illustrates various PPNI sampling methods in a two-dimensional hyperbolic space [74, 75]. We use the NetHypGeom hyperbolic embedder proposed by Alanis-Lobato et al. [74]. In Figure 2B, we visualize proteins involved in Subcellular Compartment Negatives [33, 34] (SCN, see Fig S1), Negatome 2.0 [24], and topological PPNIs involving distinct proteins. Alanis-Lobato et al. [74] have associated the radial and angular coordinates of the 2D hyperbolic plane with the evolutionary aspects of human proteins. According to their findings, proteins closer to the origin are evolutionarily older, while those with higher radial coordinates are younger (see Figure 2C). We observe that SCN PPNI involves both old and new proteins due to constraints related to subcellular localization. Negatome PPNI is primarily constrained by well-studied older proteins. In contrast, topological PPNI is constrained to younger proteins. From an evolutionary standpoint biological pathways have evolved around older proteins, resulting in their interactions with the majority of other proteins [74, 76]. Conversely, younger proteins exhibit fewer interactions among themselves. Moreover, protein families are linked to angular coordinates within the hyperbolic space, as indicated by research on latent geometry in PPNI networks [74]. The topological PPNI encompasses the entire range of angular coordinates, thereby encompassing samples from diverse human protein families, as illustrated in Figure 2B.

In Figure 3A, we present a visualization of PPI and PPNI edges in a 2D hyperbolic space. The computation of the edge (both PPI and PPNI) embedding between two proteins involves averaging the hyperbolic embeddings of the respective end proteins across each dimension. The substantial overlap observed between SCN PPNI and the experimental PPI implies that SCN fails to effectively capture the complementarity mechanism influencing protein interactions [42]. Consequently, SCN PPNI proves inadequate for training a classifier capable of

distinguishing between protein interactions and non-interactions based on complementarity. A parallel observation is made for Negatome PPNI. Conversely, topological PPNI demonstrates effective separation between PPI and PPNI in the hyperbolic space, successfully capturing complementarity-driven interaction mechanisms through negative samples. The hyperplane learned by ComPPlete, visualized in Figure 3A, effectively distinguishes PPI from PPNI in the hyperbolic space. While training a classifier on the hyperbolic embeddings would limit the complementarity-driven link prediction to the transductive scenario and would not be able to make PPI predictions for new proteins, ComPPlete demonstrates the ability to learn complementary mechanisms of protein interactions from the amino acid sequences when trained using experimental PPI and topological PPNI, and can generalize the learning to new, unseen proteins in inductive tests, hence improving both generalizability and biological interpretability of PPI prediction.

In Figure 3B, we visualize the distributions of pairwise geodesic distances [42] between protein pairs for PPI and various PPNIs. Notably, the pairwise distance distribution for topological PPNI is distinctly separated and positioned to the right of the distribution for PPIs. This indicates that, in the hyperbolic space, interacting proteins are geodesically closer compared to those in topological PPNI. Moreover, topological PPNI demonstrates superior separability between non-interacting proteins compared to SCN and Negatome.

Figure 3C illustrates the L3 paths generated by pairs in PPI and various PPNI scenarios. The visual representation reveals that both SCN and Negatome PPNI induce a significantly lower number of L3 paths in the PPI network when compared to experimentally validated PPIs. To identify topological PPNI, we focus on protein pairs that do not induce any L3 paths in the PPI network. Significantly, given that both L3 paths and the hyperbolic space capture complementarity-driven mechanisms in PPI, protein pairs in the topological PPNI category, which induce no L3 paths, are positioned farthest away in the hyperbolic space. This distinct positioning underscores their clear separation from PPI pairs that exhibit the highest number of induced L3 paths, providing a distinctive characterization.

Finally, we explore a case study involving human proteins to illustrate how topological negatives effectively capture complementary interaction mechanisms. Canonical Olfactory Receptors (ORs) represent a class of G-protein-coupled receptors (GPCRs) in mammals [77]. Monomeric ORs in mammals are activated by chemical ligands and couple to specific G-proteins. For example, ORs in the olfactory sensory epithelium activate Cyclic adenosine monophosphate (cAMP) and other second messenger signaling, leading to ion channel opening and membrane depolarization. Contrastingly, insect ORs function as heteromeric ion channels and lack homol-

ogy to G-protein-coupled chemosensory receptors found in vertebrates [78, 79]. Humans possess more than 400 ORs, while rodents have approximately 1000 OR genes [80]. Hence, it is biologically justified to observe a higher number of non-interactions between human OR proteins to other human proteins, which is reflected in our topological PPNI samples. We observe a total of ∼15,000 topological negatives involving the human OR proteins, among which ∼7,000 samples include proteins from inter-OR families. Furthermore, biological justification suggests that topological PPNIs are dependent on the species under consideration. Finally, we validate the association between OR-relevant topological PPNI and complementarity mechanisms. In Figures 3D and 3E, we visualize the distribution of the product of hydropathy [81, 82] and electrostatic charge [82] for human OR protein pairs in both PPI and topological PPNI. Divergent distributions, with the PPNI distributions being more positive, validate that topological PPNI effectively captures hydrophobic mismatch and eletrostatic complementarity mechanisms [42].

ComPPlete

Now we use the topological negatives to design ComPPlete, a generalizable, robust, interpretable, and transferable PPI prediction ML pipeline. ComPPlete uses unsupervised pre-training of node attributes for improved generalizability to unseen proteins [50].

ComPPlete Architecture

ComPPlete uses pair-wise learning [83] combined with protein representations pre-trained in an unsupervised fashion [50] for improving the generalizability of ML-based PPI prediction. Figure 4A visualizes the neural architecture of ComPPlete. We use ProtVec [84] for embedding the protein amino acid sequences. ProtVec is trained on ~ 0.5 million amino acid sequences (1.6 million trigram sequences) available in SwissProt [85].

Although machine learning (ML)-based protein-protein interaction (PPI) prediction has been investigated using 3D protein structures [86, 87], recent insights into the dynamic and self-organizing nature of the cell criticize approaches that consider the rigid structures of proteins [88]. The limitations of the current ML PPI prediction models, stemming from the scarcity of experimental 3D human protein structures and the uncertainties associated with AlphaFold [89], impede achieving high accuracy, robustness, and generalizability. The assumed rigidity of 3D structures fundamentally constrains the predictive power of PPI prediction models. Therefore, we opt to utilize amino acid sequences as a basis for learning PPI from protein representations, aiming to overcome these limitations and enhance the predictive capabilities [86–89].

The ProtVec embeddings are fed into two arms of the downstream ComPPlete decoder,

which formulates a binary classification task. We use a 3-layer multi-layer perceptron (MLP) [90] as the decoder. ComPPlete is trained in an inductive setting (see Figure S2) [50]. We use the GoldPPI (see Methods) and Negatome PPNI data in the validation and test datasets for gaining uncompromising confidence in ComPPlete predictions through harder tests. Following the inductive link prediction setup from GraIL [91], the human proteins are randomly divided into three groups for train, validation, and testing, with the proteins from the GoldPPI and Negatome PPNI predominantly residing in the validation and test groups. Then, the interactions induced by these proteins are sampled to create the train, validation, and test datasets. We implemented an early-stopping [92] on the inductive validation dataset to avoid overfitting. ComPPlete is trained and tested in a 5-fold cross-validation setting.

Inductive Link Prediction Performance

Two state-of-the-art PPI prediction models, DeepTrio [93] and PPI-GNN [55] perform exceptionally well in transductive link prediction (see Section S2). Yet, their performance significantly diminishes in inductive tests (see Table 1, Figure S3). In inductive link prediction, PPI-GNN and DeepTrio achieve significantly lower performance than ComPPlete (see Table 2, Figure 4B). Furthermore, we borrow two performance metrics from recommender systems into PPI prediction to assess the quality of the ranking made by PPI prediction models. We define PPIHits@TopK as the number of true interactions (precision) identified by an ML model in the top K predictions. We also define, PPNIHits@BottomK, which quantifies the number of true non-interactions identified in the bottom K predictions of a model. These two metrics bolster our confidence in the ranking provided by the model, which is associated with the separability of the predictions for the PPI and PPNI in test data. ComPPlete establishes itself as an excellent ranking tool for identifying PPI in terms of both metrics (PPIHits@Top1000: 0.92 ± 0.015 and PPNIHits@Bottom1000: 0.69 ± 0.08). In Figure 4C, we observe how DeepTrio creates overlapping predictions for the true interactions and true non-interactions. The incorporation of PPNI at the output of the configuration model combined with pair-wise learning improves the separation between these distributions and helps ComPPlete predict the true interactions with higher probabilities. Finally, filtering the PPNI with L3 counts (topological PPNI) separates the ComPPlete output distributions, confirming the ability of ComPPlete to learn the latent protein-protein interaction patterns driven by various molecular complementarity mechanisms. These observations provide an ablation study for the topological PPNI and showcase the strength of the CL3 approach in creating hard negatives.

On a similar token, the recent discovery of shortcut learning [52] has raised many questions

about the reliability of ML predictions and inductive tests have gained significant attention in generalizability and interpretability [50, 53]. RAPPPID [59] is a recently proposed PPI prediction model that uses inductive learning to improve the generalizability of PPI prediction. RAPPPID uses sequential neural models (LSTM) [94] to embed the protein amino acid sequences and feed them to a downstream decoder, which is trained for an inductive setting in an end-to-end fashion. Yet, the regularization of the protein embeddings limits the generalizability of RAPPPID within certain protein datasets and protein classes. Performance of RAPPPID significantly reduces in a transfer learning scenario from STRING [13] to BioLip [95]. In Table 3, we see that ComPPlete achieves comparable performance to RAPPPID for inductive link prediction within similar protein families, while performing better than RAPPPID in a transfer learning setting from STRING to BioLip. Data-specific regularization hinders RAPPID from achieving superior generalizability across datasets and protein families, which is ameliorated in ComPPlete by unregularized and unsupervised protein representations, independent of the training dataset.

In Figure 4D, we plot the ComPPlete prediction for all experimentally validated PPIs against the number of L3 paths induced by these pairs in the PPI network. A robust correlation is evident between the ComPPlete predictions and the L3 counts, with $r_{Spearman} = 0.48$. This observation confirms that ComPPlete effectively captures the complementarity-driven mechanism underlying L3 in PPIs based on protein amino acid sequences. Moreover, as depicted in Figure 4E, the ComPPlete predictions for the majority of the experimentally validated PPIs surpass the optimal classification threshold (0.476 ± 0.049).

Robustness and Sensitivity of ComPPlete

The robustness of link prediction with the introduction of noise and adversarial attacks in the network provides insight into the reputability of machine learning models [96, 97]. Robustness of PPIs have been studied from various aspects, including randomized null models [98], temperature dependence of stable protein complexes [99], and the removal of hub proteins [100]. PPI networks are topologically more robust under four common types of perturbations (i) network nodes are randomly removed (failure), (ii) the most connected node is successively removed (attack), (iii) interaction edges are rewired randomly, and (iv) edges are randomly deleted [101].

In Figure 5A, we show how L3 paths (4-cycles) in PPI make the networks robust under degree preserved edge swap [102]. In Figure 5B, we observe that the inductive link prediction performance of ComPPlete does not fluctuate under degree-preserved edge swap in train and validation datasets, while keeping the test data unchanged. This form of robustness is inherent

to the PPI network topology, and degree-preserved edge swap is insufficient for evaluating the robustness of any ML-based PPI prediction model.

Next, we test the sensitivity of ComPPlete on the protein representations. We replace the 100-dimensional ProtVec embeddings with vectors whose entries are drawn at random from a uniform distribution $U[0, 1]$. In Figure 5C, we observe that ComPPlete performance drops significantly when the protein representations are randomized and the performance is comparable to a naive Bayes' classifier. This confirms that ComPPlete learns PPI mechanisms leveraging the protein amino acid sequences and does not refuge to any form of shortcut learning.

Finally, we test the robustness of ComPPlete under random node and edge deletion. We perturb the train and validation datasets while keeping the test data unchanged. In Figures 5D and 5E, we observe slight decay in the inductive test performance of ComPPlete under random deletion of nodes and edges from the training and validation PPI. However, even after deleting 80% of training nodes and edges, ComPPlete is still able to achieve commendable inductive link prediction performance. This confirms the high generalizability of the proposed ML pipeline under data scarcity and the power of topological negatives in capturing the complementary mechanisms behind protein interactions.

Interpretability and Identifying Interaction Locations

Finally, the molecular interpretability of PPIs is crucial for understanding their biological significance, their role in various cellular processes, and their potential as drug targets. We developed an ML-based approach for identifying putative interaction regions in a given pair of putative interacting proteins by ablating their sequences [103]. Keeping the amino acid sequence of one protein fixed at a time, we mutate each amino acid trigram of the other protein by replacing it with the Out-of-Vocabulary (OOV) probing with in-vocabulary example (IVE) entry in ProtVec, and observe the change in ComPPlete output prediction probability. Note that, the OOV embedding vector in ProtVec is obtained by the average of all the trigram embedding vectors [84]. The valleys of the binding probability profiles correspond to the potential binding sites on the protein sequence. We validate this method on experimentally validated self-interacting protein pairs Drd2 [104] and ComA [105] heterodimeric complexes of proteins: Leukemia inhibitory factor (LIF) in complex with gp130, and mu-opioid receptor bound to NbE. In Figure 5 (F-I), we mark the valleys identified by ComPPlete which correspond to the binding interfaces identified by experimentally for these complexes. For Drd2 the interacting TM helices have been reported as self-interactions between a region around TM4, and another one in TM1 utilizing Cys-crosslinking and FRET [104] (Figure 5F). The transcription factor Competence

protein A (ComA) dimerization has been determined by NMR showing self-interaction mostly on C-terminal helix alpha-10 [105]. LIF interaction interface with gp130 was determined by X-ray crystallography [106]. Finally, the binding site of NbE on the orthosteric site of the mu-opioid receptor was also determined with X-ray crystallography [107].

Novel interactions for G protein-coupled receptors

Physiologically and pharmacologically relevant G-protein-coupled receptors (GPCRs) are the target of several marketed drugs [108]. They mediate physiological responses to hormones, neurotransmitters, and environmental stimuli. Recent high-resolution structural studies have shed light on the molecular mechanisms of GPCR activation and constitutive activity. Over the past three decades, significant progress has been made in understanding GPCRs, from pharmacology to in vivo function. However, only G protein-coupled receptor-G protein interactions have been explored extensively [109]. We used ComPPlete to predict potential homodimers and heterodimers between GPCRs. We consider 170 family A GPCR proteins for this prediction task. In Figure 6, we visualize the GPCR interaction network from top 100 ComPPlete predictions. The network consists of 28 GPCR protein nodes with 8 homodimers and 92 heterodimers. Furthermore, since ComPPlete is not trained on self-interactions, we evaluated the homodimeric predicted by ComPPlete with AlphaFold-Multimer predicted homodimer structures [110, 111]. In Table 4, we summarize the AlphaFold average of predicted local distance difference test (pLDDT) score, atoms at interface, and interface pLDDT score for top and bottom homodimer predictions made by ComPPlete (see Figure 6). Segregation of the potential GPCR interactions predicted by ComPPlete is in agreement with more computationally expensive results from AlphaFold-Multimer. Hence, ComPPlete can be used as a potential high throughput tool for novel PPI prediction.

DISCUSSION

While machine learning has widely been used in protein-protein interaction prediction, generalizing these models to novel proteins, transferring across protein families, and interpretability of predictions from a molecular standpoint have emerged as the major pitfalls of these models. The scarcity of high-quality, biologically relevant hard PPNI sampling has been a major hindrance for machine learning models to learn interaction mechanisms from protein structural patterns. We have proposed a novel approach to leveraging the PPI network topology in sampling PPNI contingent on the complementarity-driven generation mechanisms of protein-protein interactions. We interpret topological PPNI geometrically by leveraging the hyperbolic space

and make a connection between the complementarity of PPI and the evolution of proteins. Our PPNI sampling approach combined with unsupervised pre-training of protein representation not only improves the generalizability of PPI prediction but also improves the transferability of machine learning prediction across protein families. ComPPlete is also able to identify potential pocket locations on the amino acid sequences, bolstering the molecular interpretability of machine learning prediction. Furthermore, the robustness of ComPPlete under random node and edge removal strengthens the notion of generalizability under data scarcity [112].

Complementarity-driven networks induce even-length cycles [42]. We have developed a novel negative sampling strategy leveraging the 4-cycles in PPIs. This approach can be extended to other complementarity-driven networks by exploring the even-length cycles enriched in the networks. Furthermore, there has been much research in identifying and counting 4-cycles in directed graphs [113,114] and sparse graphs [115]. Integrating these algorithms to the topological PPNI methodology can ameliorate the need for the configuration model and help us derive more hard PPNI samples. We also hypothesize that PPI networks from different species should be treated individually to create topological PPNI. Considering the combined PPI of humans and other species would hinder us from capturing meaningful evolutionary patterns and the complementary nature of the network.

In its exploration of drug-target interaction networks, AI-Bind [53] employed network science to comprehend topological shortcuts and generate negative samples. In contrast, ComPPlete extends the application of network science to ML models operating on protein-protein interaction networks. ComPPlete introduces a novel methodology for negative sampling by leveraging higher-order network properties. Beyond supplying valuable hard negatives applicable to a diverse array of ML models, ComPPlete pioneers a new research direction that advocates the use of network topology in negative sampling. This innovative approach holds promise for enhancing the generalizability, robustness, and interpretability of graph machine learning methodologies on a wide range of networks.

METHODS

Merging and Consolidating PPI Databases

Consolidating the human interactome using the existing experimental databases is challenging [116]. While the majority of the existing PPI databases often report only the gene identifiers (gene symbols [117], Entrez ID [118], Ensembl ID [119], etc.), newer databases like BioPlex [12] report the protein IDS (UniProt ID [25], PDB ID [120], etc). Ambiguity in gene-to-protein mapping combined with the annotation inconsistency in these databases makes the task of

merging the PPI data difficult [121]. Furthermore, the majority of the computational tools for exploratory biology, including network science-inspired disease module identification algorithms [122, 123], demand a coarse-grained representation of the interactome. ML-based PPI prediction methods [55, 59, 93] use UniProt IDs to specify proteins, and hence disregard a big portion of available experimental PPI data. For the aforementioned reasons, we coarse-grain human protein IDs from UniProt to gene Entrez IDs. Establishing a one-to-one correspondence between human genes and amino acid sequences involves selecting the longest amino acid sequence from all protein isoforms associated with a given gene. We have merged multiple experimental PPI databases such as BioPlex [12], STRING [13], APID [14], BioGRID [15], CoFrac [16], CORUM [17], HuRI [18], HINT [19], HIPPIE [20], and BKB [124] to obtain our interaction data.

GoldPPI

To accumulate high-confidence PPI data, we filter the experimental PPI databases for samples with multiple experimental validation. We use three AP-MS databases, which report the number of experiments/publications validating the interaction: APID [14], HINT [19], and HIPPIE [20]. Thereafter, we filter for the PPI samples which have been observed in at least 3 experiments. Thus, we obtain the high confidence PPI, which we use in ComPPlete validation and test datasets.

Negatome PPNI

The majority of the PPI databases do not report the non-interactions. Hence, there is a lack of PPNI data in the literature, which has led the majority of the ML approaches to use random negative sampling or subcellular compartmental negatives. We use high-confidence PPNI from Negatome 2.0 [125] and NVDT [26]. These samples have been used in the validation and test of ComPPlete.

Traditional/Simple Unipartite Configuration Model

The SCM is an exponential random graph model with the probability of observing a graph configuration $P(G) = \frac{B(G)}{Z}$, where $B(G) = \exp(-\sum_{i=1}^{n} \lambda_i d_i(G))$, the Lagrange multipliers $\{\lambda_i\}$ are such that $\langle d_i \rangle = \sum_{j=1}^{n} p_{ij} = k_i$ for all nodes $i \in \{1, 2, \ldots, n\}$, the Boltzmann factor $B = \sum_G B(G)$, $d_i(G) = \sum_{j=1}^{n} G_{ij}$ is the degree of the node i, and $\{G_{ij}\}_{i,j=1}^{n}$ is the adjacency matrix of the training graph G. By entropy maximization, we get the link probability between the nodes i and j as:

$$p_{ij} = \frac{1}{e^{\lambda_i + \lambda_j} + 1}, \tag{1}$$

Data availability

QIAGEN BKB (Biomedical Knowledge Base) PPI data is available upon request to QIAGEN. The ComPPlete predictions will be shared upon reasonable request. The PPIs obtained from WikiPathways, STRING, and BioLip is available on Zenodo.

Code availability

The codes that support the findings of this study are openly available on our GitHub at https://github.com/alxndgb/ComPPlete.

Acknowledgements

We thank the Bioinformatics and Data Science group in Alexion AstraZeneca Rare Disease for funding this project. We are grateful to AstraZeneca's Scientific Computing Platform (SCP) for providing the computing resources. We thank István Kovács, Assistant Professor in the department of Physics and Astronomy at Northwestern University for fruitful discussions on the structural characteristics of PPIs and the extent of incompleteness in the human interactome. We thank Dr. Guillermo del Angel, the Rare Disease TA head at Center for Genomics Research in AstraZeneca for his constructive feedback and highlighting the importance of negative samples in modeling PPI networks. Lastly, We are grateful to Dr. Gyan Srivastava, Associate Director of Alexion Bioinformatics and Data Sciences, and Andrew Proffitt, Research Scientist at Alexion Rare Disease New Target Discovery and Validation group, for their fruitful discussions and guidance.

Author Contributions

A.C. and B.R. conceived and designed the project, wrote the manuscript, conducted data curation and preparation, generated the predictions for the network configuration model, performed experiments to identify networked-derived negatives, implemented negative sample generation, developed and testing of ComPPlete, assessed the robustness of ComPPlete, quantifying the generalizability of PPI predictions, implementation of the interpretability component to predict the active binding locations, and comparisons with SOTA models. N.H.P. prepared the dataset for interpretability analysis by mapping amino-acid sequences of protein pairs to

interaction sites and provided biological insights. M.A. and W.R.M. interpreted the results and offered biological insights. P.R. provided insights into subcellular and random negatives and offered biological insights. Y.L. provided IT support. W.D. interpreted the results and offered biological insights. J.C.M. designed the study on GPCRs, executed AlphaFold validations, conducted computational structural simulations for the interpretability of ComPPlete predictions, and provided chemical and biological insights. T.E.R. provided guidance on designing the machine-learning experiments, interpreting CL3 hypothesis through the lens of complementarity in complex networks, and writing the manuscript.

References

[1] Das, R. *et al.* Assessment of three-dimensional rna structure prediction in casp15. *Proteins: Structure, Function, and Bioinformatics* **91**, 1747–1770 (2023). URL `http://dx.doi.org/10.1002/prot.26602`.

[2] Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-protein interaction detection: Methods and analysis. *International Journal of Proteomics* **2014**, 1–12 (2014). URL `http://dx.doi.org/10.1155/2014/147648`.

[3] Lu, H. *et al.* Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy* **5** (2020). URL `http://dx.doi.org/10.1038/s41392-020-00315-3`.

[4] Stelzl, U. *et al.* A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **122**, 957–968 (2005). URL `http://dx.doi.org/10.1016/j.cell.2005.08.029`.

[5] Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005). URL `http://dx.doi.org/10.1038/nature04209`.

[6] Kuzmanov, U. & Emili, A. Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Medicine* **5**, 37 (2013). URL `http://dx.doi.org/10.1186/gm441`.

[7] Kann, M. G. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Briefings in Bioinformatics* **8**, 333–346 (2007). URL `http://dx.doi.org/10.1093/bib/bbm031`.

[8] Dimitrakopoulos, G. N., Klapa, M. I. & Moschonas, N. K. How far are we from the completion of the human protein interactome reconstruction? *Biomolecules* **12**, 140 (2022). URL https://doi.org/10.3390/biom12010140.

[9] Vidal, M. How much of the human protein interactome remains to be mapped? *Science Signaling* **9** (2016). URL http://dx.doi.org/10.1126/scisignal.aaf6030.

[10] Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B. & Peyvandi, A. A. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol. Hepatol. Bed Bench* **7**, 17–31 (2014).

[11] Jaeger, S. & Aloy, P. From protein interaction networks to novel therapeutic strategies. *IUBMB Life* **64**, 529–537 (2012). URL http://dx.doi.org/10.1002/iub.1040.

[12] Huttlin, E. L. *et al.* The BioPlex network: A systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015). URL https://doi.org/10.1016/j.cell.2015.06.043.

[13] Szklarczyk, D. *et al.* Correction to 'the STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets'. *Nucleic Acids Research* **49**, 10800–10800 (2021). URL https://doi.org/10.1093/nar/gkab835.

[14] Alonso-López, D. *et al.* APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database* **2019** (2019). URL https://doi.org/10.1093/database/baz005.

[15] Stark, C. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34**, D535–D539 (2006). URL https://doi.org/10.1093/nar/gkj109.

[16] O'Reilly, F. J. *et al.* Protein complexes in cells by ai-assisted structural proteomics. *Molecular Systems Biology* **19** (2023). URL https://doi.org/10.15252/msb.202311544.

[17] Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research* **47**, D559–D563 (2018). URL https://doi.org/10.1093/nar/gky973.

[18] Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020). URL https://doi.org/10.1038/s41586-020-2188-x.

[19] Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology* **6**, 92 (2012). URL https://doi.org/10.1186/1752-0509-6-92.

[20] Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research* **45**, D408–D414 (2016). URL https://doi.org/10.1093/nar/gkw985.

[21] Dunham, W. H., Mullin, M. & Gingras, A. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *PROTEOMICS* **12**, 1576–1590 (2012). URL http://dx.doi.org/10.1002/pmic.201100523.

[22] Coates, P. & Hall, P. The yeast two-hybrid system for identifying protein–protein interactions. *The Journal of Pathology* **199**, 4–7 (2002). URL http://dx.doi.org/10.1002/path.1267.

[23] Robinson, J. D., Chuang, C.-Y., Sra, S. & Jegelka, S. Contrastive learning with hard negative samples. In *International Conference on Learning Representations* (2021). URL https://openreview.net/forum?id=CR1XOQ0UTh-.

[24] Blohm, P. *et al.* Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research* **42**, D396–D400 (2013). URL https://doi.org/10.1093/nar/gkt1079. https://academic.oup.com/nar/article-pdf/42/D1/D396/3563500/gkt1079.pdf.

[25] Consortium, U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2020). URL https://doi.org/10.1093/nar/gkaa1100.

[26] Zhao, N., Zhuo, M., Tian, K. & Gong, X. Protein–protein interaction and non-interaction predictions using gene sequence natural vector. *Communications Biology* **5** (2022). URL https://doi.org/10.1038/s42003-022-03617-0.

[27] Srivastava, A., Mazzocco, G., Kel, A., Wyrwicz, L. S. & Plewczynski, D. Detecting reliable non interacting proteins (NIPs) significantly enhancing the computational prediction of protein–protein interactions using machine learning methods. *Molecular BioSystems* **12**, 778–785 (2016). URL https://doi.org/10.1039/c5mb00672d.

[28] You, Z.-H., Lei, Y.-K., Gui, J., Huang, D.-S. & Zhou, X. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* **26**, 2744–2751 (2010). URL http://dx.doi.org/10.1093/bioinformatics/btq510.

[29] Xia, J.-F., Zhao, X.-M. & Huang, D.-S. Predicting protein–protein interactions from protein sequences using meta predictor. *Amino Acids* **39**, 1595–1599 (2010). URL http://dx.doi.org/10.1007/s00726-010-0588-1.

[30] Xia, J.-F., Han, K. & Huang, D.-S. Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein & Peptide Letters* **17**, 137–145 (2010). URL http://dx.doi.org/10.2174/092986610789909403.

[31] Ben-Hur, A. & Noble, W. S. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* **7** (2006). URL https://doi.org/10.1186/1471-2105-7-s1-s2.

[32] et al., P. J. T. A subcellular map of the human proteome. *Science* **356** (2017). URL https://doi.org/10.1126/science.aal3321.

[33] Jansen, R. *et al.* A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003). URL http://dx.doi.org/10.1126/science.1087361.

[34] Jansen, R. & Gerstein, M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current Opinion in Microbiology* **7**, 535–545 (2004). URL http://dx.doi.org/10.1016/j.mib.2004.08.012.

[35] Gomez, S. M., Noble, W. S. & Rzhetsky, A. Learning to predict protein–protein interactions from protein sequences. *Bioinformatics* **19**, 1875–1881 (2003). URL http://dx.doi.org/10.1093/bioinformatics/btg352.

[36] Ben-Hur, A. & Noble, W. S. Kernel methods for predicting protein-protein interactions. *Bioinformatics* **21**, i38–i46 (2005). URL http://dx.doi.org/10.1093/bioinformatics/bti1016.

[37] Zhang, L. V., Wong, S. L., King, O. D. & Roth, F. P. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* **5**, 38 (2004). URL http://dx.doi.org/10.1186/1471-2105-5-38.

[38] QI, Y., KLEIN-SEETHARAMAN, J. & BAR-JOSEPH, Z. Random forest similarity for protein-protein interaction prediction from multiple sources. In *Biocomputing 2005* (WORLD SCIENTIFIC, 2004). URL http://dx.doi.org/10.1142/9789812702456_0050.

[39] Wang, H., Segal, E., Ben-Hur, A., Koller, D. & Brutlag, D. Identifying protein-protein interaction sites on a genome-wide scale. In Saul, L., Weiss, Y. & Bottou, L. (eds.) *Advances in Neural Information Processing Systems*, vol. 17 (MIT Press, 2004). URL https://proceedings.neurips.cc/paper_files/paper/2004/file/dc2b690516158a874dd8aabe1365c6a0-Paper.pdf.

[40] Martin, S., Roe, D. & Faulon, J.-L. Predicting protein–protein interactions using signature products. *Bioinformatics* **21**, 218–226 (2004). URL http://dx.doi.org/10.1093/bioinformatics/bth483.

[41] Bardes, A., Ponce, J. & LeCun, Y. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations* (2022). URL https://openreview.net/forum?id=xm6YD62D1Ub.

[42] Budel, G. & Kitsak, M. Complementarity in complex networks (2023). 2003.06665.

[43] Li, J. *et al.* Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2023). URL https://openreview.net/forum?id=YdjWXrdOTh.

[44] Peng, X., Wang, J., Peng, W., Wu, F.-X. & Pan, Y. Protein–protein interactions: detection, reliability assessment and applications. *Briefings in Bioinformatics* bbw066 (2016). URL http://dx.doi.org/10.1093/bib/bbw066.

[45] Grassmann, G. *et al.* Electrostatic complementarity at the interface drives transient protein-protein interactions. *Scientific Reports* **13** (2023). URL http://dx.doi.org/10.1038/s41598-023-37130-z.

[46] Jernigan, R. L., Khade, P., Kumar, A. & Kloczkowski, A. *Using Surface Hydrophobicity Together with Empirical Potentials to Identify Protein–Protein Binding Sites: Application to the Interactions of E-cadherins*, 41–50 (Springer US, 2022). URL http://dx.doi.org/10.1007/978-1-0716-1546-1_3.

[47] Chen, Y., Wang, W., Liu, J., Feng, J. & Gong, X. Protein interface complementarity and gene duplication improve link prediction of protein-protein interaction network. *Frontiers in Genetics* **11** (2020). URL http://dx.doi.org/10.3389/fgene.2020.00291.

[48] Veselovsky, A. V. *et al.* Protein–protein interactions: mechanisms and modification by drugs. *Journal of Molecular Recognition* **15**, 405–422 (2002). URL http://dx.doi.org/10.1002/jmr.597.

[49] Yang, Z. *et al.* Understanding negative sampling in graph representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, 1666–1676 (Association for Computing Machinery, New York, NY, USA, 2020). URL https://doi.org/10.1145/3394486.3403218.

[50] Chatterjee, A., Walters, R., Menichetti, G. & Eliassi-Rad, T. Inductive link prediction in static and temporal graphs for isolated nodes. In *Temporal Graph Learning Workshop @ NeurIPS 2023* (2023). URL https://openreview.net/forum?id=DRrSYKNhD1.

[51] Chatterjee, A., Walters, R., Menichetti, G. & Eliassi-Rad, T. Disentangling node attributes from graph topology for improved generalizability in link prediction (2023). URL https://doi.org/10.48550/arXiv.2307.08877.

[52] Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673 (2020). URL https://doi.org/10.1038/s42256-020-00257-z.

[53] Chatterjee, A. *et al.* Improving the generalizability of protein-ligand binding predictions with AI-bind. *Nature Communications* **14** (2023). URL https://doi.org/10.1038/s41467-023-37572-z.

[54] Gillis, J., Ballouz, S. & Pavlidis, P. Bias tradeoffs in the creation and analysis of protein–protein interaction networks. *Journal of Proteomics* **100**, 44–54 (2014). URL http://dx.doi.org/10.1016/j.jprot.2014.01.020.

[55] Jha, K., Saha, S. & Singh, H. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports* **12**, 8360 (2022).

[56] Hu, F., Jiang, J., Wang, D., Zhu, M. & Yin, P. Multi-PLI: interpretable multi-task deep learning model for unifying protein–ligand interaction datasets. *Journal of Cheminformatics* **13** (2021). URL https://doi.org/10.1186/s13321-021-00510-6.

[57] Bonner, S., Kirik, U., Engkvist, O., Tang, J. & Barrett, I. P. Implications of topological imbalance for representation learning on biomedical knowledge graphs. *Briefings in Bioinformatics* **23** (2022). URL https://doi.org/10.1093/bib/bbac279. Bbac279, https://academic.oup.com/bib/article-pdf/23/5/bbac279/45937606/bbac279.pdf.

[58] Ju, M., Zhao, T., Yu, W., Shah, N. & Ye, Y. Graphpatcher: Mitigating degree bias for graph neural networks via test-time augmentation. In *Thirty-seventh Conference on Neural Information Processing Systems* (2023). URL https://openreview.net/forum?id=puupdGOWUp.

[59] Szymborski, J. & Emad, A. RAPPPID: towards generalizable protein interaction prediction with AWD-LSTM twin networks. *Bioinformatics* **38**, 3958–3967 (2022). URL https://doi.org/10.1093/bioinformatics/btac429. https://academic.oup.com/bioinformatics/article-pdf/38/16/3958/49889973/btac429.pdf.

[60] Albu, A.-I., Bocicor, M.-I. & Czibula, G. Mm-stackens: A new deep multimodal stacked generalization approach for protein–protein interaction prediction. *Computers in Biology and Medicine* **153**, 106526 (2023). URL https://www.sciencedirect.com/science/article/pii/S0010482522012343.

[61] Kun, K. W., Racharak, T., Yiming, C., Cheng, P. & Nguyen, M. L. Kgwe: A knowledge-guided word embedding fine-tuning model. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 1221–1225 (2021).

[62] Yang, Y., Lichtenwalter, R. N. & Chawla, N. V. Evaluating link prediction methods. *Knowledge and Information Systems* **45**, 751–782 (2014). URL http://dx.doi.org/10.1007/s10115-014-0789-0.

[63] Lichtnwalter, R. & Chawla, N. V. Link prediction: Fair and effective evaluation. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 376–383 (2012).

[64] Dunham, B. & Ganapathiraju, M. K. Benchmark evaluation of protein–protein interaction prediction algorithms. *Molecules* **27**, 41 (2021). URL http://dx.doi.org/10.3390/molecules27010041.

[65] Li, D., Jin, R., Gao, J. & Liu, Z. On sampling top-k recommendation evaluation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, 2114–2124 (Association for Computing Machinery, New York, NY, USA, 2020). URL https://doi.org/10.1145/3394486.3403262.

[66] Craswell, N. *Mean Reciprocal Rank*, 1703–1703 (Springer US, 2009). URL http://dx.doi.org/10.1007/978-0-387-39940-9_488.

[67] Dick, K. & Green, J. R. Reciprocal perspective for improved protein-protein interaction prediction. *Scientific Reports* **8** (2018). URL http://dx.doi.org/10.1038/s41598-018-30044-1.

[68] Vihinen, M. How to evaluate performance of prediction methods? measures and their interpretation in variation effect analysis. *BMC Genomics* **13**, S2 (2012). URL http://dx.doi.org/10.1186/1471-2164-13-S4-S2.

[69] Chen, J., Shah, V. & Kyrillidis, A. Negative sampling in semi-supervised learning (2020). 1911.05166.

[70] Gallaire, H. & Minker, J. *Logic and Data Bases* (Springer, New York, NY, 1978).

[71] Kovács, I. A. *et al.* Network-based prediction of protein interactions. *Nature communications* **10**, 1–8 (2019).

[72] Barabási, A.-L. *Network Science* (Cambridge University Press, 2016).

[73] Bianconi, G., Darst, R. K., Iacovacci, J. & Fortunato, S. Triadic closure as a basic generating mechanism of communities in complex networks. *Phys. Rev. E* **90**, 042806 (2014). URL https://link.aps.org/doi/10.1103/PhysRevE.90.042806.

[74] Lobato, G., Mier, P. & Andrade-Navarro, M. The latent geometry of the human protein interaction network. *Bioinformatics* **34**, 2826–2834 (2018). URL https://doi.org/10.1093/bioinformatics/bty206. https://academic.oup.com/bioinformatics/article-pdf/34/16/2826/48918242/bioinformatics_34_16_2826.pdf.

[75] Kitsak, M., Voitalov, I. & Krioukov, D. Link prediction with hyperbolic geometry. *Phys. Rev. Res.* **2**, 043113 (2020). URL https://link.aps.org/doi/10.1103/PhysRevResearch.2.043113.

[76] Zhang, W., Landback, P., Gschwend, A. R., Shen, B. & Long, M. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome biology* **16**, 1–14 (2015).

[77] Buck, L. B. The molecular architecture of odor and pheromone sensing in mammals. *Cell* **100**, 611–618 (2000).

[78] Kaupp, U. B. Olfactory signalling in vertebrates and insects: differences and commonalities. *Nature Reviews Neuroscience* **11**, 188–200 (2010).

[79] Sato, K. *et al.* Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature* **452**, 1002–1006 (2008).

[80] Lee, S.-J., Depoortere, I. & Hatt, H. Therapeutic potential of ectopic olfactory and taste receptors. *Nature Reviews Drug Discovery* **18**, 116–138 (2019).

[81] Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **157**, 105–132 (1982).

[82] Cock, P. J. *et al.* Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (2009).

[83] Ying, Y. & Zhou, D.-X. Online pairwise learning algorithms with kernels (2015). 1502.07229.

[84] Asgari, E. & Mofrad, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE* **10**, e0141287 (2015). URL https://doi.org/10.1371/journal.pone.0141287.

[85] Bairoch, A. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Research* **24**, 21–25 (1996). URL https://doi.org/10.1093/nar/24.1.21.

[86] Liu, Y. & Li, Z. Protein-protein interaction prediction via structure-based deep learning. *bioRxiv* (2023). URL https://www.biorxiv.org/content/early/2023/05/30/2023.05.27.542552. https://www.biorxiv.org/content/early/2023/05/30/2023.05.27.542552.full.pdf.

[87] Jha, K. & Saha, S. Amalgamation of 3d structure and sequence information for protein–protein interaction prediction. *Scientific Reports* **10** (2020). URL http://dx.doi.org/10.1038/s41598-020-75467-x.

[88] Nicholson, D. J. Is the cell really a machine? *Journal of Theoretical Biology* **477**, 108–126 (2019). URL http://dx.doi.org/10.1016/j.jtbi.2019.06.002.

[89] Callaway, E. What's next for alphafold and the ai protein-folding revolution. *Nature* **604**, 234–238 (2022). URL http://dx.doi.org/10.1038/d41586-022-00997-5.

[90] Haykin, S. *Neural networks: a comprehensive foundation* (Prentice Hall PTR, 1994).

[91] Teru, K. K., Denis, E. & Hamilton, W. L. Inductive relation prediction by subgraph reasoning (2019). URL https://arxiv.org/abs/1911.06962.

[92] Prechelt, L. *Early Stopping — But When?*, 53–67 (Springer Berlin Heidelberg, 2012). URL http://dx.doi.org/10.1007/978-3-642-35289-8_5.

[93] Hu, X., Feng, C., Zhou, Y., Harrison, A. & Chen, M. Deeptrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks. *Bioinformatics* **38**, 694–702 (2021). URL http://dx.doi.org/10.1093/bioinformatics/btab737.

[94] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997). URL https://doi.org/10.1162/neco.1997.9.8.1735.

[95] Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **41**, D1096–D1103 (2012). URL https://doi.org/10.1093/nar/gks966.

[96] Zhang, P., Wang, X., Wang, F., Zeng, A. & Xiao, J. Measuring the robustness of link prediction algorithms under noisy environment. *Scientific Reports* **6** (2016). URL https://doi.org/10.1038/srep18881.

[97] Pezeshkpour, P., Tian, Y. & Singh, S. Investigating robustness and interpretability of link prediction via adversarial modifications (2019). 1905.00563.

[98] Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002). URL https://doi.org/10.1126/science.1065103.

[99] Deeds, E. J., Ashenberg, O., Gerardin, J. & Shakhnovich, E. I. Robust protein–protein interactions in crowded cellular environments. *Proceedings of the National Academy of Sciences* **104**, 14952–14957 (2007). URL https://doi.org/10.1073/pnas.0702766104.

[100] Azevedo, H. & Moreira-Filho, C. A. Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma. *Scientific Reports* **5** (2015). URL https://doi.org/10.1038/srep16830.

[101] Huang, C.-H., Fang, J.-F., Tsai, J. J. P. & Ng, K.-L. Topological robustness of the protein-protein interaction networks. In *Proceedings of the 2005 Joint Annual Satellite Conference on Systems Biology and Regulatory Genomics*, RECOMB'05, 166–177 (Springer-Verlag, Berlin, Heidelberg, 2005).

[102] Roberts, E. S. & Coolen, A. C. C. Unbiased degree-preserving randomisation of directed binary networks (2011). 1112.4677.

[103] Meyes, R., Lu, M., de Puiseau, C. W. & Meisen, T. Ablation studies in artificial neural networks (2019). 1901.08644.

[104] Guo, W. *et al.* Dopamine d2 receptors form higher order oligomers at physiological expression levels. *The EMBO Journal* **27**, 2293–2304 (2008). URL http://dx.doi.org/10.1038/emboj.2008.153.

[105] Hobbs, C. A., Bobay, B. G., Thompson, R. J., Perego, M. & Cavanagh, J. Nmr solution structure and dna-binding model of the dna-binding domain of competence protein a. *Journal of Molecular Biology* **398**, 248–263 (2010). URL http://dx.doi.org/10.1016/j.jmb.2010.03.003.

[106] Boulanger, M. J., Bankovich, A. J., Kortemme, T., Baker, D. & Garcia, K. Convergent mechanisms for recognition of divergent cytokines by the shared signaling receptor gp130. *Molecular Cell* **12**, 577–589 (2003). URL https://www.sciencedirect.com/science/article/pii/S1097276503003654.

[107] Yu, J. *et al.* Structural basis of mu-opioid receptor-targeting by a nanobody antagonist (2023). URL http://dx.doi.org/10.1101/2023.12.06.570395.

[108] Rosenbaum, D. M., Rasmussen, S. G. F. & Kobilka, B. K. The structure and function of g-protein-coupled receptors. *Nature* **459**, 356–363 (2009). URL http://dx.doi.org/10.1038/nature08144.

[109] Calebiro, D., Koszegi, Z., Lanoiselée, Y., Miljus, T. & O'Brien, S. G protein-coupled receptor-g protein interactions: a single-molecule perspective. *Physiological Reviews* **101**, 857–906 (2021). URL http://dx.doi.org/10.1152/physrev.00021.2020.

[110] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). URL https://doi.org/10.1038/s41586-021-03819-2.

[111] Evans, R. *et al.* Protein complex prediction with alphafold-multimer (2021). URL http://dx.doi.org/10.1101/2021.10.04.463034.

[112] Bansal, M. A., Sharma, D. R. & Kathuria, D. M. A systematic review on data scarcity problem in deep learning: Solution and applications. *ACM Comput. Surv.* **54** (2022). URL https://doi.org/10.1145/3502287.

[113] Eisenbrand, F. & Grandoni, F. Detecting directed 4-cycles still faster. *Inf. Process. Lett.* **87**, 13–15 (2003). URL https://doi.org/10.1016/S0020-0190(03)00252-7.

[114] Abboud, A., Khoury, S., Leibowitz, O. & Safier, R. Listing 4-cycles (2022). 2211.10022.

[115] Burkhardt, P. & Harris, D. G. Simple and efficient four-cycle counting on sparse graphs (2023). 2303.06090.

[116] Ramani, A. K., Bunescu, R. C., Mooney, R. J. & Marcotte, E. M. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology* **6**, r40 (2005). URL http://dx.doi.org/10.1186/gb-2005-6-5-r40.

[117] Bruford, E. A. *et al.* Guidelines for human gene nomenclature. *Nature Genetics* **52**, 754–758 (2020). URL http://dx.doi.org/10.1038/s41588-020-0669-3.

[118] Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research* **39**, D52–D57 (2010). URL http://dx.doi.org/10.1093/nar/gkq1237.

[119] Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Research* **51**, D933–D941 (2022). URL http://dx.doi.org/10.1093/nar/gkac958.

[120] Pdb database. www.rcsb.org/.

[121] Lehne, B. & Schlitt, T. Protein-protein interaction databases: keeping up with growing interactomes. *Human Genomics* **3** (2009). URL http://dx.doi.org/10.1186/1479-7364-3-3-291.

[122] Ghiassian, S. D., Menche, J. & Barabási, A.-L. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS computational biology* **11**, e1004120 (2015).

[123] Wang, R.-S. & Loscalzo, J. Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. *Journal of Molecular Biology* **430**, 2939–2950 (2018). URL https://www.sciencedirect.com/science/article/pii/S0022283618304273. Theory and Application of Network Biology Toward Precision Medicine.

[124] Qiagen biomedical knowledge base (2023.2).

[125] Blohm, P. *et al.* Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research* **42**, D396–D400 (2013). URL https://doi.org/10.1093/nar/gkt1079.

Table 1: **State-of-the-art PPI Prediction Models in Transductive and Inductive Link Prediction.** In transductive link prediction, nodes are shared between train and test datasets. State-of-the-art PPI prediction models, PPI-GNN and DeepTrio, achieve excellent transductive link prediction performance. However, in inductive tests, when encountered with never-before-seen proteins, the performances of both models diminish significantly.

Model	Transductive		Inductive	
	AUROC	AUPRC	AUROC	AUPRC
PPI-GNN	0.98 ± 0.044	0.98 ± 0.042	0.50 ± 0.00	0.74 ± 0.03
DeepTrio	0.93 ± 0.003	0.94 ± 0.005	0.59 ± 0.06	0.62 ± 0.07

Table 2: **State-of-the-art PPI Prediction Models vs. ComPPlete in Inductive Link Prediction.** Our proposed pipeline ComPPlete combining topological PPNI and unsupervised pre-training on protein representations, significantly improves inductive link prediction performance in PPIs compared to two state-of-the-art PPI prediction models PPI-GNN and DeepTrio.

Model	AUROC	AUPRC	Hits@Top1000	Hits@Top10,000
ComPPlete	0.79 ± 0.007	0.87 ± 0.009	0.92 ± 0.015	0.87 ± 0.010
PPI-GNN	0.50 ± 0.0	0.74 ± 0.03	0.67 ± 0.07	0.06 ± 0.006
DeepTrio	0.59 ± 0.06	0.62 ± 0.07	0.25 ± 0.008	0.024 ± 0.001

Table 3: **ComPPlete vs. RAPPPID in inductive link prediction and transfer learning across protein families and peptides.** RAPPPID is an inductive PPI prediction model which uses sequential models and regularization for protein representations. Although, ComPPlete and RAPPID achieve comparable performances in inductive link prediction, ComPPlete offers stronger prioritization as depicted by Hits@TopK. Furthermore, in the transfer learning setting of training on STRING DB and testing on BioLip, which contains proteins from different families than STRING DB, RAPPPID fails and achieves performance similar to a naive Bayes classifier on unseen data. However, ComPPlete can make meaningful PPI predictions in this transfer learning setting, showing improved generalizability across protein families.

Model	Inductive				Transfer learning
	AUROC	AUPRC	Hits@Top1000	Hits@Top10,000	AUROC
ComPPlete	0.89	0.90	0.98	0.93	0.73
RAPPPID	0.80	0.81	0.77	0.13	0.55

Table 4: **ComPPlete validation with AlphaFold Multimer.** While ComPPlete is trained on heterodimers, we test the transferability of ComPPlete by predicting homodimers. For this experiment, we use understudied GPCR proteins. The top GPCR homodimer predictions from ComPPlete include the protein Fafr, Oxyr, Ssr2, and Mtr. The bottom predictions for self-interaction prediction include the proteins Mcr, Fsh, Npy, and Ccr. We run AlphaFold-Multimer simulations for the listed proteins and summarize the observations in the table below. We observe that, on average, the top self-interaction predictions from ComPPlete show more atoms at the interaction interface and higher pLDDT (confidence) at the interface compared to the bottom predictions. Furthermore, while the average pLDDT values are comparable between top and bottom predictions, interface pLDDT is higher for the top predictions, suggesting that ComPPlete is able to learn the interfaces of interaction leveraging only the amino acid sequences.

Case	Model	Averale pLDDT	Atoms at Interface	Interface pLDDT
Ffar	rank 0	73.98	358	52.54
Oxyr	rank 0	71.79	22	63.76
Ssr2	rank 0	67.09	174	55.6
Mtr	rank 2	82.24	244	58.54
Average for Top	-	73.775	199.5	57.61
Mcr	rank 0	70.42	206	61.99
Fsh	rank 0	75.28	120	70.34
Npy	rank 0	59.09	0	0
Ccr	rank 0	76.67	0	0
Average for Bottom	-	70.365	81.5	33.0825

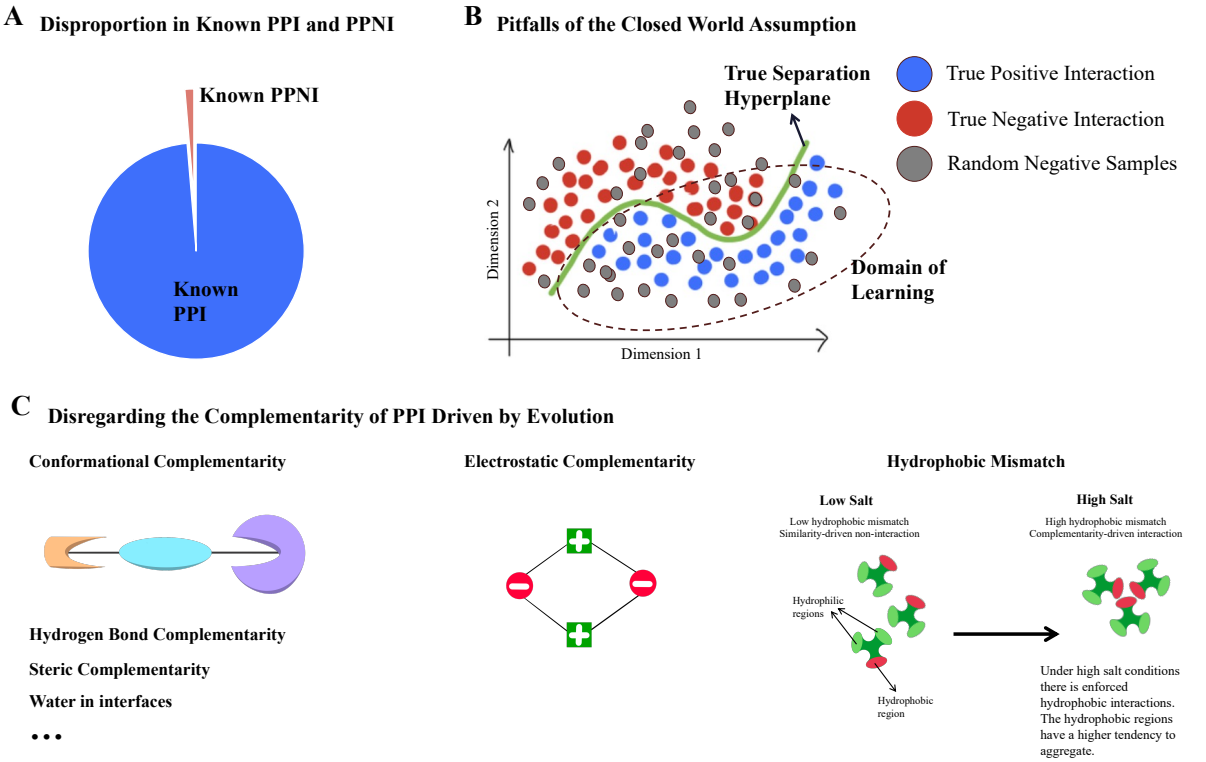


Figure 1

Figure 1: **Status quo in machine learning-based PPI prediction. (a)** We observe a substantially higher number of PPI samples in the existing databases compared to PPNI samples. The lack of true negative samples is a major obstacle in training reliable machine learning models for PPI prediction. **(b)** Traditional machine learning training methodology utilizes random sampling from the complement graph (G^c) of the training PPI network (G) to obtain negative examples. However, this approach violates the closed-world assumption, i.e., not observing a PPI does not imply non-interaction between those proteins. Hence, the traditional approach fails to create high-quality PPNI and shifts the domain of learning in the loss function manifold, hampering the machine learning models from learning the true separation hyperplane between PPI and true PPNI. **(c)** In similarity-driven networks, entities that are alike are connected by links. For example, in social networks, people with similar interests are connected. On the other hand, in complementarity-driven networks, entities with opposing properties are linked. We list six complementary properties of proteins that are major contributors to protein-protein interactions [48]. Leveraging the complementary nature of PPI networks has recently gained much attention and bifurcates the machine learning approaches for PPI prediction from those widely used in social network analysis.

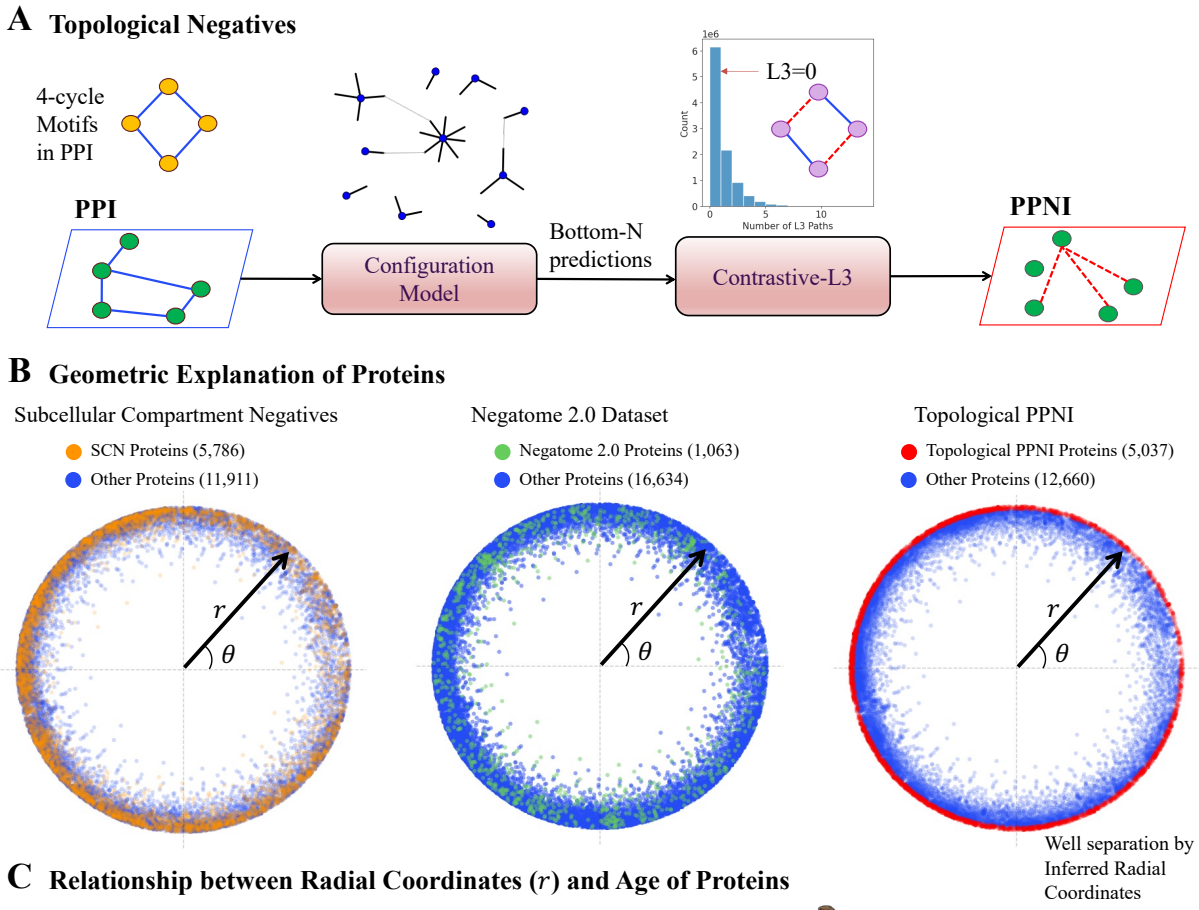

A Topological Negatives
4-cycle Motifs in PPI
PPI
Configuration Model
Bottom-N predictions
L3=0
Count
Number of L3 Paths
Contrastive-L3
PPNI
B Geometric Explanation of Proteins
Subcellular Compartment Negatives
SCN Proteins (5,786)
Other Proteins (11,911)
Negatome 2.0 Dataset
Negatome 2.0 Proteins (1,063)
Other Proteins (16,634)
Topological PPNI
Topological PPNI Proteins (5,037)
Other Proteins (12,660)
r
θ
Well separation by Inferred Radial Coordinates
C Relationship between Radial Coordinates (r) and Age of Proteins
1-Cellular organisms
2-Metazoa
3-Chordata
4-Mammalia
5-Euarchontoglires
6-Primates
0
Older Proteins
Radial Coordinate
r
Newer Proteins

Figure 2

Figure 2: **Leveraging PPI topology for PPNI sampling. (A)** We propose a novel method for sampling high-quality PPNI leveraging the topology of PPI driven by its complementary nature. First, we run a traditional configuration model to identify the topologically least probable edges via entropy maximization. The bottom-N edges are then used for computing the 4-cycles or L3 paths induced by the protein pairs. (We use N=10 million.) The configuration model helps in reducing the protein-pair search space from 156M down to 10M. Interacting pairs induce many L3 paths in the PPI network. We utilize the inverse hypothesis, namely *Contrastive L3 or CL3*, to filter the protein pairs from the bottom-N predictions which induce no L3 path in the PPI network. Hence, we obtain the topological PPNI, which is used in the training and testing of ComPPlete. **(B)** Hyperbolic embeddings have recently been used for visualizing and link prediction tasks in complementarity-driven networks like PPI networks. We visualize the proteins involved in PPI and different PPNIs in the hyperbolic space. We observe that the proteins involved in subcellular compartmental negatives (SCN) and PPNI from the Negatome database show a large overlap with proteins in the PPI. However, the proteins involved in topological PPNI show a clear separability from the proteins in PPI. Proteins involved in topological PPI are towards the circumference of the hyperbolic disc and are hence constituted by evolutionarily younger proteins. **(C)** Alanis-Lobato et al. [74] established the relationship between the radial coordinates $\boldsymbol{r}$ of human proteins and their age by assigning proteins to six different age groups through grouping proteins based on their ancient relatives in other species (subfigure courtesy of Alanis-Lobato et al. [74]). The clear selection of younger proteins by *CL3* hypothesis indicates that *CL3* can only identify negative interactions for younger proteins, which are less central in biological pathways. Hence biology would optimize less competition between the complementarity mechanisms to form a biologically relevant function for younger proteins, leaving patterns in PPI topology that can distinguish younger and older proteins as captured by *CL3*.

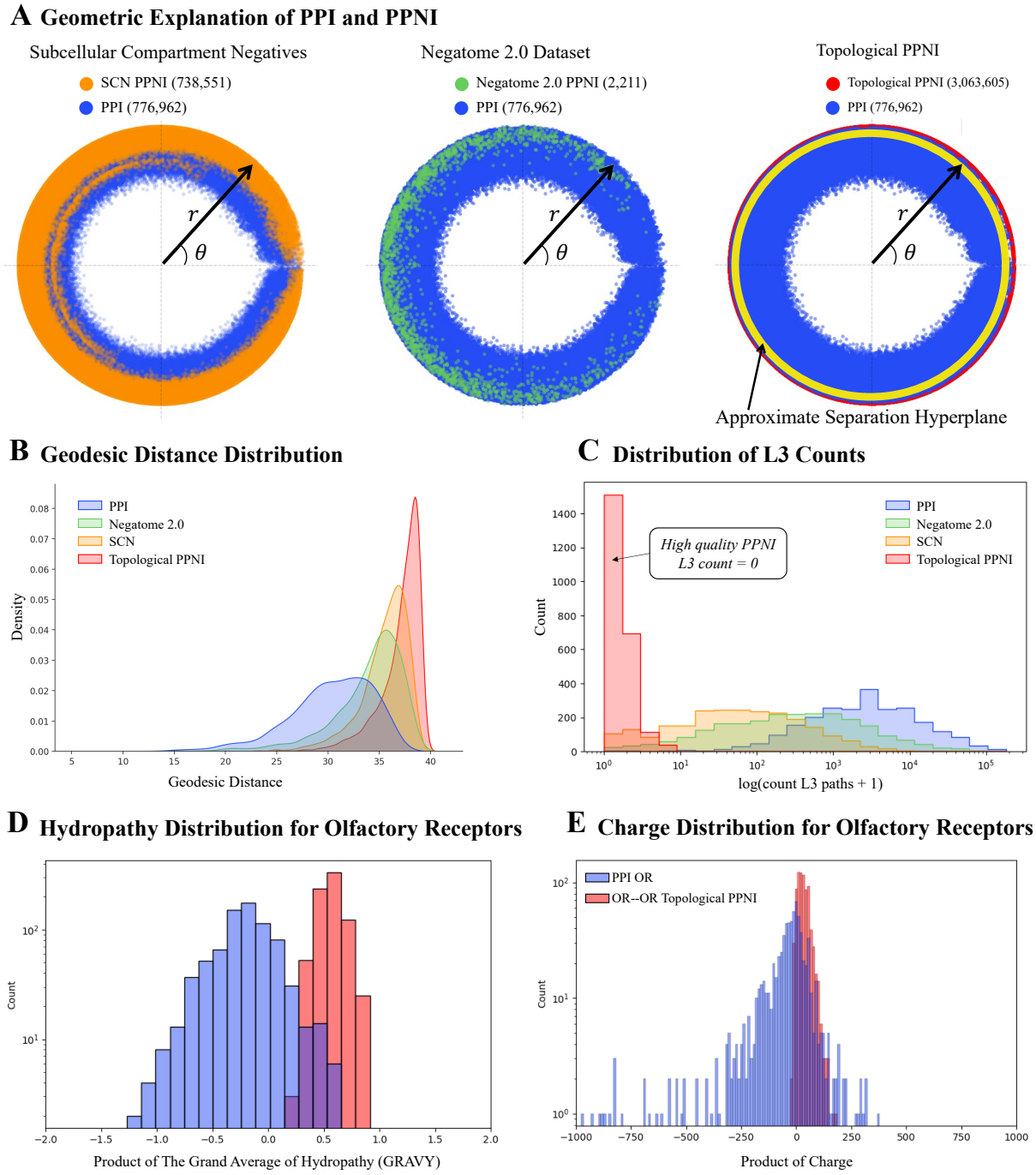

A Geometric Explanation of PPI and PPNI
Subcellular Compartment Negatives
SCN PPNI (738,551)
PPI (776,962)
Negatome 2.0 Dataset
Negatome 2.0 PPNI (2,211)
PPI (776,962)
Topological PPNI
Topological PPNI (3,063,605)
PPI (776,962)
r
θ
Approximate Separation Hyperplane
B Geodesic Distance Distribution
PPI
Negatome 2.0
SCN
Topological PPNI
Density
Geodesic Distance
C Distribution of L3 Counts
High quality PPNI
L3 count = 0
Count
log(count L3 paths + 1)
D Hydropathy Distribution for Olfactory Receptors
Count
Product of The Grand Average of Hydropathy (GRAVY)
E Charge Distribution for Olfactory Receptors
PPI OR
OR--OR Topological PPNI
Count
Product of Charge

Figure 3

Figure 3: **Geometric and Chemical Relevance of topological PPNI. (A)** We visualize each PPI and PPNI in in the hyperbolic space by averaging $\boldsymbol{r}$ and θ for each protein pair. Topological PPNI provides the best-separating hyperplane that can be learned by downstream tasks only from the secondary structure of proteins. **(B)** We plot the distributions of geodesic distances between the protein pairs in PPI and different PPNIs (number of samples for each type n=2,311). Topological PPNI identifies pairs of proteins that are significantly far from each other in the hyperbolic geometry of the human PPI network. Hence, compared to SCN and Negatome, topological PPNI offers negative pairs with unique patterns allowing ML to decode and learn such patterns. **(C)** We plot the L3 count distributions for PPI and various PPNIs. While the protein pairs in PPNI induce a significantly lower number of L3 paths in the PPI network, we select the pairs inducing no L3 path in the topological PPNI for training and testing of ComPPlete. **(D)** The Grand Average of Hydropathy (GRAVY) is a measure of the hydrophobicity or hydrophilicity of a peptide or protein [81]. The more negative the score, the more hydrophilic the amino acid is, and the more positive the score, the more hydrophobic. Humans have approximately 400 olfactory receptors (ORs) and we observed 7,076 negative inter-family links among ORs in topological PPNI. Also, we have 773 positive links between ORs and other proteins in our PPI, none of the 773 PPI OR links are inter-family. The distribution of the product of GRAVY scores between PPI ORs and OR–OR PPNI (randomly sampled 773 topological PPNI) shows that the *CL3* resulted in identifying pairs of ORs that show higher hydrophobic mismatch, and hence decrease the chance of interacting with each other. **(E)** Similarly, we calculated the product of charge for the 773 PPI OR links and OR–OR topological PPNI (charge at $pH = 7$). The distribution of the product of charge shows that *CL3* identified inter-family OR proteins with the same sign of charge hence repelling each other and decreasing the chance of interaction.

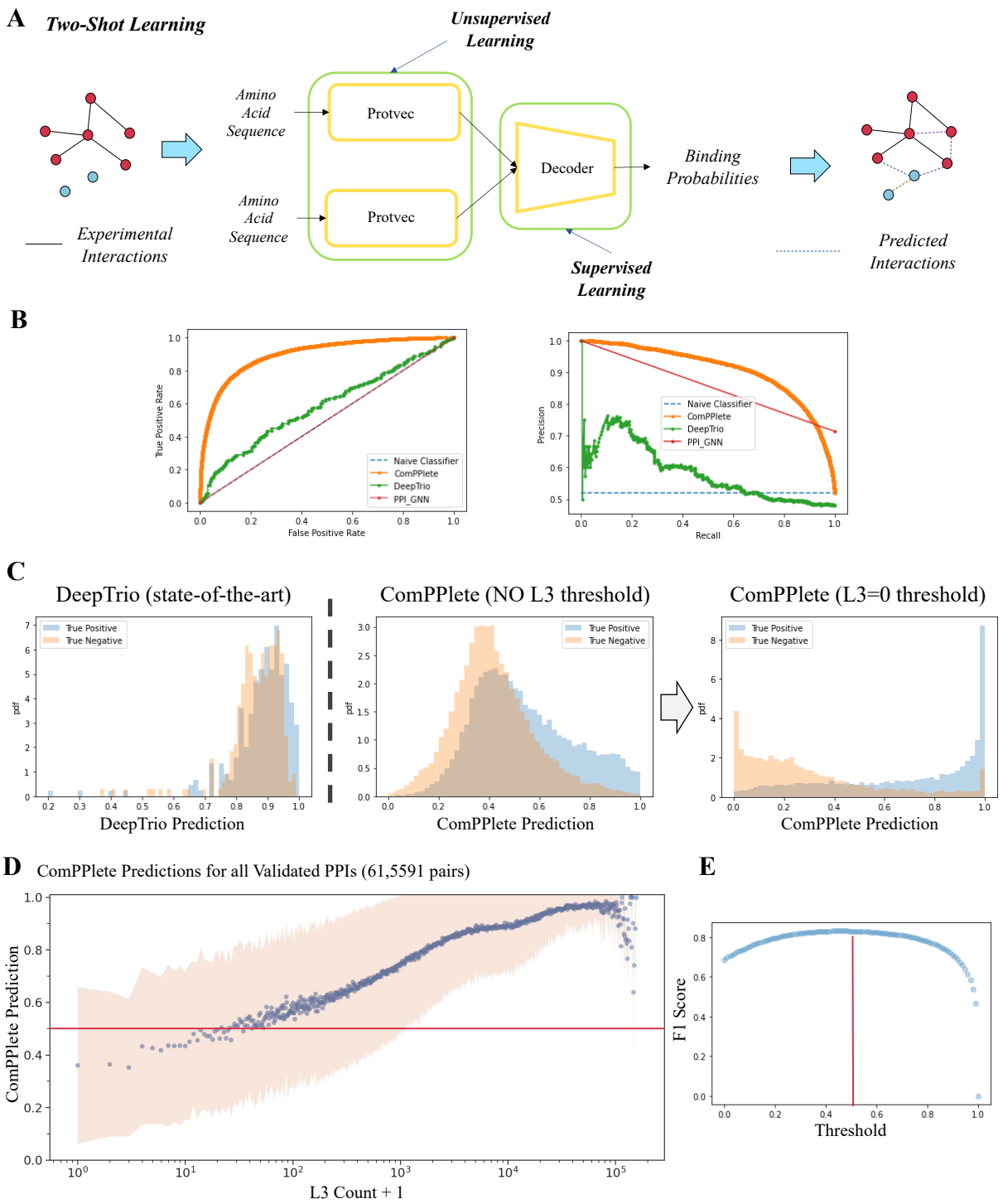

A
Two-Shot Learning
Unsupervised Learning
Amino Acid Sequence
Protvec
Amino Acid Sequence
Protvec
Decoder
Binding Probabilities
Experimental Interactions
Supervised Learning
Predicted Interactions
B
True Positive Rate
False Positive Rate
Naive Classifier
ComPPlete
DeepTrio
PPI_GNN
Precision
Recall
C
DeepTrio (state-of-the-art)
ComPPlete (NO L3 threshold)
ComPPlete (L3=0 threshold)
True Positive
True Negative
pdf
DeepTrio Prediction
ComPPlete Prediction
D
ComPPlete Predictions for all Validated PPIs (61,5591 pairs)
ComPPlete Prediction
L3 Count + 1
E
F1 Score
Threshold

Figure 4

Figure 4: **ComPPlete architecture, inductive link prediction, and transfer learning.** **(A)** ComPPlete architecture: ComPPlete embeds the protein amino acid sequences into 100-dimensional vectors using ProtVec. For each protein pair, ComPPlete concatenates the ProtVec embeddings and feeds it to a decoder (multi-layer perceptron). **(B)** ComPPlete achieves better performance both in terms of receiver operating characteristics and precision-recall compared to two state-of-the-art models PPI-GNN and DeepTrio in inductive link prediction. **(C)** DeepTrio predicts overlapping prediction values for the PPI and PPNI test samples. ComPPlete shows better separation between the predictions for PPIs and PPNIs, becoming a superior ranking tool for both PPI and PPNI. **(C)** Predictions at the output of DeepTrio are overlapping for the PPI and PPNI data. Hence, DeepTrio is unable to separate the predictions from PPIs vs. PPNIs. ComPPlete trained on potential negatives from the configuration model (no *CL3*) shows better separation between the predictions for PPIs vs. PPNIs. Finally, after introducing CL3 thresholding, ComPPlete shows a clear separation between the predictions for PPIs and PPNIs. **(D)** We computed the number of L3 simple paths between all PPI pairs to investigate if ComPPlete has learned the complementarity mechanisms from protein sequences. Indeed, we observe a strong Spearman's rank correlation coefficient of 0.48 between L3 counts and ComPPlete predictions, indicating that the topological PPNI enforced ComPPlete to learn complementary mechanisms that drive protein-protein interactions only from the amino acid sequences. **(E)** We plot the F1-score on the test dataset (first fold) by changing the binary classification threshold. We observe that the test F1-score is maximized for an optimal classification threshold ≈ 0.5. Furthermore, in subplot (D), we observe that ComPPlete predicts interaction probabilities greater than the optimal interaction threshold for the majority of the experimentally validated PPIs. Similar observations are made for other folds of ComPPlete and the average optimal threshold is 0.476 ± 0.049.

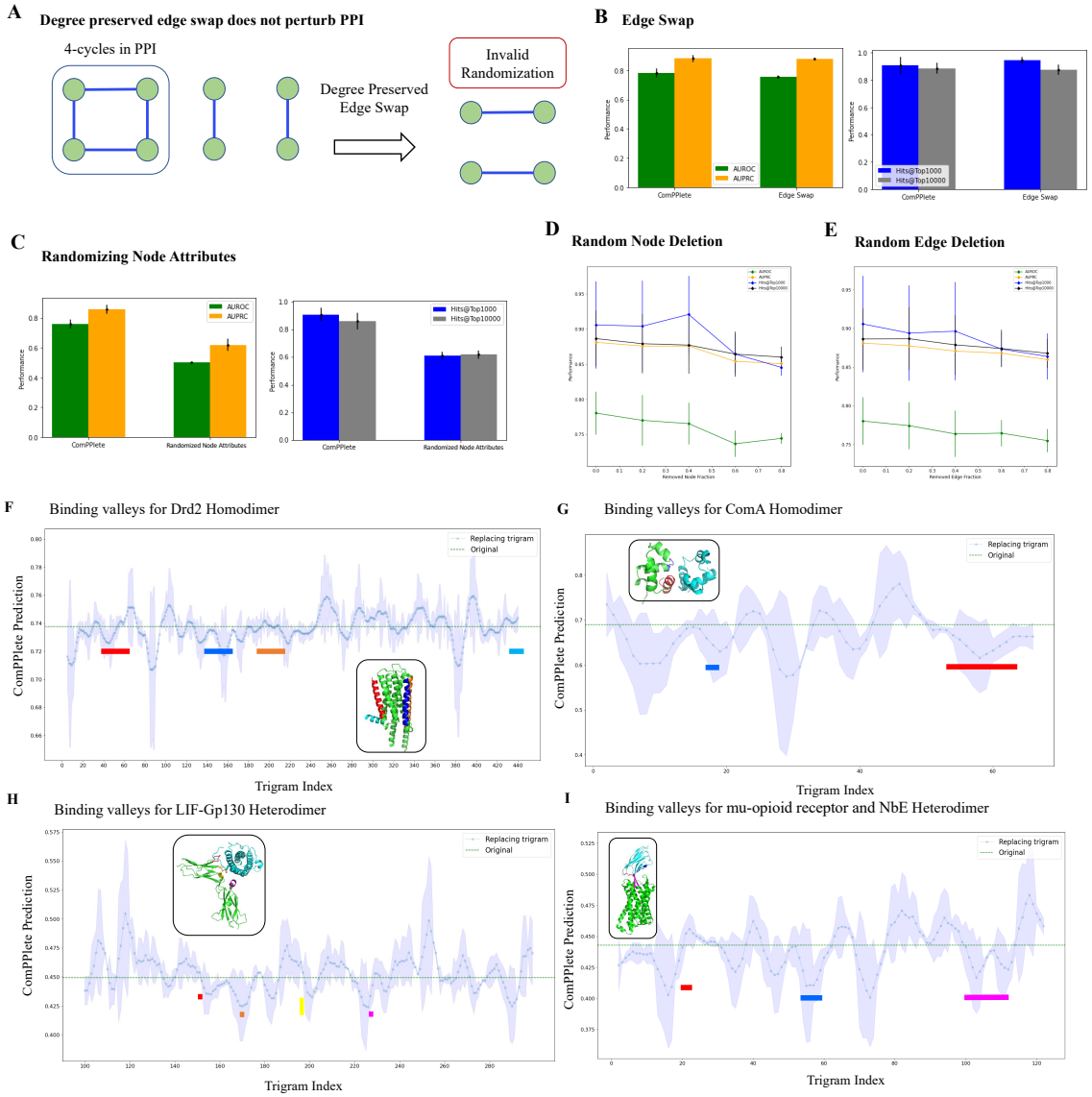

A
Degree preserved edge swap does not perturb PPI
4-cycles in PPI
Degree Preserved Edge Swap
Invalid Randomization
B
Edge Swap
AUROC
AUPRC
Hits@Top1000
Hits@Top10000
ComPPlete
Edge Swap
Performance
C
Randomizing Node Attributes
Randomized Node Attributes
D
Random Node Deletion
E
Random Edge Deletion
F
Binding valleys for Drd2 Homodimer
G
Binding valleys for ComA Homodimer
H
Binding valleys for LIF-Gp130 Heterodimer
I
Binding valleys for mu-opioid receptor and NbE Heterodimer
Replacing trigram
Original
ComPPlete Prediction
Trigram Index

Figure 5

Figure 5: **Robustness and Interpretability of ComPPlete. (A)** Degree preserved edge swap is a widely used method to randomize graphs for studying robustness. However, since PPI networks are enriched with 4-cycles, degree preserved randomization frequently duplicates existing interactions in PPI instead of creating random edges. **(B)** In the robustness study we perturb train and validation datasets while keeping the test dataset unchanged. The inductive performance of ComPPlete is invariant to degree-preserved randomization. **(C)** We replace ProtVec embeddings with 100-dimensional random vectors where the value for each dimension is randomly selected from a uniform distribution $U[0, 1]$. We observe inductive performance similar to a naive Bayes classifier under such randomization, confirming that ComPPlete learns interactions leveraging the embeddings of the amino acid sequences. **(D)** We randomly remove nodes (proteins) from the training and validation datasets. We do not observe significant fluctuation in ComPPlete inductive performance, which confirms the ability of ComPPlete to learn from limited data and generalize to new proteins. **(E)** Similar to (D), we randomly remove edges (interactions) from the training and validation datasets. We do not observe significant fluctuation in ComPPlete inductive performance, which confirms the ability of ComPPlete to learn from limited data and generalize to new proteins. **(F)** We ran an ablation study on the amino acid sequence of DRD2 to identify potential trigrams where interaction takes place for creating Drd2 homodimer. In this ablation study, each amino acid trigram is replaced with the Out-of-Vocabulary embedding from ProtVec, while keeping the other amino acid sequence in the input of ComPPlete unchanged. We observe multiple valleys in the binding probability profiles. These valleys correspond to the interfeces TM4/TM5 and TM1/H8 which have been identified experimentally using Cys-crosslinking and FRET. **(G)** We repeat a similar process for another homodimer of transcription factor ComA. **(H)-(I)** We run abltation study on two heterodimers consisting of protein pairs LIF-Gp130 and Mu-opioid receptor-NbE. The interaction locations on protein complexes LIF-Gp130 and mu-Opioid receptor-NbEHH are marked in the figures, which overlap with valleys predicted by ComPPlete. In all of the above scenarios, we observe that the valleys with lower standard deviation obtained from 5-folds of ComPPlete correspond to the true binding locations. Therefore, the valleys on which 5-fold of ComPPlete agree are the binding locations with a higher confidence.

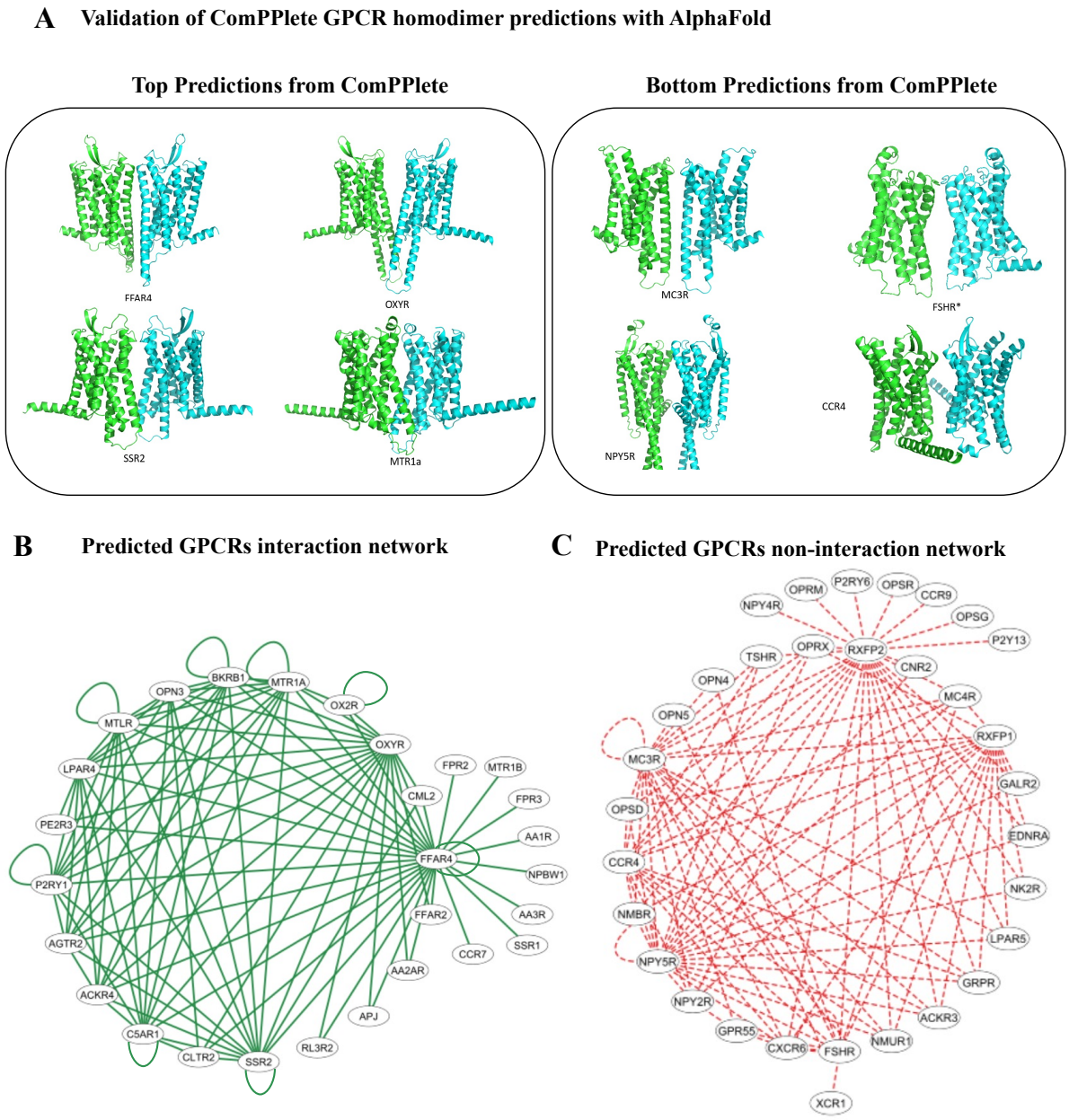


Figure 6

Figure 6: **ComPPlete Predictions for G-protein coupled receptors (GPCRs). (A)** We validate top and bottom self-interaction predictions for GPCRs from ComPPlete with AlphaFold multimer. We observe that AlphaFold predicts a higher number of atoms at the interaction surface and pLDDT at the surface for the top predictions compared to the bottom predictions (see Table 4), which validates the agreement between ComPPlete and AlphaFold. **(B)** Interaction network of GPCRs. From top 100 ComPPlete predictions we construct the interaction network of GPCRs. We predict 8 homodimers and 92 heterodimers involving 28 GPCR proteins. **(C)** Similarly, we construct a non-interaction network with bottom 100 predictions from ComPPlete.

Supplementary Information

Topology-Driven Negative Sampling Enhances Generalizability in Protein-Protein Interaction Prediction

Ayan Chatterjee[1,2,†] , Babak Ravandi[1,2,3,†,*], Naomi H. Philip[1], Mario Abdelmessih[1], William R. Mowrey[1], Piero Ricchiuto[1], Yupu Liang[1], Wei Ding[1], Juan C. Mobarec[4], Tina Eliassi-Rad[2,5,6,7]

[1]Alexion AstraZeneca Rare Disease, Boston, MA, USA
[2]Network Science Institute, Northeastern University, Boston, MA, USA
[3]Department of Physics, Northeastern University, Boston, MA, USA
[4]Mechanistic and Structural Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
[5]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[6]Santa Fe Institute, Santa Fe, NM, USA
[7]The Institute for Experiential AI, Northeastern University, Boston, MA, USA

[*]Corresponding author e-mail: Babak.Ravandi@Alexion.com

Contents

1 Revisiting Topological Shortcuts in PPI Network **1**
1.1 State-of-the-art PPI Prediction Models in Transductive Tests 2
1.2 Sealing Information Leakage, Inductive Tests, and Generalizability 3
2 Unipartite Duplex Configuration Model **4**
2.1 Overview 4
2.2 Mathematical Formulation 4
3 Annotation Imbalance **6**
4 Coefficient of Variation for GPCR Predictions. **6**
5 Alternate protein Embeddings **6**

1 Revisiting Topological Shortcuts in PPI Network

Most real-world networks, such as Protein-Protein Interaction (PPI) [1] and Drug-Target Interaction (DTI) [2] networks, typically exhibit a degree distribution that follows a heavy-tailed distribution [3]. This means that the majority of nodes in these networks have only a small number of links (interactions or samples) in the training datasets, while a small number of hub nodes have a large number of connections. During the training phase of a link prediction model, the model tends to focus more on these hub nodes than on the nodes with lower degrees. In traditional machine learning scenarios, train and test datasets are created by randomly dividing

[†]Ayan Chatterjee and Babak Ravandi equally contributed.

the edges within the network [4]. As a result, the majority of edges in both the training and test datasets involve these hub nodes. Machine learning models become proficient at learning the neighborhood characteristics of these hub nodes in terms of their degree information [5,6]. Consequently, they can make accurate predictions for the edges associated with these hubs, leading to excellent test performance. Figure S2 demonstrates the need for the inductive scenario in link prediction tasks. The transductive scenario, most frequently used to evaluate the performance of AI models, favors learning topological shortcuts instead of the Mechanisms of Actions (MoA) behind the emergence of structural topology in networked data (Figure S2B). In other words, state-of-the-art (SOTA) models leverage the neighborhood topology of a node (protein or drug) to make new interaction predictions, which are often truant of biological and molecular interpretability. However, in an inductive setting (see Figure S2A), the SOTA models struggle when dealing with the low-degree nodes, since they are forced to leverage the molecular patterns for learning interactions. Hence, hubs play the main role in the misleading high performance of the SOTA models. However, in Erdős-Rényi (ER) graphs [7], there are no hubs, and all nodes are indistinguishable in terms of their neighborhood degrees (see Figure S2B). Thus, link prediction models achieve performance similar to a naive Bayes classifier (AUROC ≈ 0.5) when trained and tested on ER graphs (see SI Section 1 of [5]). Furthermore, AI models trained under transductive scenarios are unable to perform well even in transductive link prediction tasks for low-degree nodes [5]. This highlights the models' limitations in extracting meaningful patterns from the molecular structures of proteins, resulting in poor inductive performance, revealing the misleading nature of transductive tests, and insufficient biological interpretability of the machine learning predictions. Likewise, current machine learning models demonstrate outstanding effectiveness when analyzing common diseases that benefit from abundant data available in established databases [8]. Nonetheless, when it comes to rare diseases [9], the limited data availability restricts the utility of these models in making valuable predictions (see Figure 2C).

1.1 State-of-the-art PPI Prediction Models in Transductive Tests

Protein-protein interactions (PPIs) exhibit a complex hierarchy of generation mechanisms. At the top level, we have Preferential Attachment (PA) [3], a concept that plays a crucial role in a link prediction model's ability to excel when it comes to predicting connections involving hub proteins. However, it is important to note that PA itself arises as a consequence of certain proteins having a greater number of binding pockets and findings in experimental studies. SOTA models, despite their sophistication, often fall short of understanding the true underlying mechanisms by leveraging the molecular structures of proteins. It's during inductive tests that

these models are compelled to uncover this genuine generative process, which, in turn, enhances the model's interpretability. This assertion is substantiated through an experiment involving the introduction of random amino acid sequences onto a DTI network, where the state-of-the-art binding prediction model DeepPurpose [10] model achieves exceptional transductive performance, showcasing the capacity to discern the authentic generation mechanism.

In Figure S3B we visualize the heavy-tailed degree distribution of the QIAGEN BKB PPI [11]. We consider two state-of-the-art PPI prediction models, DeepTrio [12] and PPI-GNN [13]. We develop a unipartite duplex configuration model (see Section S2) [14], which takes as input only the degree sequences from the protein-protein interaction and non-interactions (random negative sampling) used in the training of DeepTrio and PPI-GNN (see Figure S3B). We observe that using only the degree information of the proteins and being completely blind to the protein structures (amino acid sequences), the configuration model achieves comparable performance to the SOTA models (Figure S3C). This experiment confirms that the SOTA models are unable to learn from the protein amino acid sequences and resort to topological shortcuts [5], hence lack biological interpretability in the predictions.

1.2 Sealing Information Leakage, Inductive Tests, and Generalizability

Following the common formulation of GraIL [15], Park et al. [16], and Chatterjee et al. [6,17], we evaluate the performance of the SOTA models in different test scenarios. We have already seen that in the transductive setting, proteins are shared between train and test datasets in a transductive setting (Figure S2A). In a semi-inductive setting, one protein of the test edge is present in train data, and the other one is unseen during training. In an inductive link prediction setting, both proteins of the test edge are absent in training. In Table 1 in the main text, we observe a gradual lowering of the test performance of the SOTA models as we move from transductive to inductive tests.

This observation is primarily driven by the reduction of topological information leakage between train and test data. To develop a generalizable AI model, learning PPI from the protein structures instead of leveraging the topology that inflates the performance, we need to mitigate shortcut learning and make protein representation learning independent of the PPI training data. Furthermore, regularization has been proven to be an important tool for improving generalizability in computer vision [18]. Yet, regularization is based on the assumption that similarity between a pair of entities implies interaction, an assumption that is true in image processing and social sciences [19], yet similar molecules do not necessarily always interact in PPI and other biological systems [20]. Although [19] is applicable for identifying similarity between images, the similarity of proteins does not imply their interaction, and hence the regularization-

based approach to overcome negative sampling is not applicable in PPI prediction. Thus, we leverage the network topology of the PPI network to create high-quality PPNI for training our AI pipeline.

2 Unipartite Duplex Configuration Model

2.1 Overview

Protein-protein annotations naturally form an integral part of a unipartite duplex network, which can be conceptualized as a network comprising a set of nodes representing all proteins. This network operates on two distinct layers, with each layer corresponding to a specific type of interaction occurring between the same pair of nodes, as described in Menichetti et al. [14]. Specifically, Layer 1 characterizes positive interactions, while Layer 2 records negative interactions (see Figure S3B).

The multilink notation, denoted as $\mathbf{m}$, is employed to encode the pattern of links connecting two nodes across different layers. Notably, $\mathbf{m} = (1, 0)$ signifies positive interactions, $\mathbf{m} = (0, 1)$ represents negative interactions, $\mathbf{m} = (0, 0)$ signifies the absence of any interaction, and $\mathbf{m} = (1, 1)$ is mathematically prohibited, as it is impossible for both positive and negative interactions to coexist for the same pair of proteins.

In our analysis, we utilize the canonical unipartite duplex null model, which ensures the conservation, on average, of the number of positive and negative annotations associated with each node. This model also appropriately rewires positive and negative links while avoiding the occurrence of forbidden configurations. By applying entropy maximization with constraints, we derive analytical expressions for the probability of each multilink configuration and the conditional probability of observing a positive binding event once an annotation is reported.

2.2 Mathematical Formulation

Let $A_{ij}^{\mathbf{m}}$ denote the multi-adjacency matrix representing the unipartite duplex network of proteins $\{i\}$ and $\{j\}$, with elements equal to 1 indicating the presence of a multilink $\mathbf{m}$ between proteins i and j, and 0 otherwise. We define the multidegree of protein i as:

$$k_i^{\mathbf{m}} = \sum_{j=1}^{N} A_{ij}^{\mathbf{m}} \tag{2}$$

Here, N represents the total number of proteins.

A unipartite duplex network ensemble encompasses all duplexes that adhere to specific constraints, such as the expected multidegree sequences defined in Equation (S2). The probability of observing a unipartite duplex network, denoted as $P(\vec{G})$, is determined using entropy

maximization with multidegree constraints $\{k_i^{(1,0)}\}$ and $\{k_i^{(0,1)}\}$, along with their corresponding Lagrangian multipliers $\{\lambda_i^{(1,0)}\}$ and $\{\lambda_i^{(0,1)}\}$, as outlined in [14, 21]. The probability $P(\vec{G})$ factorizes as follows:

$$P(\vec{G}) = \frac{1}{Z} \prod_{ij} \exp\left[- \sum_{\mathbf{m} \neq (0,0),(1,1)} (\lambda_i^{\mathbf{m}} + \lambda_i^{\mathbf{m}}) A_{ij}^{\mathbf{m}} \right] \tag{3}$$

Where:

$$Z = \prod_{ij} \left[1 + \sum_{\mathbf{m} \neq (0,0),(1,1)} e^{-(\lambda_i^{\mathbf{m}} + \lambda_j^{\mathbf{m}})} \right] \tag{4}$$

The multilink probabilities $p_{ij}^{\mathbf{m}}$ are determined by the derivatives of the logarithm of Z with respect to $(\lambda_i^{\mathbf{m}} + \lambda_j^{\mathbf{m}})$. For instance, the probability of observing a positive annotation is expressed as:

$$p_{ij}^{(1,0)} = \frac{e^{-(\lambda_i^{(1,0)} + \lambda_j^{(1,0)})}}{1 + e^{-(\lambda_i^{(1,0)} + \lambda_j^{(1,0)})} + e^{-(\lambda_i^{(0,1)} + \lambda_j^{(0,1)})}} \tag{5}$$

Similarly, the probability of observing a negative annotation is given by:

$$p_{ij}^{(0,1)} = \frac{e^{-(\lambda_i^{(0,1)} + \lambda_j^{(0,1)})}}{1 + e^{-(\lambda_i^{(1,0)} + \lambda_j^{(1,0)})} + e^{-(\lambda_i^{(0,1)} + \lambda_j^{(0,1)})}} \tag{6}$$

It is important to note that $p_{ij}^{(1,0)} + p_{ij}^{(0,1)} + p_{ij}^{(0,0)}$ equals 1, indicating a normalization of probabilities.

In this theoretical framework, binding prediction is inherently conditional, focusing on the presence of positive and negative annotations for proteins i and j. Therefore, $p_{ij}^{(1,0)}$ and $p_{ij}^{(0,1)}$ are normalized by the probability of observing a generic annotation, i.e., $p_{ij}^{(1,0)} + p_{ij}^{(0,1)}$. For unseen edges, binding prediction is determined by:

$$p_{ij}^{\text{conditional}} = \frac{p_{ij}^{(1,0)}}{p_{ij}^{(1,0)} + p_{ij}^{(0,1)}} \tag{7}$$

In the case of an unseen protein j^*, the binding probability toward a known compound i is calculated as follows:

$$p_{ij^*}^{\text{conditional}} = \frac{\langle p_{ij}^{(1,0)} \rangle_j}{\langle p_{ij}^{(1,0)} \rangle_j + \langle p_{ij}^{(0,1)} \rangle_j} \tag{8}$$

Where $\langle \cdot \rangle_j$ represents the average over all known proteins.

For unseen proteins i^* and j^*, the binding probability is determined based on the overall

number of positive ($L^{(1,0)}$) and negative ($L^{(0,1)}$) annotations:

$$p_{i^*j^*}^{\text{conditional}} = \frac{\langle p_{ij}^{(1,0)} \rangle_{ij}}{\langle p_{ij}^{(1,0)} \rangle_{ij} + \langle p_{ij}^{(0,1)} \rangle_{ij}} = \frac{L^{(1,0)}}{L^{(1,0)} + L^{(0,1)}} \tag{9}$$

Here, $\langle \cdot \rangle_{ij}$ indicates the average over all known protein pairs.

3 Annotation Imbalance

The absence of negative samples in PPI databases results in a node-wise class imbalance, where only positive examples are available for the majority of proteins during training. This limitation leads to overpredictions of interactions for these proteins, as the ML models lack exposure to both positive and negative examples [5]. Annotation imbalance is a pervasive issue in interaction databases, as discussed in the context of drug-target binding databases in AI-Bind [5].

To address the annotation imbalance in protein-protein interactions, we utilize topological PPNI. As depicted in Figure S4A, topological PPNI introduces a significant number of negative samples (3,063,605) compared to subcellular compartmental negatives (738,551) and Negatome (2,211). This approach effectively tackles the challenge of annotation imbalance by providing both positive and negative samples for the majority of the proteins in the training, validation, and test sets of ComPPlete (see Figure S4B).

4 Coefficient of Variation for GPCR Predictions.

To further investigate ComPPlete output, we looked into the coefficient of variation (for the 5-folds) as illustrated in Figure S5. Coefficient of variation is able to capture the combined effect of the standard deviation and mean of the predictions from 5-folds of ComPPlete.

5 Alternate protein Embeddings

We evaluate ComPPlete against another alternative protein representation learning methodology. Variational autoencoders (VAE) [22] are utilized to derive protein embeddings from amino acid sequences by minimizing both reconstruction loss and distributional loss. We implement VAE-based embeddings following the methodology proposed by Hawkins-Hooker et al. [23]. Initially, we generate one-hot encodings for the amino acid residues and employ these encodings to represent the sequences. Subsequently, a VAE is trained to produce 100-dimensional representations of the sequences. Alternatively, we employ a message-passing neural network (MPNN) [24] on 3D protein structures sourced from the Protein Data Bank (PDB) and those predicted by AlphaFold [25]. However, our examination reveals that Protvec outperforms both

VAE and MPNN in inductive testing scenarios (refer to Table S1). Furthermore, we create a simple embedding of the 3D protein structures from the PDB files. We one hot encode the atoms and flatten out the information in the PDB files (atoms, coordinates, and electron densities). We use zero-padding at the end for making the embeddings of same length for different proteins. We observe that these simpler 3D embeddings, although performs poorly compared to Protvec, show improved performance compared to the MPNN approach.

Table S1: **Comparing Different Protein Embeddings Methods in Inductive Tests.** We compare Protvec-based protein representations with VAE and MPNN-based approaches. In both cases, Protvec performs better. Interestingly, a simpler one-hot encoding-based approach on the PDB files provides better performance compared to the MPNN-based protein embeddings.

Model	AUROC	AUPRC	Hits@Top1000	Hits@Top10,000
Protvec	0.79 ± 0.007	0.87 ± 0.009	0.92 ± 0.015	0.87 ± 0.010
VAE	0.64 ± 0.05	0.63 ± 0.05	0.74 ± 0.13	0.64 ± 0.05
MPNN	0.59 ± 0.06	0.62 ± 0.07	0.25 ± 0.008	0.024 ± 0.001
PDB	0.74 ± 0.03	0.72 ± 0.03	0.68 ± 0.01	0.55 ± 0.02

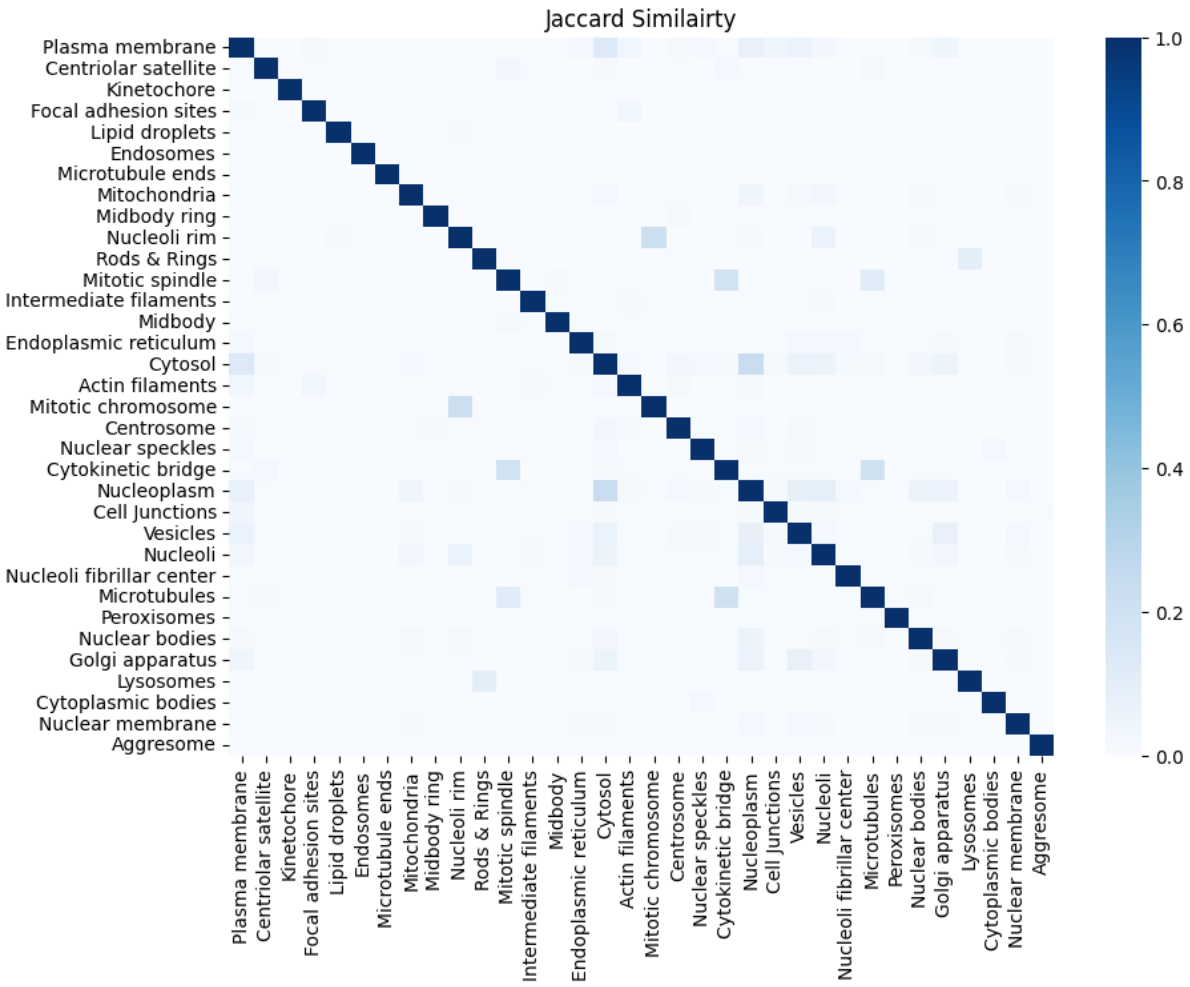


Figure S1: **Subcellular Compartmental Negatives (SCN).** Protein overlap among 34 subcellular compartments in terms of Jaccard similarity [26]. PPNI pairs are samples from a pair of compartments that have no overlapping protein in the heatmap.

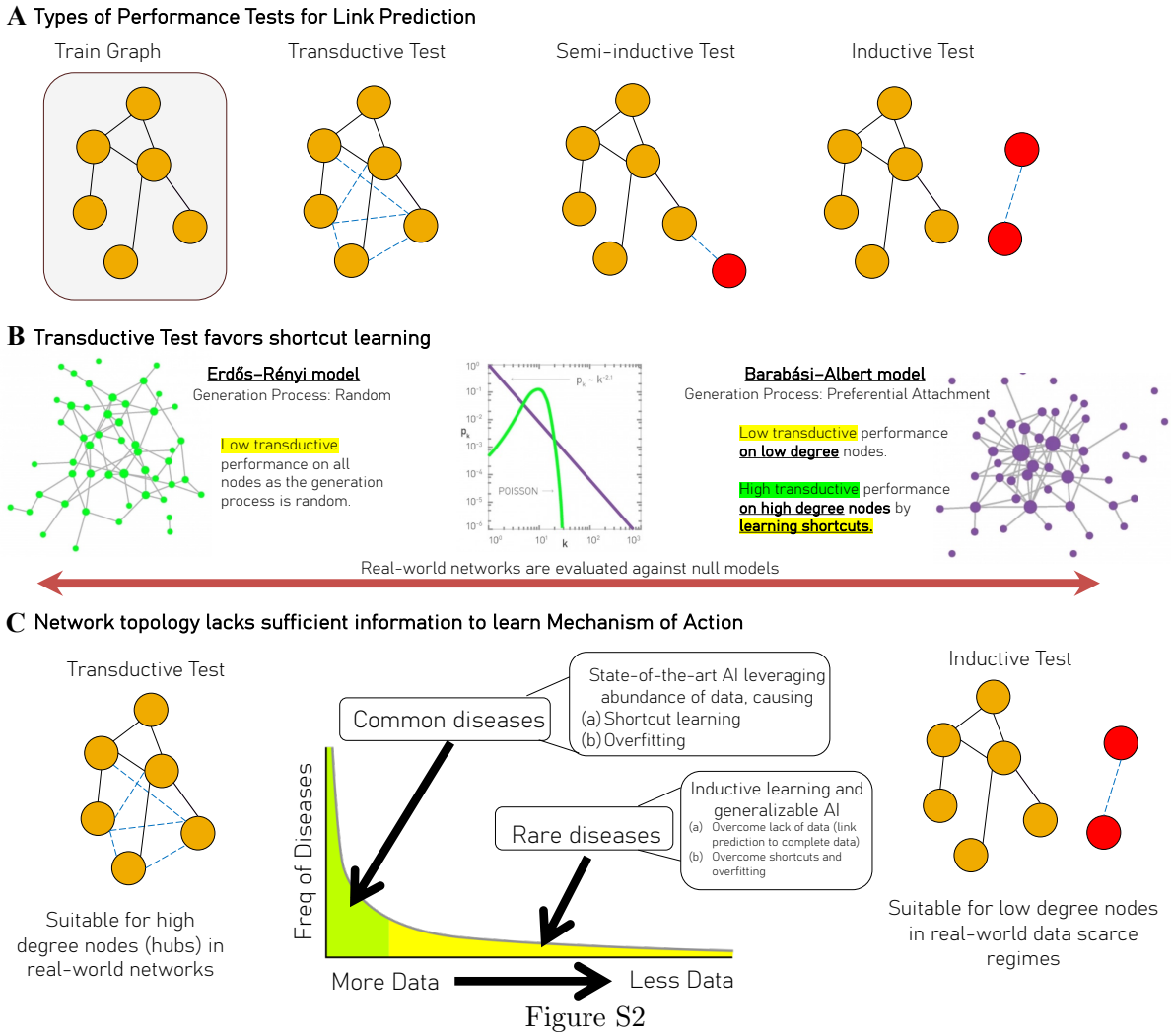
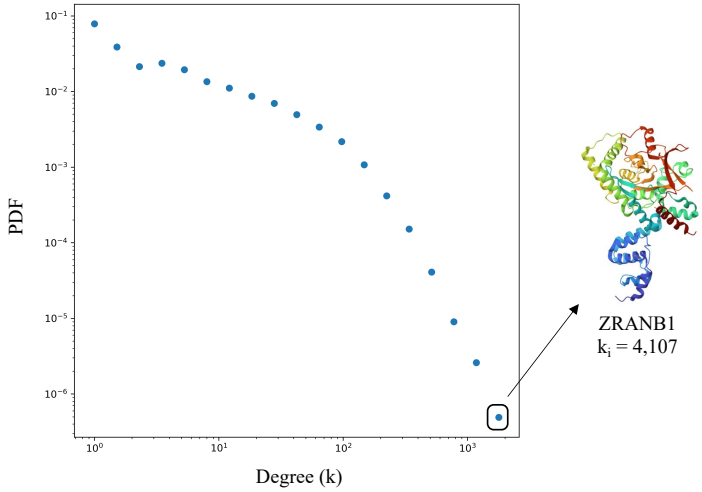

A Types of Performance Tests for Link Prediction
Train Graph
Transductive Test
Semi-inductive Test
Inductive Test
B Transductive Test favors shortcut learning
Erdős–Rényi model
Generation Process: Random
Low transductive performance on all nodes as the generation process is random.
POISSON
k
Barabási–Albert model
Generation Process: Preferential Attachment
Low transductive performance on low degree nodes.
High transductive performance on high degree nodes by learning shortcuts.
Real-world networks are evaluated against null models
C Network topology lacks sufficient information to learn Mechanism of Action
Transductive Test
Suitable for high degree nodes (hubs) in real-world networks
Common diseases
State-of-the-art AI leveraging abundance of data, causing
(a) Shortcut learning
(b) Overfitting
Rare diseases
Inductive learning and generalizable AI
(a) Overcome lack of data (link prediction to complete data)
(b) Overcome shortcuts and overfitting
Freq of Diseases
More Data
Less Data
Inductive Test
Suitable for low degree nodes in real-world data scarce regimes

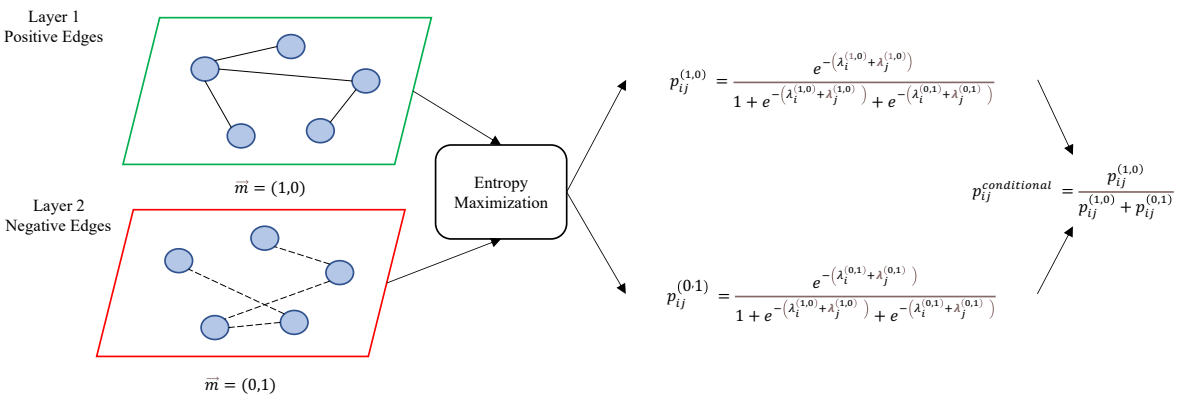
Figure S2

Figure S2: **Transductive Test Favors Shortcut Learning.** The selection of test scenario can significantly affect the patterns that ML models learn from data. Intuitively, we expect deep models to learn the Mechanisms of Actions (MoA) behind the generation of data, yet it is common that ML models learn shortcuts instead. **(a)** In transductive scenario both the topological structure (patterns of connections) and node attributes are included in the training phase. The semi-inductive scenario partially hides the topological structure from one side of link prediction, by predicting interaction between a node never-seen-before (annotated with red color) and another node that has been seen during the training phase. Lastly, the inductive test prevents the leakage of topological information from train to test/validation sets, conducting link prediction strictly on data-never-seen before, which enforces an AI to learn MoA from node attributes rather than patterns of connections between nodes (shortcuts). **(b)** Real-world networks are often evaluated against two extremes, one is Erdős–Rényi (ER) model where the graph is generated randomly, and the other is Barabási–Albert (BA) model where the graph is generated by enforcing preferential attachment based on the degree of nodes on each iteration of BA model. In ER, the transductive test performs poorly as the graph is generated purely by randomness. In BA model the transductive test only performs well on high degree needs as the classifier learns high degree nodes have a higher chance of having connections. Yet the classifier does not perform well on low degree nodes since the network topology alone does not have enough information to learn the MoA (preferential attachment), which in this case is the parameters used to generate a BA model. **(c)** Transductive scenario may offer insightful predictions when abundant of data is available by facilitating learning the structural topology. Yet, in the data scarce regime, like rare disease area, the performance of transductive scenario is low due to lack of topological structure. In the data scarce regime, learning from node attributes is essential for link prediction, hence we need to enforce learning the MoA behind emergence of structural topology that is domain specific (e.g., social and biological interactions are driven through different MoAs).

A Degree Distribution of BKB PPI Network



B Duplex Configuration Model



C

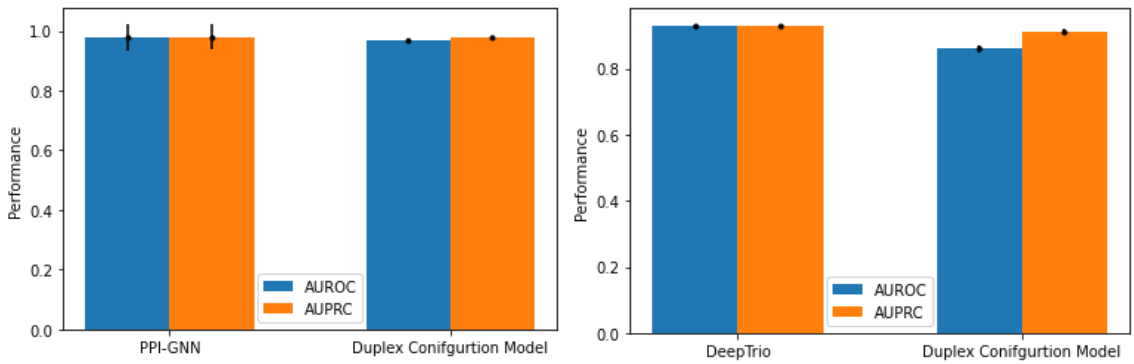
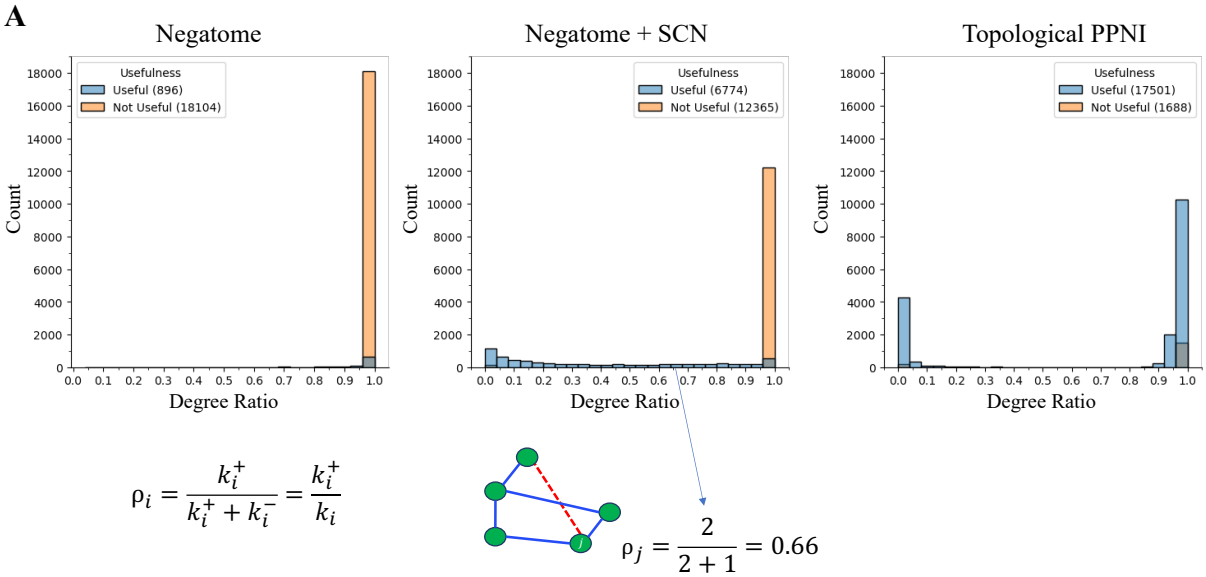


Figure S3

Figure S3: **State-of-the-art and Duplex Configuration Model. (A)** We observe a heavy-tailed degree distribution in the QIAGEN BKB [11] PPI network, which confirms the existence of hubs (high-degree nodes) in the PPI network. For example, ZRANB1 interacts with 4,107 other genes, constituting a hub. **(B)** The unipartite duplex configuration model takes as input PPI training data and random negative samples. By an entropy maximization, we first predict the link probabilities in each layer, which we then combine to obtain the output probability $p_{ij}^{conditional}$. **(C)** The duplex configuration model achieves performance comparable to two state-of-the-art PPI prediction models PPI-GNN and DeepTrio in the transductive setting.



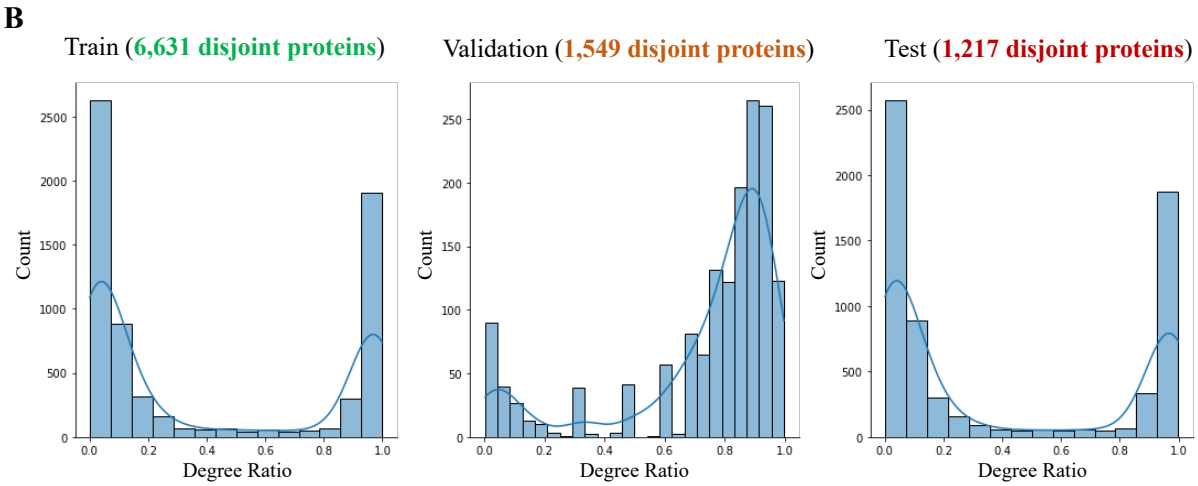


Figure S4

Figure S4: **Removing annotation imbalance and topological shortcuts. (A)** Combining the PPI databases and Negatome with only 2k PPNI samples creates a large annotation imbalance in the training data. The majority of proteins only have positive samples and lack negative examples. This imbalance is quantified by degree ratio ($\rho = \frac{k^+}{k}$) [5], which is measured by the ratio of positive annotations (k^+) to total annotations ($k = k^+ + k^-$) associated with a protein. Subcellular Compartmental Negatives (SCN) introduces 778k PPNI examples, rectifying annotation imbalance partially. Finally, topological PPNI fixes the issue of annotation imbalance by introducing a substantial number (3 million) of non-interactions. We then filter the interactions and non-interactions involving the proteins having both positive and negative samples for training and testing of ComPPlete. Removing proteins with only positive or negative samples helps ComPPlete circumvent learning shortcuts related to predicting that some proteins always interact or do not interact. **(B)** Using only topological PPNI we remove annotation imbalance from all train, validation, and test data of ComPPlete. Here, we visualize the degree ratio distributions in train, validation, and test datasets for the first fold.

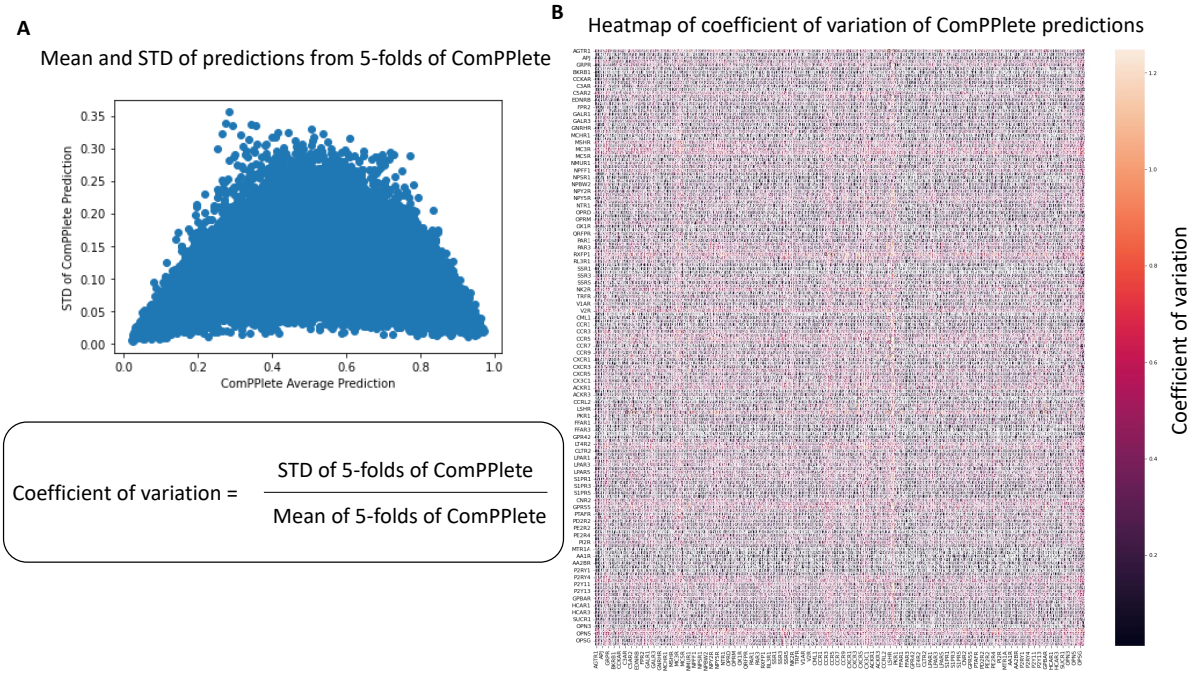


Figure S5

Figure S5: **Coefficient of variation. (A)** We plot the standard deviation (STD) vs. mean of the predictions from the 5-folds of ComPPlete for GPCR proteins. We observe that the STD is lower for the predictions near 0 and 1, indicating that ComPPlete models from 5-folds agree upon the top and the bottom predictions. Whereas, the predictions in between have higher STD. Thus, we define coefficient of variation of ComPPlete predictions as the ratio of the STD and mean from the 5-folds. Lower value of this coefficient suggests lower STD with high average, this high confidence interaction predictions. **(B)** We plot the heatmap of the coefficient of variation from ComPPlete predictions for 170 GPCR proteins. The darker dots correspond to the high confidence interaction predictions.

References

[1] Qi, Y., Bar-Joseph, Z. & Klein-Seetharaman, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins:*

Structure, Function, and Bioinformatics **63**, 490–500 (2006). URL https://doi.org/10.1002/prot.20865.

[2] Feng, Y., Wang, Q. & Wang, T. Drug target protein-protein interaction networks: A systematic perspective. *BioMed Research International* **2017**, 1–13 (2017). URL https://doi.org/10.1155/2017/1289259.

[3] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999). URL https://doi.org/10.1126/science.286.5439.509.

[4] Tan, J., Yang, J., Wu, S., Chen, G. & Zhao, J. A critical look at the current train/test split in machine learning (2021). 2106.04525.

[5] Chatterjee, A. *et al.* Improving the generalizability of protein-ligand binding predictions with AI-bind. *Nature Communications* **14** (2023). URL https://doi.org/10.1038/s41467-023-37572-z.

[6] Chatterjee, A., Walters, R., Menichetti, G. & Eliassi-Rad, T. Disentangling node attributes from graph topology for improved generalizability in link prediction (2023). URL https://doi.org/10.48550/arXiv.2307.08877.

[7] ERDdS, P. & R&wi, A. On random graphs i. *Publ. math. debrecen* **6**, 18 (1959).

[8] Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* (2019). URL https://doi.org/10.1093/nar/gkz1021.

[9] Genes, G. Rare diseases, common challenges. *Nature Genetics* **54**, 215–215 (2022). URL https://doi.org/10.1038/s41588-022-01037-8.

[10] Huang, K. *et al.* DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* (2020). URL https://doi.org/10.1093/bioinformatics/btaa1005.

[11] Qiagen biomedical knowledge base (2023.2).

[12] Hu, F., Jiang, J., Wang, D., Zhu, M. & Yin, P. Multi-PLI: interpretable multi-task deep learning model for unifying protein–ligand interaction datasets. *Journal of Cheminformatics* **13** (2021). URL https://doi.org/10.1186/s13321-021-00510-6.

[13] Jha, K., Saha, S. & Singh, H. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports* **12**, 8360 (2022).

[14] Menichetti, G. & Remondini, D. Entropy of a network ensemble: Definitions and applications to genomic data. *Theoretical Biology Forum* **107**, 77–87 (2014).

[15] Teru, K. K., Denis, E. & Hamilton, W. L. Inductive relation prediction by subgraph reasoning (2019). URL https://arxiv.org/abs/1911.06962.

[16] Park, Y. & Marcotte, E. M. Flaws in evaluation schemes for pair-input computational predictions. *Nature Methods* **9**, 1134–1136 (2012). URL https://doi.org/10.1038/nmeth.2259.

[17] Chatterjee, A., Walters, R., Menichetti, G. & Eliassi-Rad, T. Inductive link prediction in static and temporal graphs for isolated nodes. In *Temporal Graph Learning Workshop @ NeurIPS 2023* (2023). URL https://openreview.net/forum?id=DRrSYKNhD1.

[18] Meng, R. *et al.* Attention diversification for domain generalization (2022). 2210.04206.

[19] Bardes, A., Ponce, J. & LeCun, Y. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations* (2022). URL https://openreview.net/forum?id=xm6YD62D1Ub.

[20] Alberts, B. & etc. (eds.) *Molecular biology of the cell* (CRC Press, Boca Raton, FL, 2002), 4 edn.

[21] Menichetti, G. & Remondini, D. Entropy of a network ensemble: definitions and applications to genomic data. *Theor Biol Forum* **107**, 77–87 (2014).

[22] Kingma, D. P. & Welling, M. Auto-encoding variational bayes (2022). 1312.6114.

[23] Hawkins-Hooker, A. *et al.* Generating functional protein variants with variational autoencoders. *PLOS Computational Biology* **17**, e1008736 (2021). URL http://dx.doi.org/10.1371/journal.pcbi.1008736.

[24] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1263–1272 (JMLR.org, 2017).

[25] Wang, L., Liu, Y., Lin, Y., Liu, H. & Ji, S. ComENet: Towards complete and efficient message passing for 3d molecular graphs. In Oh, A. H., Agarwal, A., Belgrave, D. & Cho, K. (eds.) *Advances in Neural Information Processing Systems* (2022). URL https://openreview.net/forum?id=mCzMqeWSFJ.

[26] Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist* **11**, 37–50 (1912). URL https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.