# Multiview Contrastive Autoencoder-Transformer Approach for Protein-Protein Interface Representation: Unveiling Biological and Functional Insights

by

**Damla Övek**

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Computer Science and Engineering

KOÇ ÜNİVERSİTESİ

July 24, 2024

**Multiview Contrastive Autoencoder-Transformer Approach for Protein-Protein Interface Representation: Unveiling Biological and Functional Insights**

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

**Damla Övek**

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Attila Gürsoy (Advisor)

_____

Prof. Z. Özlem Keskin Özkaya [Co-Advisor]

_____

Prof. Erkut Erdem

_____

Assoc. Prof. Nurcan Tunçbağ

_____

Assoc. Prof. Aykut Erdem

_____

Assoc. Prof. Öznur Taştan

Date: _____

*Dedicated to my family.*

# ABSTRACT

**Multiview Contrastive Autoencoder-Transformer Approach for Protein-Protein Interface Representation: Unveiling Biological and Functional Insights**

**Damla Övek**

**Doctor of Philosophy in Computer Science and Engineering**

**July 24, 2024**

Protein-protein interactions (PPIs) play pivotal roles in various biological processes, orchestrating cellular functions essential for life. The interfaces where these interactions occur serve as focal points for understanding the mechanisms underlying disease pathways. Accurate representation of these interfaces is crucial for deciphering their biological significance and designing therapeutic interventions. This thesis introduces a novel approach for representing protein-protein interfaces using a graph-based multiview contrastive autoencoder combined with a transformer, which learns representations from a large dataset. Comprehensive evaluations demonstrate the method's effectiveness in capturing the structural and functional characteristics of protein-protein interfaces. The learned representations are applied to tasks such as biological relevance prediction, biological vs. crystal classification, and Gene Ontology term prediction, showcasing their versatility and utility in understanding PPIs. By integrating explainable AI techniques, key features contributing to model predictions are identified, enhancing the interpretability of the results. A detailed case study illustrates the practical application of these methods, highlighting their potential to provide actionable insights for biological research and drug discovery. Overall, this thesis advances the understanding of protein-protein interactions by providing interpretable representations that capture the complex structural and functional characteristics of interfaces, thereby facilitating biomedical studies and therapeutic developments.

# ÖZETÇE


**Protein-Protein Arayüzü Gösterimi için Çoklu Görünümlü Karşılaştırmalı Otomatik Kodlayıcı-Transformatör Yaklaşımı: Biyolojik ve İşlevsel İçgörülerin Ortaya Çıkarılması**

**Damla Övek**

**Bilgisayar Bilimleri ve Mühendisliği, Doktora**

**24 Temmuz 2024**

Protein-protein etkileşimleri, çeşitli biyolojik süreçlerde hayati roller oynayarak yaşam için gerekli olan hücresel işlevleri düzenler. Bu etkileşimlerin gerçekleştiği arayüzler, hastalık yollarının altında yatan mekanizmaları anlamak için odak noktalarıdır. Protein-protein arayüzlerinin doğru temsili, biyolojik önemlerini çözmek ve terapötik müdahaleler tasarlamak için zorunludur. Bu tez, büyük bir veri setinden temsiller öğrenmek için bir graf tabanlı çoklu görüş karşılaştırmalı otomatik kodlayıcı ve bir transformer kullanan yeni bir yaklaşım sunmaktadır. Kapsamlı değerlendirmeler, yöntemin protein-protein arayüzlerinin yapısal ve işlevsel özelliklerini doğru bir şekilde yakalamadaki etkinliğini göstermektedir. Öğrenilen temsiller, biyolojik alaka tahmini, biyolojik ve kristal sınıflandırma ve Gen Ontolojisi terim tahmini gibi görevlerde uygulanarak, PPI'ları anlamadaki çok yönlülüğünü ve faydasını ortaya koymaktadır. Açıklanabilir yapay zeka tekniklerini entegre ederek, model tahminlerine katkıda bulunan anahtar özellikler belirlenmekte ve sonuçların yorumlanabilirliği artırılmaktadır. Ayrıntılı bir vaka çalışması, bu yöntemlerin pratik uygulamasını göstermekte ve biyolojik araştırma ve ilaç keşfi için sağladığı eyleme geçirilebilir içgörüleri vurgulamaktadır. Genel olarak, bu tez, protein-protein etkileşimlerini anlamayı ilerleterek, arayüzlerin karmaşık yapısal ve işlevsel özelliklerini yakalayan yorumlanabilir temsiller sağlayarak biyomedikal çalışmaları ve terapötik gelişmeleri kolaylaştırmaktadır.

# ACKNOWLEDGMENTS

First and foremost, I extend my deepest gratitude to my advisors, Prof. Ozlem Keskin and Prof. Attila Gursoy, whose expertise, understanding, and patience added considerably to my graduate experience. Their guidance helped me throughout the research and writing of this thesis.

Besides my advisors, I would like to thank the rest of my thesis committee: Assoc. Prof. Nurcan Tuncbag, Prof. Erkut Erdem, Assoc. Prof. Aykut Erdem, and Assoc. Prof. Öznur Taştan, for their time, insightful comments, and encouragement.

My sincere thanks must also go to the members of the COSBI lab for their collaboration and companionship throughout my time there. The lab has been a second home, and I am indebted to all its members for their help and all the fun we have had in the last four years.

I am deeply grateful to my husband, Batuhan Teoman Baydar, for his unwavering support and love throughout this journey. His understanding and personal sacrifices have made it possible for me to complete this work. Thank you for always being there, for the reassurance, and for believing in me even when I doubted myself.

Lastly, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you, Mom and Dad, for always believing in me.

This journey has been a transformative phase of my life, and I appreciate everyone who has been part of it, knowingly or unknowingly.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABBREVIATIONS

| | |
|---|---|
| PPI | Protein-Protein Interaction |
| GCL | Graph Convolution Layer |
| GNN | Graph Neural Network |
| ASA | Accessible Surface Area |
| CAPRI | Critical Assessment of Prediction of Interactions |
| XAI | Explainable Artificial Intelligence |
| SHAP | Shapley Additive Explanations |
| LIME | Local Interpretable Model-agnostic Explanations |

# Chapter 1

# INTRODUCTION

Proteins are large, complex molecules essential for numerous functions within living cells. They transport nutrients, accelerate chemical reactions, and form the structural framework of organisms. Antibodies, which protect the body from harmful entities such as bacteria and viruses, are also proteins [Keskin et al., 2008a]. Typically, proteins do not act in isolation; they interact with other proteins to perform their cellular functions [Keskin et al., 2008b]. These interactions, known as protein-protein interactions (PPIs), involve highly specific physical contacts between two or more protein molecules, driven by biochemical activities such as electrostatic forces, hydrogen bonding, and hydrophobic effects[Keskin et al., 2008a].

PPIs form the foundation of intricate signaling networks within cells. Signaling proteins transmit extracellular signals, like hormones or growth factors, from the cell membrane to the nucleus, thereby regulating gene expression, cell proliferation, differentiation, and apoptosis [Mayer, 1999]. Enzymatic proteins engage in PPIs to catalyze biochemical reactions within metabolic pathways, coordinating processes essential for cellular and organismal function [Williamson and Sutcliffe, 2010]. Transcription factors, which control gene expression, often interact with other proteins to modulate their activity[Stallcup et al., 2003], binding to specific DNA sequences to regulate target gene transcription in response to internal and external stimuli [Georgiev and Maksimenko, 2014].

Additionally, PPIs are crucial for maintaining cellular architecture and motility, mediating interactions between structural proteins. These interactions are vital for cell division, tissue morphogenesis, and maintaining cellular integrity [Poluri et al., 2021]. In the immune system, PPIs play key roles in antigen recognition,

signal transduction, and immune cell activation, triggering responses that eliminate pathogens and regulate immune tolerance [Merwe and Davis, 2003, Tamir and Cambier, 1998].

Overall, protein-protein interactions are integral to the functioning of biological systems, facilitating communication, coordination, and regulation at the molecular level [Williamson and Sutcliffe, 2010]. Given the integral role of PPIs in biological systems, understanding their complexity and dynamics is essential for elucidating cellular processes, disease mechanisms, and therapeutic targets in fields such as molecular biology, medicine, and drug discovery [Chakraborty et al., 2014]. Any disruption in protein functions can have detrimental effects on the organism, with many diseases resulting from protein mutations that alter binding sites. With over 645,000 disease-relevant PPI interfaces reported among human interactions [Mabonga and Kappo, 2019], targeting these interactions with drug-like molecules presents a significant therapeutic strategy [Vassilev et al., 2004].

PPIs are found experimentally via co-immunoprecipitation, affinity purification, and yeast-two hybrid techniques. Mutagenesis techniques, NMR spectroscopy, and X-ray crystallography are used to determine a protein's binding site. Although these experimental methods are very accurate, they are expensive, time-consuming, and and not applicable on a large scale [Phizicky and Fields, 1995].

Protein-protein interactions and binding sites are predicted using a variety of computational techniques. Protein binding sites and structural interactions between proteins, for instance, can be studied using molecular dynamics simulations [Rakers et al., 2015, Kuzmanic et al., 2020]. Nevertheless, modeling conformational changes in these investigations is challenging and computationally expensive for simulation-based approaches. For PPI prediction, other computational methods make advantage of known interaction data. These techniques rely on gene expression profile, coevolution, and homology [Keskin et al., 2016b].

Computational methods utilize information obtained from proteins' sequence and structure. There are billions of known protein sequences; however, structures of only a tiny fraction of these proteins have been identified through enormous ex-

perimental efforts. Recently, AlphaFold [Jumper et al., 2021] has been developed by DeepMind, which accurately predicts protein 3D structure from the sequence using deep learning. Thanks to AlphaFold, protein structure data has been accumulating. In this study, protein structure data is used to study protein-protein interactions.

Protein-protein interfaces represent the spatial regions where two or more protein molecules interact. These interfaces serve as the molecular battlegrounds for intricate molecular recognition events. Within these interfaces, amino acid residues from distinct protein molecules come into close proximity, engaging in a complex network of non-covalent interactions, including hydrogen bonds, van der Waals forces, hydrophobic interactions, and electrostatic attractions. The topography of protein-protein interfaces is defined by the arrangement of atoms and functional groups on the molecular surfaces, dictating the interactions' specificity, affinity, and dynamics [Abali et al., 2022]. Understanding the structural and chemical properties of protein-protein interfaces is essential for unraveling the molecular basis of cellular processes and designing targeted interventions for various biomedical applications.

Patterns of chemical and geometric properties on molecular surfaces offer information about the protein's interactions with other biomolecules. Proteins with no sequence homology and globally different structures that engage in similar biomolecular interactions may share architectural motifs [Gainza et al., 2020a, Tuncbag et al., 2011a]. Therefore, encoding architectural motifs and functional groups is significant in dictating the specificity and selectivity of protein-protein interactions, regardless of the sequence or overall fold of the participating proteins. Accurate representation of protein-protein interfaces becomes paramount in this context. Computational methods for representing proteins mainly rely on two approaches: sequence-based representations and structure-based representations.

Sequence-based representation methods encode information solely based on the amino acid sequence of proteins. Sequence-based representations include one-hot encoding [Agrawal et al., 2022, Cui et al., 2021], position-specific scoring matrices (PSSMs) [Wang et al., 2019], and amino acid composition profiles[Liu et al., 2013]. One-hot encoding represents each amino acid as a binary vector, where each element

corresponds to a specific amino acid. PSSMs capture evolutionary information by quantifying the occurrence of amino acids at each protein sequence position derived from multiple sequence alignments. Amino acid composition profiles count the occurrences of each amino acid in a protein sequence.

Sequence-based representations are computationally efficient and easy to generate. They can be applied to various proteins, including those with unknown or unresolved structures. Moreover, they leverage evolutionary information, providing insights into conserved regions and functional motifs. However, they lack three-dimensional structural information, limiting their ability to capture spatial relationships and interactions within proteins. They may overlook subtle variations in protein structures that affect function and binding specificity and struggle to accurately predict interactions involving structurally divergent proteins or complex binding mechanisms.

Structure-based representation methods utilize the three-dimensional structure of proteins to encode spatial arrangements of atoms and functional groups. Structure-based representations capture detailed structural features, including binding sites, secondary structures, and solvent accessibility. They provide insights into protein-protein interactions, ligand binding, and conformational changes and can accurately predict binding affinity, specificity, and functional consequences of mutations.

The emergence of deep learning approaches has revolutionized the analysis of proteins, offering powerful tools for extracting complex features and patterns from protein sequences and structures [Jumper et al., 2021, Gainza et al., 2020a, Réau et al., 2023]. Deep learning approaches enable data-driven representation learning, allowing models to automatically extract relevant features directly from raw protein data, overcoming the limited expressivity of handcrafted features or predefined descriptors [Tubiana et al., 2022].

Pre-trained on large datasets and fine-tuned on specific tasks, deep learning models have become increasingly popular in protein representation learning. Pre-trained models, such as language models trained on protein sequences or protein structure prediction models, provide valuable starting points for downstream tasks and can

significantly reduce the need for large labeled datasets [Jha et al., 2023, Van Kempen et al., 2024]. State-of-the-art performance in various protein-related tasks can be achieved by leveraging pre-trained representations with minimal computational resources.

Despite the importance of protein-protein interfaces, existing methods for representing them have several limitations that hinder their effectiveness. They may need more scalability and generalizability, mainly when dealing with large and diverse data sets. Most existing approaches rely on manual feature engineering or predefined structural descriptors [Gainza et al., 2020a], limiting their ability to adapt to novel protein structures or accommodate heterogeneous data sources. Moreover, the interpretability of representations generated by existing methods may need to be improved, hindering biological interpretation and hypothesis generation.

There is a need for more effective approaches to learning representations directly from protein-protein interface data by leveraging advanced deep-learning techniques to address these challenges and overcome the limitations. This dissertation proposes a method to capture the complex structural and functional characteristics of protein-protein interfaces in a data-driven and interpretable manner.

To address these challenges, this dissertation proposes a novel method for capturing the complex structural and functional characteristics of protein-protein interfaces in an interpretable and data-driven manner. The approach combines a graph-based multiview contrastive autoencoder with a transformer architecture, offering several advantages, including the ability to capture the network of interactions between amino acid residues, incorporate multiple data modalities, and leverage contrastive learning techniques to disentangle informative features from noisy data.

The objectives of this dissertation are to design and implement a graph-based multiview contrastive autoencoder combined with a transformer architecture tailored for protein-protein interface representation learning, train the proposed model on large unlabeled data set of protein-protein interfaces to learn informative and interpretable representations, evaluate the learned representations on benchmark data sets and assess their performance in tasks such as biological relevance validation, bi-

ological vs. crystal classification, gene ontology term prediction, and functional annotation, and comparing the performance of the proposed method with existing state-of-the-art approaches to demonstrate its superiority in terms of accuracy, scalability, and interoperability.

The contributions of this work to the field of protein-protein interactions include addressing the limitations of existing methods by developing a novel deep-learning approach capable of learning representations from protein-protein interface data, advancing the understanding of protein-protein interactions by providing interpretable representations that capture the structural and functional characteristics of interfaces comprehensively, demonstrating the effectiveness of the proposed method in various tasks, including biological relevance validation and functional annotation, thereby facilitating biomedical studies and drug discovery efforts.

The remaining part of the thesis is formed from the following chapters: Chapter 2 presents an extensive review of related work in the field of protein-protein interactions. It covers the significance of PPIs for pharmaceutical studies, protein function prediction, and genome-wide systems biology research. The chapter discusses the latest structure-based computational protein representation learning models, state-of-the-art PPI validation, docking scoring, and crystal interface discrimination studies. It concludes with a summary of successful and promising deep learning approaches used in PPI studies, such as geometric deep learning, graph neural networks, and contrastive learning.

In Chapter 3, the novel representation learning method is explained. This chapter details the data set used, input preparation, and graph construction for the proposed model. It describes the model architecture, including the graph-based contrastive autoencoder and the transformer. The results and discussion section presents an ablation study and t-SNE and UMAP analysis to evaluate the learned representations' quality and robustness. The chapter highlights the proposed method's effectiveness in capturing complex structural and functional characteristics of protein-protein interfaces.

Chapter 4 focuses on validating the biological relevance of PPI interfaces. The

proposed method is evaluated using various datasets, including DeepInterface, GNN-DOVE, and DeepRank-GNN. It presents the model architecture and discusses the performance comparison results with existing methods. The chapter concludes by emphasizing the proposed method's superior performance in validating protein-protein interfaces.

Chapter 5 introduces the accurate classification of biological and crystallographic protein interfaces. It describes the data and methods used, including the MANY/DC benchmark datasets. The results and discussion section presents the performance evaluation of the proposed method, highlighting its high accuracy, precision, and F1 score compared to established classifiers.

Chapter 6 explains the comprehensive analysis conducted to evaluate the proposed approach's effectiveness in predicting gene ontology (GO) terms. It describes the dataset used, the study's primary objectives, and the methods for predicting various facets of protein functionality. The results and discussion section presents the performance comparison with existing models.

Chapter 7 employs explainable AI techniques, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), to analyze the importance of input features that contribute to the model's prediction. A detailed case study demonstrates that the important residues that contribute to the classification model's learning are hot spots. The chapter concludes by emphasizing the importance of interpretability in AI models for protein-protein interactions.

Finally, the thesis concludes by summarizing the contributions and findings of the proposed method, explaining how it enhances the understanding of protein-protein interaction interfaces.

# Chapter 2

# LITERATURE REVIEW

This chapter presents an extensive literature review of related work in the field. First, we give general information about protein-protein interaction interfaces, emphasizing their significance for pharmaceutical studies, prediction of protein function, and genome-wide systems biology research. Then, we explain the most recent structure-based computational protein representation learning models. We present state-of-the-art protein-protein interface validation, docking scoring, and crystal interface discrimination studies. This chapter concludes with a comprehensive summary of previous successful and promising deep learning approaches used in protein-protein interaction studies, such as geometric deep learning, graph neural networks, and contrastive learning, as well as the integration of explainable artificial intelligence methods to interpret the predictions of these approaches.

## *2.1 Protein-Protein Interaction Interfaces*

Protein-protein interactions (PPIs) are essential to cellular functions, with their interfaces playing a vital role in signal transmission, molecular complex formation, and pathway regulation [Keskin et al., 2016a]. These interfaces are central to understanding cellular mechanisms and have significant implications for understanding diseases and developing drugs [Abalı, 2021]. Here, the state-of-the-art PPI prediction techniques are explained, and related challenges and current issues on this topic are discussed.

Experimental techniques identify the physiochemical interactions of proteins to predict the functional relationships between them. These techniques are yeast two-hybrid based methods [Bartel and Fields, 1997], mass spectrometry [Gavin et al., 2002], Tandem Affinity Purification [Rigaut et al., 1999], protein chips [Zhu et al.,

2001], and hybrid approaches [Tong et al., 2002]. However, they are expensive, time-consuming, labor-intensive, and can only cover part of the PPI network. Therefore, computational techniques are needed. Computational PPI prediction is an extensively studied topic, and many approaches have been proposed. We can categorize these approaches as sequence- and structure-based techniques.

Existing studies address protein-protein interactions from various points by predicting protein binding regions [Jimenez et al., 2017, Khan et al., 2022, Tubiana et al., 2022, Zhao et al., 2022], identifying individual residues that contribute significantly to the interaction [Deng et al., 2019, Moreira et al., 2017, Tuncbag et al., 2009a, Zhu et al., 2020], and eventually designing specific interfaces. Protein-protein interactions can also be predicted using computational docking approaches [Andrusier et al., 2008, de Vries et al., 2010, Tsaban et al., 2022].

Several computational techniques for PPI identification have been presented in recent decades with fully sequenced genomes and proteomes. Because many functionally relevant proteins are conserved across species, early computational approaches rely heavily on statistical properties and conserved patterns of proteins. Proteins with homologous sequence patterns or structures are likelier to have similar interaction characteristics. Some PPIs can be predicted using homologous proteins from different species [Ding and Kihara, 2018]. As a result, numerous techniques use 'interologs' (conserved PPIs) to predict PPIs across a wide range of species [Huang et al., 2004, Lee et al., 2008, Matthews et al., 2001], and some of the projected PPIs have been confirmed experimentally.

Machine learning approaches were later used in PPI prediction, dating back to 2001 [Bock and Gough, 2001]. Supervised learning (including Bayesian inference, decision trees, support vector machines (SVMs), and artificial neural networks (ANNs)), unsupervised learning (such as K-means and spectral clustering), and reinforcement learning are the three primary categories. Among these methods, SVM seeks an ideal hyperplane that separates the distinct labeled samples with the greatest possible margin. SVM-based techniques can use several protein properties, such as conserved sequence patterns, 3D architectures, domain compositions, and cor-

responding gene expression [Bock and Gough, 2001, Guo et al., 2008, Shen et al., 2007]. Decision tree-based approaches iteratively partition the sample space depending on the many features of proteins. These features are obtained from the primary sequences [Chen and Jeong, 2009, Xia et al., 2010, You et al., 2015, Zahiri et al., 2013], 3D structures [Li et al., 2012] and domain composition [Chen and Liu, 2005, Rodgers-Melnick et al., 2013].

One of the existing studies, PRISM, predicts PPIs following the hypothesis that if the complementary partners of a known interface are similar to the surface regions of any two unknown proteins, these two proteins can interact with each other through these regions. PRISM employs geometric complementarity detected by structural alignments and evolutionary conservation of hot spots while looking for similar spatial patterns on an unknown protein surface [Tuncbag et al., 2011a].

Recent studies have used deep learning models to predict PPIs. Their architectures are composed of convolutional neural networks (CNNs) [Hashemifar et al., 2018, Hu et al., 2021, Yang et al., 2021a], recurrent neural networks (RNNs) [Chen et al., 2019], multilayer perceptrons (MLPs) [Mahapatra and Sahu, 2021, Yao et al., 2019], and graph convolutional networks (GCNs) [Liu et al., 2020b, Yang et al., 2020]. Due to their powerful non-linear transformation ability, deep models perform better than conventional machine learning-based approaches in PPI prediction [Hu et al., 2022a]. Deep learning models are divided into three categories: sequence-based [Hashemifar et al., 2018], structure-based [Gainza et al., 2020a], and combined methods [Song et al., 2022, Yang et al., 2020]. Some sequence-based methods also use the network topology information to enhance the prediction performance [Yang et al., 2020].

In recent decades, convolutional neural networks(CNNs) and graph neural networks(GNNs) are the most popular techniques used in computational protein-protein interaction studies. However, existing studies [Réau et al., 2023, Wang et al., 2021] have proven that GNNs perform better on the protein tertiary structure data than the CNNs since the CNNs are not rotation invariant.

MaSIF (molecular surface interaction fingerprinting) [Gainza et al., 2020a] uses

geometric deep learning to identify interaction fingerprints of proteins from the surface. MaSIF divides a surface into overlapping radial patches. Each patch point is assigned an array of geometric and chemical input features. MaSIF then learns to incorporate the input features of the surface patch into a numerical vector descriptor. Each descriptor is then processed further with application-specific neural network layers. One specific application is MaSIF-site. It predicts a score for each surface vertex, which shows the likelihood of the vertex being in the binding site. Another application is MaSIF-search, which learns patterns in interacting pairs of surface patches and predicts potential binding partners.

Both MaSIF and PRISM depend on the calculations of various time-consuming external tools. MaSIF is built on top of handcrafted features and is limited by the expressivity of the handcrafted features. PRISM makes all-against-all structural alignment comparisons, which is computationally expensive.

According to PRISM's working principle, identifying interfaces of unknown proteins and comparing them with the known interfaces are two crucial steps to be taken to predict protein-protein interactions. Several studies have been conducted to predict proteins' binding sites and compare protein structures using deep learning methods.

One of the pioneering works in this field by Jiménez et al. introduced a convolutional neural network (CNN) based method for predicting protein-ligand binding affinities [Jiménez et al., 2017]. Their model, DeepSite, uses 3D CNNs to scan the surface of proteins and identify potential binding pockets. The success of DeepSite highlighted the effectiveness of CNNs in capturing spatial relationships within protein structures, a critical factor in identifying binding sites. Following this, several studies have sought to enhance the accuracy and efficiency of binding site prediction using various deep learning architectures.

ScanNet is a recent method to predict protein binding sites from structure using geometric deep learning [Tubiana et al., 2022]. There are four main stages in the architecture. These are atomic neighborhood embedding, atom-to-amino acid pooling, amino acid neighborhood embedding, and neighborhood attention. It outputs

residue-wise scores showing the probability of that residue being located in the protein binding site. However, ScanNet reaches an accuracy of 87.7% for protein-protein binding site prediction and could not beat AlphaFold-multimer for partner-specific protein-protein binding site prediction. Considering AlphaFold-multimer is also not highly accurate in the case of heteromers, there is still a need for improvement.

Another method predicts protein binding residues by harnessing the power of language models adapted for bioinformatics. The technique, MsPBRsP, employs a multi-scale analysis strategy, examining protein sequences at various levels, from individual amino acids to larger structural motifs. By leveraging language models trained on extensive protein sequence databases, MsPBRsP can identify key residues involved in binding interactions, considering both the local context around each residue and the broader structural domains [Li et al., 2023].

PeSTo [Krapp et al., 2023] represents another significant contribution to binding site prediction, applying parameter-free geometric deep learning to predict PPI interfaces with high accuracy. This model distinguishes itself by its ability to directly process the geometric information inherent in protein structures without relying on extensive parameter tuning, which is a common challenge in traditional deep learning models. By focusing on the spatial relationships and orientations of amino acids in 3D space, PeSTo captures the complex patterns that govern protein interactions.

DoGSite3 [Graef et al., 2023] focuses on improvements like enhanced detection of binding sites in the presence of ligands, optimized parameters for robust predictions, and faster computations. The tool uses a grid-based method and a Difference-of-Gaussian filter for cavity detection, offering significant advancements in binding site prediction crucial for drug discovery.

In summary, the fast-paced evolution of deep learning technologies offers a revolutionary framework for predicting Protein-Protein Interactions (PPIs). The variety and complexity of deep learning methods applied to PPI prediction highlight the field's vigorous and inventive development. Although deep learning applications have seen significant advances, there remain obstacles that necessitate further exploration and resolution.

## *2.2 Protein Representation Learning*

Proteins are typically represented by their sequence and/or structures. One way to encode protein sequences in sequence-based studies is one hot vector [Liu et al., 2020a]. Amino acid-level evolutionary information may be encoded using position-specific scoring matrices (PSSM) [Li et al., 2021b].

The representation of protein structures may vary as primary (1D), secondary (2D), or tertiary structures (3D) based on the specific context and intended application. Three-dimensional structures capture the arrangement of atoms in a protein and provide valuable information about its folding, functional sites, and interaction surfaces. Protein structures can be represented as point clouds [Chen et al., 2023] or graphs [Derry and Altman, 2023]. Each protein atom is placed in a space according to its coordinates in point cloud representation [Chen et al., 2023]. Alternatively, only the carbon alpha atom of each residue may be used [Derry and Altman, 2023]. Using a binary two-dimensional matrix, a protein contact map represents the distance between all possible amino acid residue pairs of a 3-D protein structure. For two residues $i$ and $j$, the matrix's $ij^(th)$ element is 1 if the two residues are closer than a predetermined threshold and 0 otherwise. Another common representation is to use graphs. In graph representation, each node represents a residue[Réau et al., 2023] or an atom[Wu et al., 2023a], and if two nodes are closer to each other than a predefined threshold, there is an edge between them[Réau et al., 2023, Wu et al., 2023a, Derry and Altman, 2023].

Deep learning approaches are then employed to analyze and extract meaningful features from these 3D structures. One commonly used approach is convolutional neural networks (CNNs) [Hashemifar et al., 2018, Hu et al., 2021, Yang et al., 2021a], which can learn hierarchical representations by applying filters to local regions of the protein structure. In CNN-based studies, 3-D structures can be converted into 2D contact maps that are translation and rotation invariant. Recurrent neural networks (RNNs)[Chen et al., 2019] are also utilized to capture sequential dependencies among residues in the primary sequences of proteins. Additionally, graph neural

networks (GNNs) have gained popularity for modeling proteins as graphs, where nodes represent amino acids and edges denote spatial, sequential, and/or functional relationships [Réau et al., 2023, Wang et al., 2021].

Renowned for their proficiency in handling spatial data, CNNs are recognized for their utility in PPI research [Hu et al., 2022b, Chen et al., 2022, Gao et al., 2023]. The convolutional layers within these networks are capable of identifying and processing spatial hierarchies in data, making them well-suited for analyzing protein structures and sequences.

RNNs and long short-term memory networks (LSTMs) are lauded for their sequential data processing capabilities. They excel in capturing temporal dependencies and patterns within protein sequences, pivotal for understanding the sequential and interdependent nature of proteins and their interactions[Alakus and Turkoglu, 2021, Aybey and Gümüş, 2023, Li et al., 2021a].

GNNs are expert in leveraging the graph-based nature of protein interaction networks and are commended for their ability to model the intricate web of PPIs effectively, capturing the complex relationships and interactions within these biological networks [Quadrini et al., 2022, Réau et al., 2023].

Autoencoders focus on the transformation of protein sequences or structures into insightful representations [Hasibi and Michoel, 2021, Czibula et al., 2021, Zhang et al., 2022, Yue et al., 2022, Soleymani et al., 2023]. They distill complex biological data into compressed, informative features, facilitating the analysis and prediction of PPIs by reducing data dimensionality and complexity.

Attention mechanisms and transformers, initially rooted in natural language processing, have found significant applications in PPI prediction. Their ability to focus on relevant parts of protein sequences and to model complex dependencies without the constraints of sequence length positions them as powerful tools for PPI analysis [Zhu et al., 2022, Asim et al., 2022, Nambiar et al., 2023, Wu et al., 2023b].

Multi-task and multi-modal learning approaches involve models that can simultaneously address multiple learning tasks or integrate various types of biological data [Pan et al., 2022]. Such models offer a holistic view of PPIs by leveraging

diverse data sources and predictive tasks, enriching the analysis with multifaceted biological insights [Dong et al., 2021, Capel et al., 2022].

Transfer learning is recognized for its efficiency in utilizing pre-trained models to expedite and enhance PPI prediction tasks. Transfer learning allows for the leveraging of learned features from one context to be applied to another, reducing the need for extensive data and computational resources in PPI analysis [Yang et al., 2021b, Derry and Altman, 2023].

DeepFold is a deep learning method that uses a convolutional neural network (CNN) encoder to extract descriptors from intra-residue distance matrices [Liu et al., 2018]. However, the number of DeepFold's parameters is high, which makes it computationally ineffective. Furthermore, CNNs may struggle to accurately capture the spatial interactions between residues in protein structures [Xia et al., 2021].

Another recent deep learning study is graph-based protein structure representation learning (GraSR) for fast and accurate comparison of protein structures [Xia et al., 2022]. This method builds a graph based on the intra-residue distance derived from the protein 3D structure. Then, graph representations of the protein structures are learned by a deep graph neural network (GNN) architecture with a short-cut connection under a contrastive learning framework. Given a query protein, and other protein(s), it builds the graph from the structure, converts it to a fixed-length vector representation with contrastive learning, and measures the similarity between the query protein's vector and the other protein(s)' vectors. Length-scaling cosine distance is used to measure the similarity. Although GraSR can provide fast and accurate structure comparison, it has difficulty getting the superposition information of the protein structures, causing a poor understanding of the structural similarity at the atomic level. Hence, there is still a gap for further improvement.

Zhang et. al. proposed an approach of pretraining protein representations based on their 3D structures [Zhang et al., 2022]. This study introduces a structure-based encoder, GearNet, designed to encode the spatial and geometric features of proteins effectively. The encoder employs a Geometry-Aware Relational Graph Neural Network (GearNet) to facilitate this process, enhanced further by an edge message

passing layer for capturing interactions between protein residues more accurately. To harness the potential of GearNet, the study presents geometric pretraining methods that leverage multiview contrastive learning and various self-prediction tasks, aiming to enrich the protein representations by capturing the underlying biological and structural nuances. The pretraining framework maximizes the mutual information between biologically correlated views of protein substructures, thereby embedding similar substructures closer in the latent space while distinguishing unrelated ones.

In conclusion, protein representation learning has witnessed remarkable improvements, with advanced deep learning algorithms offering unprecedented insights into protein structures and functions. Techniques such as GNNs and transformers have revolutionized the understanding of the intricate relationships and spatial configurations of proteins, facilitating breakthroughs in drug discovery, disease understanding, and synthetic biology. Despite these advancements, the intrinsic complexity of proteins, including their dynamic and multi-scale nature, poses significant representation challenges that current models struggle to fully capture. Furthermore, interpretability remains a critical issue; understanding the 'why' behind model predictions is crucial for trust and applicability in clinical settings. Addressing these challenges requires not only methodological innovations but also interdisciplinary collaboration to integrate domain-specific knowledge, thus paving the way for the next generation of deep learning applications in protein science.

## *2.3 Protein-Protein Interface Validation and Docking Model Scoring*

Experimentally derived protein structures have been instrumental in elucidating the physicochemical underpinnings of protein complex functionalities. Nonetheless, such experimental techniques are often hindered by high costs and lengthy timeframes. To complement these efforts, the last two decades have seen significant advancements in computational modeling. Computationally, docking methodologies are commonly employed to predict protein-protein structures [Lensink et al., 2016]. The execution of docking simulations between two protein entities typically yields a voluminous set of potential configurations, known as decoys, numbering in the hundreds or thou-

sands. Consequently, the capacity to distinguish between native-like configurations and erroneous predictions generated by docking procedures is paramount. A pivotal strategy in achieving this distinction is precisely predicting the interface regions where the interacting proteins converge.

Protein docking is broadly categorized into template-based [Tuncbag et al., 2011b, Anishchenko et al., 2015], which employs existing structures as the foundational framework for modeling, and ab initio techniques that aggregate individual structural data and evaluate generated models to identify the most credible configurations. Within the scope of ab initio strategies, a multitude of techniques have been applied for the representation of molecular structures [Venkatraman et al., 2009, Pierce et al., 2011], encompassing conformational search algorithms such as the fast Fourier transform [Katchalski-Katzir et al., 1992, Padhorny et al., 2016], geometric hashing [Fischer et al., 1995, Venkatraman et al., 2009], and particle swarm optimization [Moal and Bates, 2010], in addition to accounting for protein flexibility [Gray et al., 2003, Oliwa and Shen, 2015]. The advent of novel techniques aims to enhance and extend the functionalities beyond simplistic binary docking, incorporating advancements such as multichain docking [Schneidman-Duhovny et al., 2005, Esquivel-Rodríguez et al., 2012, Ritchie and Grudinin, 2016], peptide-protein interactions [Kurcinski et al., 2015, Alam et al., 2017, Kurcinski et al., 2020], engagement with disordered proteins [Peterson et al., 2017], predictions of docking sequences [Peterson et al., 2018b, Peterson et al., 2018a], and applications to cryo-EM structural maps [Esquivel-Rodríguez and Kihara, 2012, van Zundert et al., 2015]. Recent incorporations of deep learning advancements have further augmented docking efficacy [Akbal-Delibas et al., 2016, Degiacomi, 2019, Gainza et al., 2020b].

Notwithstanding the significant enhancements in ab initio protein docking, the discernment of near-native configurations from an extensive array of generated models, often termed decoys, persists as a formidable challenge. This challenge is exacerbated by the notable disproportion between near-native models and erroneous decoys within the pool of generated decoys. The precision in scoring these decoys is pivotal for the overarching efficacy of protein docking, catalyzing the active

development of scoring algorithms [Moal et al., 2013] for docking models. Acknowledging the criticality of accurate scoring, the Critical Assessment of PRediction of Interactions (CAPRI) [Lensink et al., 2018], a communal endeavor in protein docking predictions, has designated a specific category for the evaluation of scoring methodologies, wherein participants are tasked with selecting ten viable decoys from a multitude provided by the organizers. Over the preceding two decades, a diverse array of scoring methodologies has been developed, encompassing physics-based potentials [Akbal-Delibas et al., 2016, Degiacomi, 2019, Gainza et al., 2020b], assessments based on interface morphology [Akbal-Delibas et al., 2016, Kingsley et al., 2016, Degiacomi, 2019, Gainza et al., 2020b], statistical potentials derived from knowledge [Lu et al., 2003, Huang and Zou, 2008], machine learning techniques [Fink et al., 2011], analyses of evolutionary profiles at interface residues [Nadaradjane et al., 2018], and deep learning algorithms leveraging interface structural data [Wang et al., 2020, Balci et al., 2019, Renaud et al., 2021, Wang et al., 2021, Réau et al., 2023].

DeepInterface [Balci et al., 2019] is a CNN-based method that takes two proteins as input and predicts whether they form a biologically valid complex interface. It utilizes voxelized representations of proteins. DOVE [Wang et al., 2020] is another CNN model that examines the protein-protein interfaces using 3D voxels, considering atomic interaction types and their energetic contributions as input characteristics applied to the neural network. However, CNNs are not rotation invariant, requiring data augmentation to mitigate orientation sensitivity, which affects computational performance during training.

Alternatively, representing protein-protein interfaces as rotation-invariant graphs enables the use of GNNs to overcome these limitations. GNNs handle rotation-invariant graph representations without extensive data augmentation. GNN-DOVE [Wang et al., 2021] employs a graph neural network model, representing atoms' chemical properties as node features and inter-atom distances as edge features. Similarly, DeepRank-GNN [Réau et al., 2023] is a GNN-based approach that scores docking models by converting interface residues into two graphs: one representing internal

edges between residues from the same chain and another for external edges between residues from different chains. These graphs are then sequentially processed by a dedicated graph interaction neural network.

Furthermore, the potential of using transformers to investigate protein-protein interfaces has been demonstrated by a recent approach, PIsToN [Stebliankin et al., 2023]. This approach differentiates native-like protein complexes from incorrect conformations by transforming protein interfaces into two-dimensional interface maps. It combines Vision Transformer [Dosovitskiy et al., 2020] architecture with hybrid and multi-attention networks.

This study presents a graph-based approach for validating the biological relevance of protein-protein interfaces using learned interface representations. The generation of adequate representations for protein-protein interfaces is accomplished through a graph autoencoder. To enhance the representation quality, a transformer component is incorporated to predict edge features. The model incorporates a contrastive layer to capture the similarity between correlated substructures within proteins and an edge message passing layer to capture the interdependence among interactions involving a residue and its neighboring residues.

## *2.4 Biological vs. Crystal Classification*

Significant attempts have been made for several decades to determine protein-protein complexes' three-dimensional (3D) structures. The majority of the high-resolution structures have been generated by X-ray crystallography, accounting for 84.5% of all structures deposited in the Protein Data Bank (PDB) as of February 2024 [Sussman et al., 1998]. Despite being the most widely utilized experimental technique, X-ray crystallography poses two significant problems [Elez et al., 2020]. The first is technical, referring to the difficulties of producing high-quality protein crystals of sufficient size and form to allow structural determination [Wernimont and Edwards, 2009]. The second pertains to the analysis and interpretation of the generated structures. At the molecular level, a crystal is formed by an infinite repetition of unit cells, which can result in the formation of two types of protein-protein interfaces

(PPIs): biologically relevant ones, which correspond to the interface occurring in solution and eliciting the biological function, and crystallographic ones, which are simply artifacts of crystal packing. Specifically, assemblies with a Kd value in the low micromolar range or lower are typically considered biological, unlike weak interactions with larger Kd values [Ali and Imperiali, 2005, Capitani et al., 2016, Krissinel and Henrick, 2007].

Given the propensity for structural ambiguities within X-ray crystallography, various computational techniques have been developed to distinguish biologically significant interfaces from those that are artifacts of crystallization [Bliven et al., 2018, Jiménez-García et al., 2019].

Biological and crystallographic interfaces exhibit distinct characteristics. Parameters such as interface size, estimated solvation energy, the number of salt bridges, knowledge-based atomic pair-wise potentials, and evolutionary conservation are typically used to differentiate them. No single property can distinguish between the two types of interfaces; a combination of these features is necessary for accurate classification [Elez et al., 2020].

Existing computational methods can be grouped into three main categories: energy-based, empirical knowledge-based, and machine learning-based methods. Energy-based approaches use scoring functions to estimate the stability of an interface by approximating its energy. The other two categories consider different properties of the interface, using either heuristic rules (empirical knowledge-based) or specific algorithms (machine learning-based) [Elez et al., 2020].

### *2.4.1 Energy-Based Classification Approaches*

ClusPro-DC [Yueh et al., 2017] extends the traditional docking algorithm ClusPro [Kozakov et al., 2017] to classify interfaces. It separates the subunits of a dimer, docks them numerous times without restrictions, and retains the top 1000 poses with the lowest energy. The number of near-native structures is determined by counting docked poses whose C$\alpha$ interface RMSD is less than 7 Å from the crystal structure. A threshold of 33 was found optimal, leading to a classification accuracy of 74.5%.

PISA [Krissinel and Henrick, 2007] identifies stable assemblies by calculating the free energy change upon dissociation. All stable assemblies are ranked using preferences such as larger assemblies, single-assembly sets, and higher free energy of dissociation. The method achieves an overall accuracy of 90%.

### *2.4.2 Empirical, Knowledge-Based Approaches*

Liu et al. [Liu et al., 2014] investigated the relevance of B-factors for classifying biological interfaces, finding that B-factor-related features outperformed interface area in cross-dataset classification performance. The average B-factor score showed higher accuracy and specificity.

EPPIC [Duarte et al., 2012] uses core residue count and evolutionary criteria such as core-to-rim and core-to-surface entropy ratios. It applies a majority voting scheme unless a specific threshold is met. The method achieves an accuracy of 81

PreBI [Tsuchiya et al., 2006] and COMP [Tsuchiya et al., 2008] evaluate the complementarity and area of an interface. They use heuristic rules based on hydrophobicity, electrostatic potential, and shape. COMP identifies 84.8% of biological interfaces in the training dataset.

### *2.4.3 Machine Learning-Based Approaches*

Recent advancements have seen the integration of machine learning algorithms within tools like EPPIC 3 [Bliven et al., 2018], PRODIGY-CRYSTAL [Jiménez-García et al., 2019], and PIACO [Fukasawa and Tomii, 2019], while methodologies such as DeepRank-GNN [Réau et al., 2023] leverage the power of deep learning to refine interface classification.

PIACO [Fukasawa and Tomii, 2019] uses covariation signals along with other features to train a random forest classifier, achieving an accuracy of 85%.

PRODIGY-CRYSTAL [Jiménez-García et al., 2019] combines residue contact properties and interaction energies to distinguish interfaces, achieving an accuracy of 92%.

RPAIAnalyst [Hu et al., 2018] utilizes co-evolutionary features and residue pair frequency among other properties, achieving an accuracy of 84.6%.

NOXclass [Zhu et al., 2006] uses interface area, amino acid composition, and the ratio of interface area to protein surface area to train an SVM, achieving an accuracy of 97.9%.

IChemPIC [Da Silva et al., 2015] uses random forest classifiers based on 45 features related to interface size, chemical complementarity, and buriedness, achieving an accuracy of 75%.

Luo et al. [Luo et al., 2014] employ a random forest classifier using 46 features, achieving an accuracy of 86.9%.

IPAC [Mitra and Pal, 2011] uses a naive Bayes classifier to evaluate interfaces, achieving an accuracy of 90%.

DiMoVo (DIscrimination between Multimers and MOnomers by VOronoi tessellation) [Bernauer et al., 2008] uses Voronoi tessellation and SVM to classify interfaces, achieving an accuracy of 95%.

Valdar and Thornton [Valdar and Thornton, 2001] use size and conservation to classify interfaces, showing comparable performance to linear heuristics.

### *2.4.4 Datasets*

Reliable datasets are essential for training and comparing new predictors. Notable datasets include those compiled by Ponstingl et al. [Ponstingl et al., 2003], Bahadur et al. [Bahadur et al., 2004], and the manually annotated DC dataset. These datasets cover various interface sizes and types, providing a solid foundation for developing and validating classification methods.

The emergence of machine learning and deep learning technologies has introduced a new era in this field, propelled by the increasing availability of extensive datasets. These advanced techniques have enabled the development of more sophisticated models that can capture the subtle nuances and complex patterns inherent in protein interfaces. Consequently, the application of these models has significantly enhanced the precision and reliability of interface classification, leading to improved predictive

performance.

A recent study by Schweke et al. [Schweke et al., 2023], addresses the challenge of reliably scoring and ranking candidate models of protein complexes and assigning their oligomeric state from the structure of the crystal lattice. A community-wide effort was launched to tackle these challenges by exploiting the latest resources on protein complexes and interfaces.

The research involved creating a benchmark dataset of 1677 homodimer protein crystal structures, which included a balanced mix of physiological and non-physiological complexes. The non-physiological complexes were chosen to bury a similar or larger interface area than their physiological counterparts, making it more difficult for scoring functions to differentiate between them. The study collected and evaluated 252 functions for scoring protein-protein interfaces developed by 13 groups to determine their ability to discriminate between physiological and non-physiological complexes.

Two approaches were used to evaluate the performance of these scoring methods: a simple consensus score generated using the best-performing score from each group, and a cross-validated Random Forest (RF) classifier. Both approaches showed excellent performance, with an area under the Receiver Operating Characteristic (ROC) curve of 0.93 and 0.94, respectively, outperforming individual scores developed by different groups. Additionally, AlphaFold2 engines recalled the physiological dimers with significantly higher accuracy than the non-physiological set, supporting the reliability of the benchmark dataset annotations.

The study highlights that optimizing the combined power of interface scoring functions and evaluating it on challenging benchmark datasets appears to be a promising strategy. The consensus and RF classifier scores outperformed individual scoring functions, demonstrating the complementarity of different features captured by various scoring methods. These findings suggest that the combination of different scores and their integration into machine learning models can significantly improve the prediction and classification of protein interfaces, enhancing our understanding of protein-protein interactions and their roles in biological processes.

Over the years, significant research has concentrated on the detailed analysis of protein interfaces, largely driven by the wealth of data generated through X-ray crystallography. These tools have achieved notable success rates in classification tasks, underscoring the effectiveness of their underlying methodologies.

Despite these advancements, there remains a need for further accuracy improvements. The complexity of biological systems and the diversity of protein interactions present ongoing challenges that require continuous refinement of computational methods. Enhancing the training datasets with more diverse and high-quality examples, integrating multi-modal data sources, and developing more robust algorithms are crucial steps toward achieving higher accuracy.

Moreover, the integration of these computational tools with experimental methods can provide complementary insights, leading to a more comprehensive understanding of protein interactions. As research progresses, the synergy between computational advancements and experimental data is expected to drive further breakthroughs in classifying protein interfaces, ultimately contributing to a deeper understanding of biological processes and developing novel therapeutic strategies.

## *2.5 Explainable Artificial Intelligence (XAI) Methods in Protein-Protein Interactions*

Explainable Artificial Intelligence (XAI) has gained significant traction in the field of bioinformatics, particularly in understanding complex protein-protein interactions (PPIs). Traditional machine learning and deep learning models, although powerful, often operate as "black boxes," providing predictions without transparent reasoning. XAI addresses this limitation by providing insights into the decision-making process of AI models, making them more interpretable and trustworthy.

One of the primary XAI methods involves determining feature importance, where the contribution of each input feature to the model's prediction is assessed. For instance, methods like SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017] and LIME (Local Interpretable Model-agnostic Explanations) [Ribeiro et al., 2016] have been widely adopted. SHAP values, based on cooperative game theory,

provide a unified measure of feature importance, indicating how each feature impacts the prediction positively or negatively. LIME, on the other hand, explains individual predictions by approximating the black box model locally with an interpretable model.

In deep learning models, particularly those using neural networks, attention mechanisms have been employed to highlight which parts of the input data are most relevant for making a prediction [Vaswani et al., 2017]. In the context of PPIs, attention layers can help identify critical amino acids or interaction sites that contribute most significantly to the interaction prediction. This approach not only enhances the model's performance but also provides an intuitive understanding of which features are most important.

ExplainableFold(xFold) [Tan and Zhang, 2023] proposed an innovative framework to demystify the black-box nature of AlphaFold's predictions in protein structure prediction. This framework employs counterfactual learning to generate explanations by systematically deleting or substituting amino acids in the protein sequence. Through rigorous experiments on the CASP-14 protein dataset, the study demonstrates that ExplainableFold provides high-quality explanations for AlphaFold's predictions, thereby enabling a near-experimental understanding of the effects of amino acids on 3D protein structure. This approach bridges the gap between AI predictions and biological interpretability, offering significant potential for advancing our understanding of protein structures.

MaTPIP [Ghosh and Mitra, 2024] presents a deep learning architecture integrated with XAI for predicting protein-protein interactions using sequence-based features. MaTPIP combines pre-trained Protein Language Model (PLM)-based features with manually curated protein sequence attributes, leveraging both two-dimensional granular amino acid-level features and one-dimensional whole protein-level features. The architecture includes a configuration of CNN with Transformer components to handle these mixed features. The XAI assessment revealed the contribution of different feature families to each prediction, highlighting the effectiveness of integrating PLM-based features with manually curated data for accurate PPI

predictions.

KGsim2vec is a method to generate explainable vector representations for protein-protein interaction prediction by leveraging knowledge graphs and semantic similarity features [Sousa et al., 2024]. The approach employs aspect-oriented semantic similarity to capture different semantic perspectives within a knowledge graph, providing a more interpretable alternative to traditional black-box models like knowledge graph embeddings and graph neural networks.

Despite the advancements, several challenges remain in applying XAI to PPIs. One of the primary challenges is the trade-off between model complexity and interpretability. More complex models often provide better predictions but are harder to interpret. Striking a balance between these two aspects is crucial. Additionally, there is a need for standardized benchmarks and evaluation metrics to assess the effectiveness of XAI methods in the context of PPIs.

# Chapter 3

# PROTEIN-PROTEIN INTERFACE REPRESENTATION LEARNING

This chapter explains the methodology and framework developed for protein-protein interface representation learning. It begins by detailing the dataset used for training and evaluating the proposed model. This includes the preparation of input data, specifically how protein structures are converted into graph representations.

The core of this chapter focuses on the model architecture. A graph convolutional network-based encoder-decoder framework, combined with a transformer, is introduced to generate detailed and informative embeddings of protein-protein interfaces. Following the architectural description, the training process and the methods employed to optimize the model are presented. This includes a discussion on the hyperparameters, loss functions, and training strategies used to ensure the model learns robust and interpretable representations.

The chapter also includes an evaluation of the model's performance through various experiments. Results from ablation studies, t-SNE, and UMAP analyses are presented to demonstrate the model's effectiveness in accurately representing protein-protein interfaces. Finally, it concludes with a discussion of the implications of the findings and potential directions for future research.

## *3.1 Data set*

A comprehensive and extensive set of unlabeled protein-protein interfaces is employed for both the training and testing phases of the graph-based contrastive autoencoder, which is further integrated with a transformer model to generate detailed protein representations. This data set encompasses a substantial total of 534,203

samples, ensuring a robust and diverse source of information for model development. For optimal training and evaluation, 80% of the data is allocated for training purposes, while the remaining 20% is reserved for testing. This strategic partitioning allows for a thorough assessment of the model's performance.

The data set originates from the research conducted by Abali et *et al* (2021) [Abalı, 2021], and it is readily accessible at https://interactome.ku.edu.tr:8443/PPInt/. Furthermore, detailed information regarding the Protein Data Bank (PDB) and chain IDs associated with the samples within the data set is available through the GitHub repository at `https://github.com/ku-cosbi/ProInterVal`.

A comprehensive analysis of the structural properties of these protein-protein interfaces has been conducted, focusing on three key characteristics: size, hydrophobicity, and polarity. By examining the distribution of these features across the dataset, the aim is to uncover patterns and trends that can provide insights into the structural and functional aspects of protein-protein interactions.

The size of an interface corresponds to the number of residues, while it is polarity is defined by the number of polar interface residues. Additionally, the hydrophobicity of an interface is calculated by summing the hydrophobicity values of the interface residues given in *hydrophobicity_scale* dictionary (See Appendix A.1).

The size distribution of the protein interfaces, as shown in Figure 3.1, indicates that the majority of interfaces have relatively small sizes. The histogram reveals a sharp peak at lower sizes, with most interfaces having sizes below 200 residues. There is a steep decline in frequency as the size increases, with very few interfaces exceeding 500 residues. This distribution suggests that smaller interfaces are much more common, which could be due to the prevalence of smaller proteins or modular domains within larger proteins.

Figure 3.2 displays the distribution of polarity across the protein interfaces. The polarity distribution of the protein interfaces shows a similar trend to the size distribution. The majority of interfaces have low polarity values, with a significant peak at the lower end of the scale. The frequency decreases rapidly as polarity increases, indicating that highly polar interfaces are relatively rare. This pattern suggests

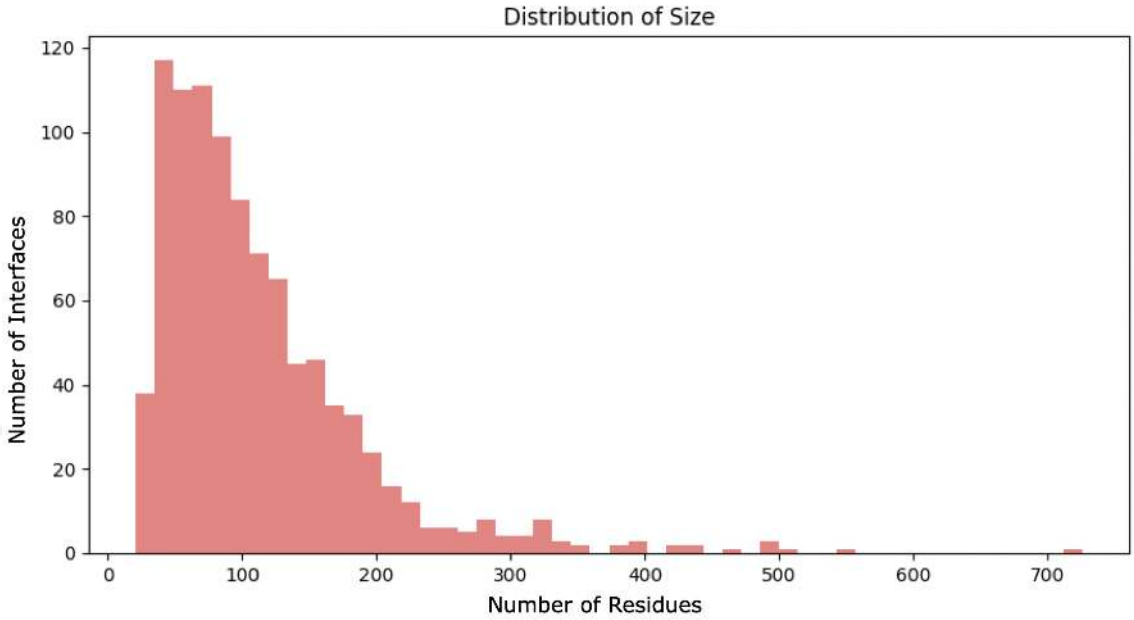


Figure 3.1: Size distribution of protein-protein interface dataset.

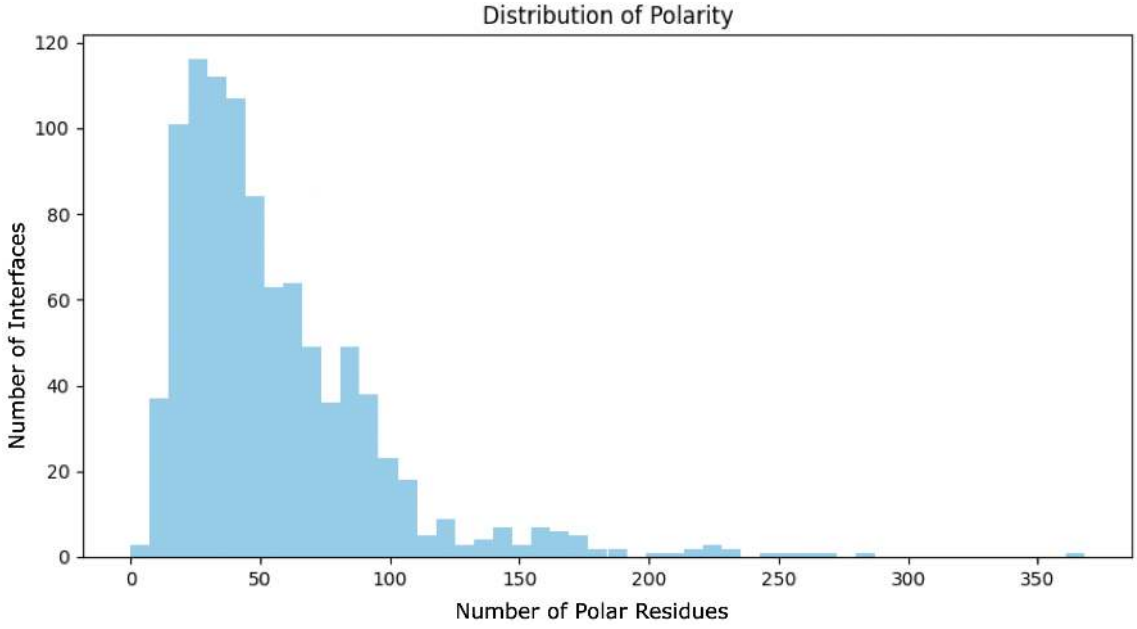


Figure 3.2: Polarity distribution of protein-protein interface dataset.

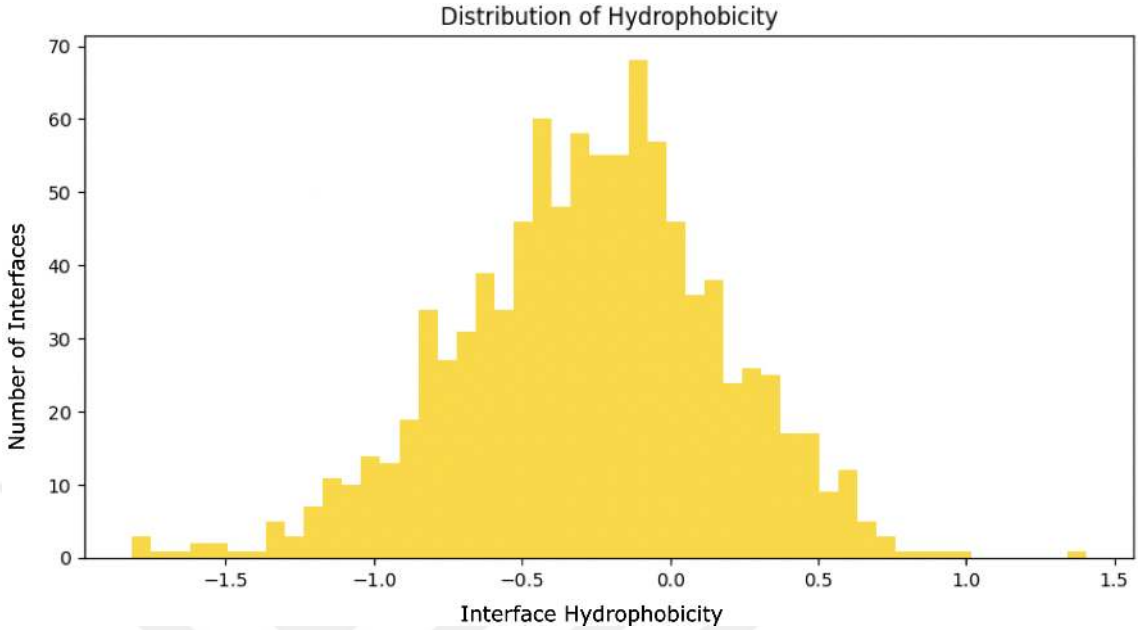


Figure 3.3: Hydrophobicity distribution of protein-protein interface dataset.

that most protein interfaces tend to have a balanced or low polarity, which may be conducive to stable protein-protein interactions.

The histogram in Figure 3.3 shows the distribution of hydrophobicity values for the protein interfaces. The distribution of hydrophobicity follows a bell-shaped curve centered around zero. This indicates that most protein interfaces exhibit a moderate hydrophobic character. The histogram shows that hydrophobicity values are symmetrically distributed around the mean, with a slight skew towards negative values. This suggests that while there is a balance between hydrophobic and hydrophilic residues, there is a slight tendency towards hydrophobicity in protein interfaces, which aligns with the role of hydrophobic interactions in stabilizing protein structures.

The utilization of this extensive data set aims to capture a wide and diverse spectrum of protein-protein interactions and their corresponding interfaces. The significant volume of unlabeled data facilitates the proposed model's ability to learn and generalize robust representations. This approach enhances the model's capacity to identify and understand the intricate patterns and relationships within protein

interfaces.

In the subsequent sections of this study, a detailed explanation is provided of how this data was meticulously processed and utilized in training the graph-based framework. The integration of the contrastive autoencoder and transformer model is explored, highlighting the methodological steps taken to investigate protein-protein interfaces. The process involves pre-processing the data to ensure its suitability for training, followed by the implementation of the graph-based contrastive autoencoder and transformer to learn meaningful representations.

## *3.2 Input Preparation & Graph Construction*

In this study, the input protein structures are represented as graphs, wherein residues are represented as nodes, and their interactions are captured through edges. Each node is encoded as a vector of length 30, encompassing the features: residue type, polarity, residue charge, relative accessible surface area (relASA) in the monomer and in the complex form, knowledge-based pair potential (PP), and backbone dihedral angles ($\phi$ and $\psi$). Naccess is used to calculate relASA [Ding and Arnold, 2006], and knowledge-based PP is obtained from Keskin *et. al.*[Keskin et al., 1998].

Naccess has various hyperparameters such as z-slices, probe size, Van der Waals values file, and parameters, whether to ignore HETATM records, hydrogens, and waters [Ding and Arnold, 2006]. In this study, default parameter values are used while running Naccess.

As in the previous study of Zang *et. al.*[Zhang et al., 2022], three distinct edge types are used in the graph representation:

- Sequential edges: An edge is established between the $i^{th}$ and $j^{th}$ residues if their sequential distance, denoted as $|j-i|$, falls below a predefined threshold. Specifically, an edge is created when $|j-i| < 3$, signifying proximity within the sequence.

- Radius edges: Besides sequential edges, edges between two nodes, i and j, are introduced, if the Euclidean distance between them is less than a speci-

fied threshold of 10 Å. This criterion ensures that residues in close physical proximity are interconnected.

- K-nearest neighbor edges: Recognizing the potential variations in spatial scales across different proteins, edges are incorporated by connecting each node to its k-nearest neighbors based on the Euclidean distance. K is set to 10 ($k = 10$). This strategy ensures a comparable density of spatial edges across diverse protein graphs, facilitating meaningful graph representations.

The detailed descriptions of the node and edge features can be found in 3.1. The size of the graph and, hence, the adjacency matrix vary according to the size of the input protein-protein interface. However, each node and each edge are represented with fixed-size vectors of length 30 and 3, respectively. The embedding dimension is, on the other hand, fixed to 128x512.

Table 3.1: Descriptions of node and edge features.

| **Attribute** | **Description** | **Size** |
|---|---|---|
| ***Node*** | | |
| Residue Type | Twenty different amino acids | 20 |
| Polarity | Polar or non-polar | 2 |
| Charge | Positively charged, negatively charged, or neutral | 3 |
| relMonASA | Relative accessible surface area in the monomer form | 1 |
| relCompASA | Relative accessible surface area in the complex form | 1 |
| Pair Potential | Knowledge-based pair potential | 1 |
| Phi ($\phi$) | Backbone dihedral angle | 1 |
| Psi ($\psi$) | Backbone dihedral angle | 1 |
| ***Edge*** | | |
| Radius Edge ($\psi$) | Closeness of two residues in Euclidean space ($<$ 10 Å) | 1 |
| Sequential Edge ($\psi$) | Closeness of two residues in the protein sequence ($< 3$) | 1 |
| KNN Edge ($\psi$) | K-nearest neighbors in Euclidean space (k=10) | 1 |

By employing these strategies, proposed graph-based approach enables comprehensive and informative encoding of the interface region, capturing sequential and

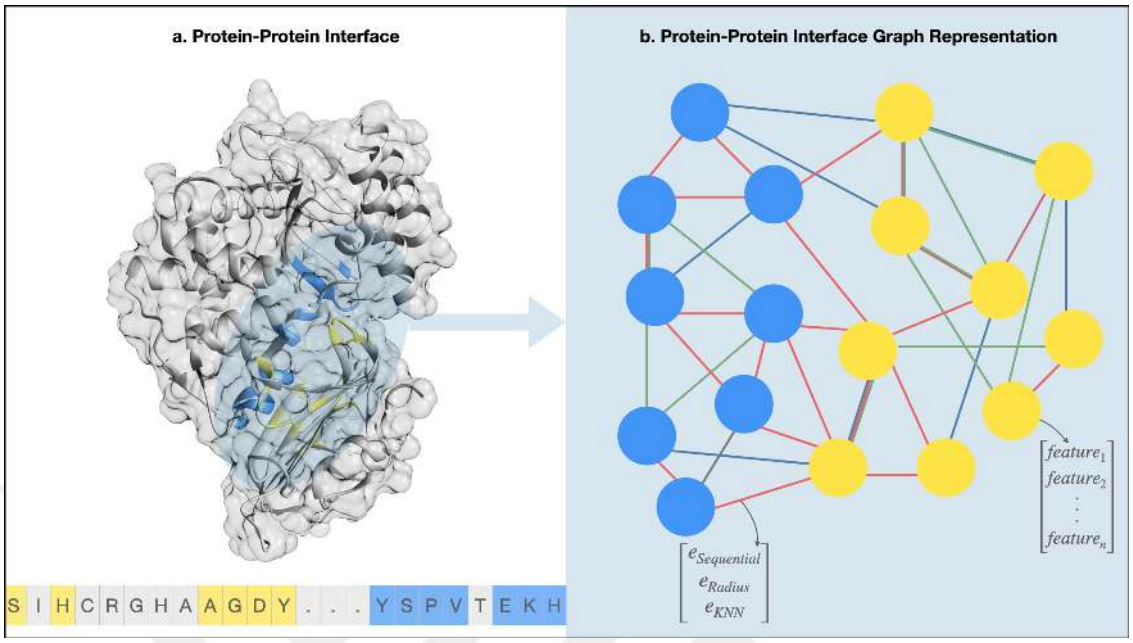


Figure 3.4: Protein-protein interface structure, sequence, and graph representation. (a) Visualization of a protein-protein complex (PDB ID: 1JTG_A_B) in three-dimensional (3D) structure, depicted as a surface and ribbon model. The sequence of the complex is displayed at the bottom, with interface residues highlighted in color. Residues from chain A are shown in blue, while residues from chain B are depicted in yellow. (b) The corresponding graph representation of the complex is presented. The graph consists of nodes, each representing a residue and its associated features, and edges, representing different types of interactions. Red: Radius Edges, Green: Sequence Edges, Blue: KNN Edges.

spatial relationships between residues within the protein complex. Figure 3.4 shows the 3D structure and the sequence of a protein and its graph representation. The nodes are represented as tensors, encompassing residue-specific features, while the edges are encoded as vectors of 1s and 0s, indicating the presence or absence of specific edge types.

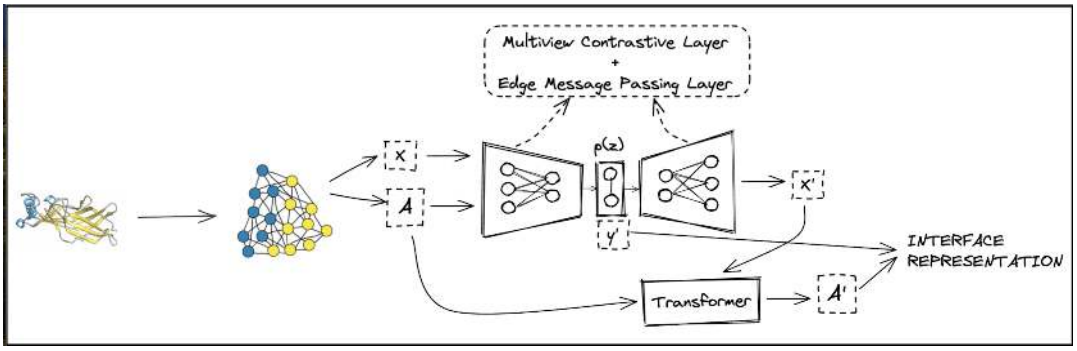


Figure 3.5: Protein-protein interface representation learning framework. The protein-protein interface structure is represented as a residue-level graph. The graph is inputted into a graph-based autoencoder combined with a transformer to learn the representation.

## *3.3 Model Architecture*

### *3.3.1 Graph-based Contrastive Autoencoder*

After the construction of the graph, it is passed through a novel graph autoencoder architecture combined with a transformer module to obtain a latent representation. The model is visually depicted in Figure 3.5. The encoder component inputs the node feature matrix $X$ and the adjacency feature tensor $A$, effectively encoding them into a continuous latent representation denoted as $z$. Conversely, given a point in the latent space, the decoder generates an output node feature matrix $X'$. Additionally, utilizing a disentangled latent space, the model predicts graph-level properties, yielding the graph property prediction denoted as $y'$.

*Multiview Contrastive Layer*

A multiview contrastive learning layer was added to the autoencoder to enhance the model's performance. The same approach taken by Zhang *et. al.*[Zhang et al., 2022] is employed. The multiview contrastive learning involved the application of two cropping functions, namely subsequence, and subspace.

1. A consecutive protein segment was randomly sampled in the subsequence crop-

ping, and the corresponding subgraph was extracted.

2. In the subspace cropping, a center residue was first sampled, and then all residues within a specified distance threshold were selected to form the subgraph.

Furthermore, a random transformation function was applied to the output subgraph, with options including identity (no transformation) and random edge masking (random removal of a fixed ratio of edges from the graph).

*Edge Message Passing Layer*

A message-passing layer is added to enhance the model performance further, as in a prior study[Zhang et al., 2022]. In message passing, a relational graph, commonly called a line graph, establishes connections among edges. Each node in the line graph represents an edge from the original graph. Specifically, an edge $(i, j, r_1)$ in the original graph is linked to an edge $(w, k, r_2)$ in the line graph if and only if $j = w$ and $i \neq k$. Subsequently, a relational graph convolutional network is applied to the line graph, enabling the derivation of the message function for each edge. The message function of an edge is updated by aggregating features from its adjacent edges.

### *3.3.2 Transformer*

The model incorporates a transformer component for edge prediction in graph-based architectures inspired by the study of J. Mitton *et al* [Mitton et al., 2021]. The transformer takes the adjacency feature tensor and the predicted node feature matrix as input and generates a predicted edge feature tensor. The utilized transformer architecture follows a vanilla implementation [Vaswani et al., 2017].

The vanilla transformer architecture is a neural network framework initially designed for natural language processing tasks. At its core, it relies on a multi-head self-attention mechanism, which allows the model to dynamically weigh the importance of different input elements. It consists of an encoder-decoder structure,

with both components comprising identical layers. Each layer contains two main sub-layers: a multi-head self-attention mechanism and a position-wise feedforward neural network that processes the attended representations, thus enabling the transformer to capture complex relationships and patterns in data.

During the generation phase, the transformer sequentially adds edges to the graph. At each step, the input to the transformer consists of the node encoding of adjacent nodes to the predicted edge and the previously predicted edges in the graph. This iterative process allows the model to progressively construct the graph by making informed edge predictions based on the available information.

The model effectively captures dependencies and interactions between nodes in the graph by employing a transformer-based approach, enabling accurate edge predictions. The iterative generation of edges ensures that the model incorporates the evolving graph structure while maintaining consistency with the existing edges in the graph.

The objective function of the transformer model can be expressed as:

$$L = \lambda_0 L_n + \lambda_1 L_{KL} + \lambda_2 L_h + \lambda_3 L_t \tag{3.1}$$

where $\lambda_i$ represents scaling constants with the values $\lambda_0 = 0.2$, $\lambda_1 = 0.3$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.2$; $L_n$ denotes the cross entropy loss for reconstructing node features, $L_{KL}$ corresponds to the KL divergence loss to calculate the statistical distance between the input graph and the latent space representation, $L_h$ represents the mean squared error loss for predicting graph properties from the latent space, and $L_t$ represents the cross entropy loss for reconstructing edge features.

Cross-entropy loss measures the difference between two probability distributions: the true labels and the predicted probabilities. Specifically, cross-entropy loss quantifies how well the predicted probabilities match the actual labels by calculating the negative log-likelihood of the true labels given the predicted probabilities. In this study, the cross-entropy loss is calculated between the original node features (the

true labels) and the reconstructed features (the predicted probabilities). For each node, the loss measures how well the predicted probability distribution matches the actual distribution of the node features.

Kullback-Leibler divergence loss, or KL divergence loss, measures the difference between two probability distributions. It is used to quantify how much one probability distribution (the predicted distribution) diverges from a reference probability distribution (the true distribution). Mathematically, it is expressed as:

$$D_{KL}(P||Q) = \sum_i P(i) \log \left( \frac{P(i)}{Q(i)} \right) \tag{3.2}$$

where $P$ is the true distribution and $Q$ is the predicted distribution.

Mean Squared Error (MSE) loss is a commonly used loss function that measures the average squared difference between the actual (true) values and the predicted values by the model. Mathematically, it is defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.3}$$

where $y_i$ represents the actual values, $\hat{y}_i$ represents the predicted values, and $n$ is the number of data points. MSE penalizes larger errors more heavily due to the squaring of differences, making it sensitive to outliers.

The starting values of lambdas in the objection function given in Equation 3.1 were $\lambda_0 = \lambda_1 = \lambda_2 = \lambda_3 = 0.25$. As it is aimed to balance multiple objectives in the loss function, i.e., cross-entropy loss for reconstructing node features, KL divergence loss to calculate the statistical distance between the input graph and the latent space representation, the mean squared error loss for predicting graph properties from the latent space, and the cross entropy loss for reconstructing edge features, lambda values are normalized to sum up to 1. Then, different combinations of lambda values within a specified range were tried, i.e., lambda values were increased and decreased by 0.05 each time, and their impact on model performance was evaluated.

The model performs best with the values $\lambda_0 = 0.2, \lambda_1 = 0.3, \lambda_2 = 0.3, and \lambda_3 = 0.2$.

To optimize the model, an Adam optimizer is employed with an initial learning rate of $5e^{-4}$.

## *3.4 Results and Discussion*

### *3.4.1 Ablation Study*

An ablation study is a systematic analysis technique used to evaluate the performance of different components within a model or system. By selectively removing or modifying certain parts of the model, researchers can identify the contributions and importance of each component to the overall performance. This method helps in understanding which features, layers, or parameters are most critical and which may be redundant or less impactful.

In order to assess the individual contributions of specific components in the model, ablation experiments were conducted on the PPI validation task. Key elements were systematically removed, namely the multiview contrastive layer and the edge message passing layer. Then, each model was trained with the DeepInterface training set and the prediction accuracy of learned representations for the PPI validation was compared on the test set. This rigorous analysis allowed us to evaluate the impact of each component on the overall performance of the model and gain insights into their respective roles in the learning process. As shown in Table 3.2, the results can be significantly improved by using multiview contrastive learning and the prediction performance further increases after performing edge message passing.

**Multiview contrastive layer** is designed to capture the similarity between correlated protein substructures by constructing informative views of the substructures. The layer's objective is to preserve the similarity between these substructures before and after mapping them to a latent space. By creating multiple views that reflect different aspects of the protein substructures, the multiview contrastive layer enables the model to capture and preserve the intricate relationships and similarities among the substructures. This contributes to the overall effectiveness of the model

Table 3.2: Ablation experiment results of graph autoencoder, multiview contrastive layer, and edge message passing layer on the PPI validation task.

| **Model** | **Accuracy** |
|---|---|
| Graph autoencoder | 0.764 |
| Multiview contrastive graph autoencoder | 0.888 |
| Graph autoencoder with edge message passing | 0.879 |
| ProInterVal | 0.923 |

in capturing the complex nature of protein interactions and representations.

**The edge message passing layer** can be regarded as a modified version of the pair representation update specifically tailored for graph neural networks. Drawing inspiration from AlphaFold2 [Jumper et al., 2021], which utilizes the triangle attention mechanism in transformers to capture pair representations, the edge message passing layer aims to capture the interdependence among various interactions involving a residue and its sequentially or spatially adjacent residues. Unlike the triangle attention in AlphaFold2, the proposed approach incorporates angular information to effectively model diverse types of edge interactions, enabling more efficient sparse edge message passing operations.

### *3.4.2 t-SNE and UMAP Analysis*

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique used to reduce the dimensionality of data in order to visualize the similarities among high-dimensional data points. It starts by calculating pairwise similarities between points in the high-dimensional space, modeled using a Gaussian distribution. These similarities are then mirrored in the low-dimensional space using a Student's t-distribution, which helps maintain the local structure of the data. The algorithm minimizes the Kullback-Leibler divergence between the high-dimensional and low-dimensional similarities through iterative optimization, ensuring that points close in the original space remain close in the reduced space, thus revealing clusters and patterns in the

data.

To analyze the learned protein interface representations, t-SNE was employed to project them onto a two-dimensional (2D) space. A random sample set of 10,000 samples was selected, representing five function classes: antibodies, enzymes, receptors, hormones, and transporters. Within this sample set, there were 1,879 antibodies, 849 enzymes, 1,572 receptors, 2,012 hormones, and 3,688 transporters. A t-SNE plot was generated to visualize the clustering patterns of these classes.

Figure 3.6 showcases the t-SNE visualization of protein function clusters, offering valuable insights into the arrangement of protein functions. This visualization demonstrates the potential of the proposed approach in comprehending protein interactions and verifying their biological significance. As depicted in Figure 3.6, the proteins exhibit clustering based on their functional classes in the embedding space, indicating that the model has captured the functional fingerprints of the proteins to some extent.

Uniform Manifold Approximation and Projection (UMAP) is another dimensionality reduction technique that has been developed as an effective tool for visualizing high-dimensional data in a low-dimensional space, typically two or three dimensions. UMAP is particularly known for preserving both the local and global structure of the data, which makes it suitable for various applications, including clustering and classification tasks.

UMAP works by constructing a high-dimensional graph representation of the data, where each point represents a data sample, and edges represent the relationships or similarities between these points. This graph is then optimized to reflect the manifold structure of the data, capturing both local neighborhoods and global patterns. Once the high-dimensional graph is constructed, UMAP projects this graph into a lower-dimensional space while preserving the relationships between the points as closely as possible.

The primary advantages of UMAP over other dimensionality reduction techniques, such as t-SNE, are its computational efficiency and its ability to maintain both local and global structures. UMAP tends to be faster and more scalable,

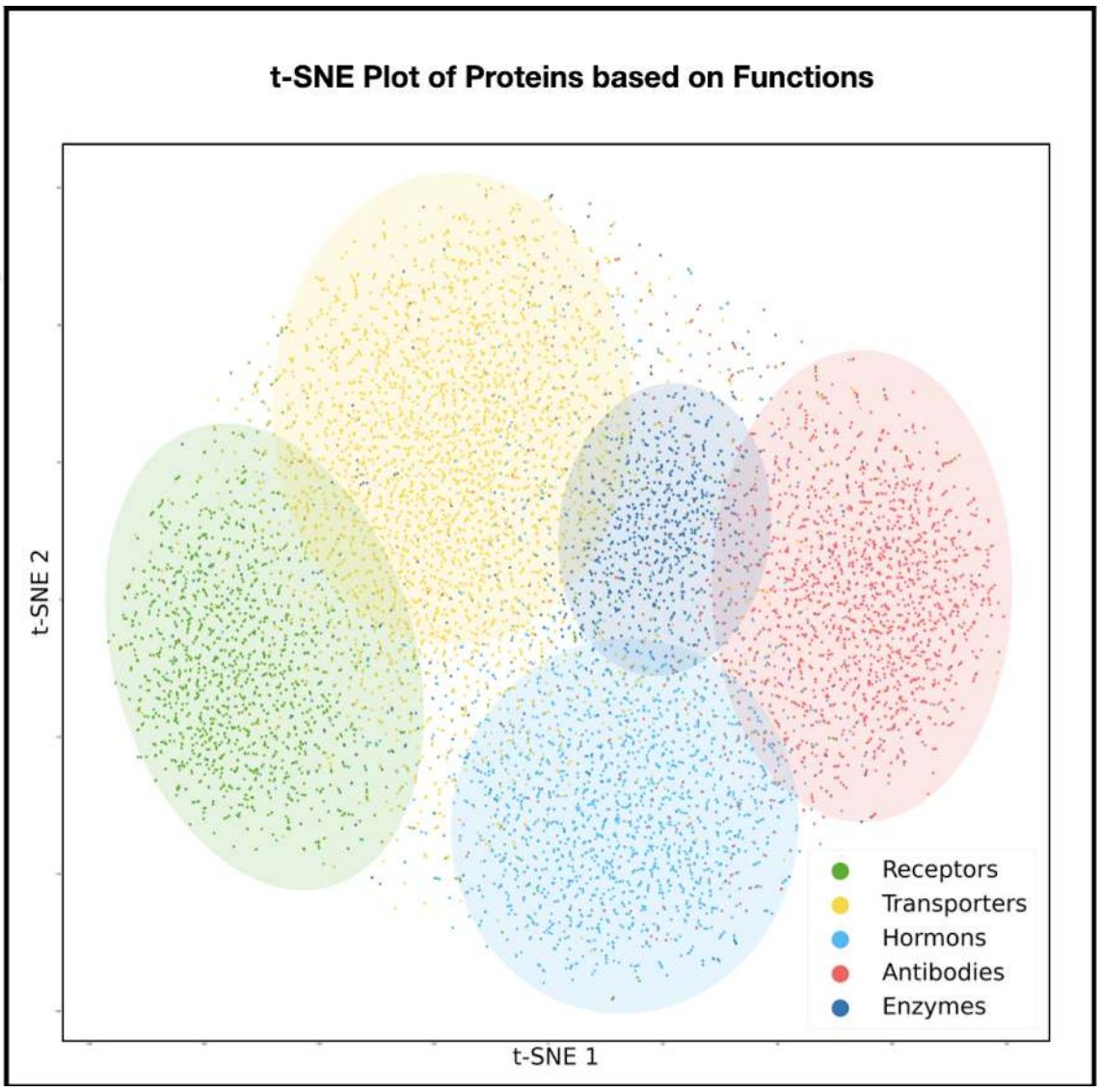


Figure 3.6: t-SNE graph of representations learned by the model. The t-SNE plot represents the clustering of protein functions based on learned representations. Each point corresponds to a protein and is color-coded according to its functional category.

making it suitable for large datasets. Moreover, UMAP often produces more interpretable visualizations by preserving the global structure of the data, allowing for better identification of clusters and patterns within the data.

For further analysis, a random subset of 1000 protein-protein interface samples was selected from the overall dataset, representing PPIs of varying sizes, ranging from 19 to 659 residues. Out of these samples, 597 interfaces have fewer than 100 residues, 282 interfaces contain between 100 and 200 residues, 72 interfaces have between 200 and 300 residues, and 49 interfaces consist of more than 300 residues. Among these, 417 interfaces are primarily composed of non-polar residues, meaning less than 50% of the residues in each interface are polar. The remaining 583 interfaces are primarily composed of polar residues. Additionally, the dataset includes 489 hydrophilic interfaces and 511 hydrophobic interfaces. Two dimensionality reduction techniques, t-SNE, and UMAP, were employed to visualize the learned representations of these interfaces.

The visualizations of the protein-protein interfaces based on size and polarity, as presented in the tSNE and UMAP plots in Figure 3.7 and Figure 3.8 respectively, reveal that similar-sized and similar-polarity PPIs do not form distinct clusters in two-dimensional space. This observation can be attributed to several underlying factors.

Firstly, the multifaceted nature of protein-protein interactions must be considered. Size and polarity are among many attributes that characterize PPIs, but they do not singularly define their structure and function. PPIs of similar size or polarity can exhibit vastly different structural properties, interaction dynamics, and functional roles. These differences are likely captured by the embeddings produced by the proposed model, leading to a dispersion rather than distinct clustering based on these simple characteristics.

Additionally, the characteristics of the learned embedding space play a crucial role. The model is designed to represent data by capturing the most significant features for accurate reconstruction. These features might not be directly associated with the size or polarity of the PPIs. Instead, the embeddings might emphasize other

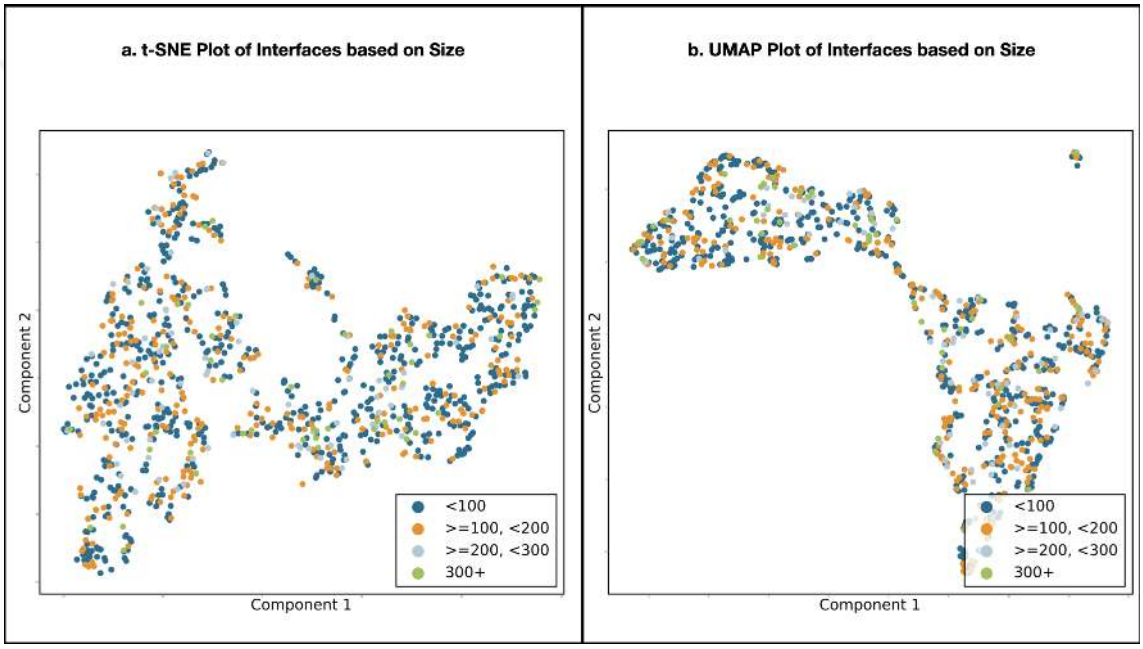


Figure 3.7: t-SNE and UMAP graph of representations learned by the model. These plots represent the clustering of interface sizes based on learned representations. Each point corresponds to an interface and is color-coded according to its size category.

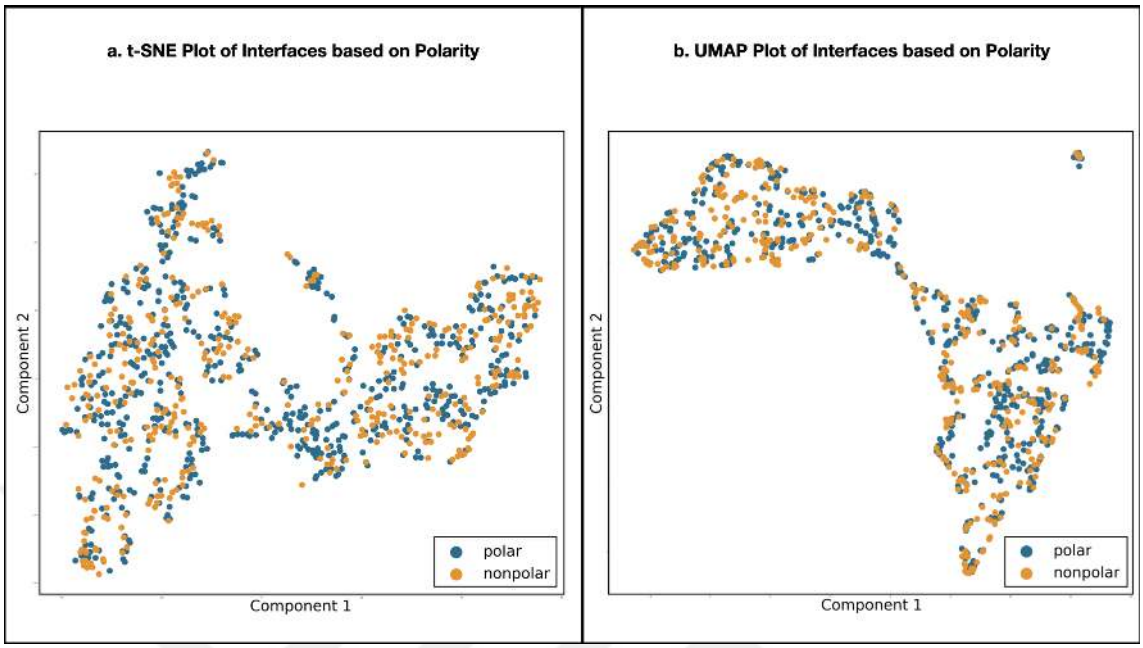


Figure 3.8: t-SNE and UMAP graph of representations learned by the model. These plots represent the clustering of interface polarity based on learned representations. Each point corresponds to an interface and is color-coded according to its polarity.

structural motifs, interaction patterns, or biologically relevant features that are not directly linked to size or polarity. The model may prioritize capturing other subtle but biologically significant patterns essential for understanding PPIs, such as binding affinities, conformational changes, or interaction types, rather than emphasizing size or polarity alone.

The role of dimensionality reduction techniques, such as tSNE and UMAP, further contributes to the observed distribution. Both techniques are non-linear dimensionality reduction methods that aim to preserve the local structure of the data while projecting it into a lower-dimensional space. They are sensitive to the high-dimensional relationships intrinsic to the data and may not cluster similar-sized or similar-polarity PPIs together if those PPIs differ significantly in other critical aspects. Consequently, the global arrangement in the reduced-dimensional space might not reflect simple characteristics like size or polarity.

The tSNE and UMAP visualizations of PPIs based on hydrophobicity are de-

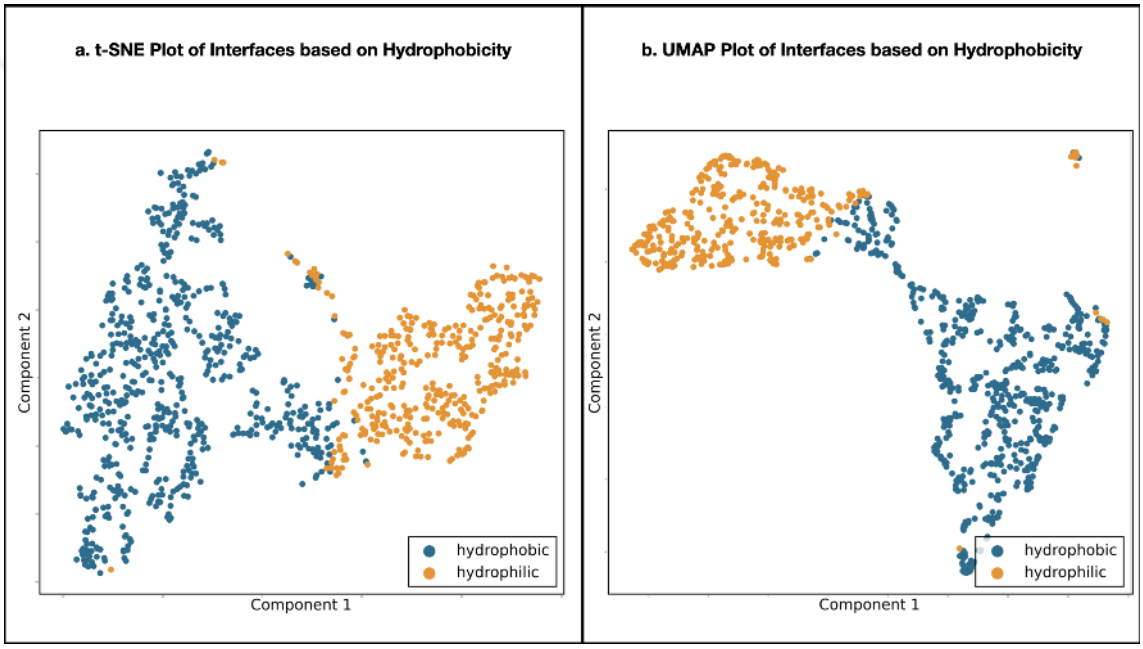


Figure 3.9: t-SNE and UMAP graph of representations learned by the model. These plots represent the clustering of interface hydrophobicity based on learned representations. Each point corresponds to an interface and is color-coded according to its hydrophobicity.

picted in Figure 3.9. These graphs demonstrate a clear clustering of hydrophobic and hydrophilic PPIs in the two-dimensional space. This distinct clustering, in contrast to the dispersion observed with size and polarity, can be because of several key factors, including hydrophobicity is a fundamental property influencing the behavior and interaction of proteins.

Hydrophobic and hydrophilic regions in proteins tend to drive specific structural arrangements and interaction patterns. Hydrophobic regions often aggregate together to minimize exposure to the aqueous environment, leading to distinct structural features that the model consistently captures. Consequently, the embeddings reflect these pronounced structural tendencies, resulting in clear clustering based on hydrophobicity.

Moreover, the model likely prioritizes hydrophobicity as a significant feature during the learning process. Given that hydrophobic interactions play a crucial role in protein folding, stability, and interaction, the model's embeddings are heavily influenced by this property. This prioritization leads to a more coherent and distinguishable representation of hydrophobic and hydrophilic PPIs in the reduced-dimensional space.

Additionally, hydrophobicity impacts the overall physicochemical environment within which proteins operate. Hydrophobic and hydrophilic regions create distinct local environments that affect protein dynamics and function. These environments are captured by the model and preserved during the dimensionality reduction process, resulting in the observed clustering.

In summary, the distinct clustering of PPIs based on hydrophobicity in the tSNE and UMAP visualizations can be attributed to the fundamental role of hydrophobic interactions in protein structure and function. The model's emphasis on hydrophobicity, the preservation of local structures by dimensionality reduction techniques, and the robust influence of hydrophobicity on protein behavior collectively contribute to the observed clustering. This analysis highlights the importance of hydrophobicity as a critical factor in understanding protein-protein interactions and the effectiveness of embedding and visualization techniques in capturing biologically

meaningful features.

## *3.5 Concluding Remarks*

The multiview contrastive autoencoder-transformer approach introduced in this chapter represents a significant advancement in the field of protein-protein interface representation learning. This method successfully captures the intricate structural and functional nuances of protein-protein interfaces by leveraging the strengths of graph-based models combined with the transformative capabilities of autoencoders and transformers.

The proposed method offers several potential advantages for representing protein-protein interfaces:

1. Graph-based representations can capture the complex network of interactions between amino acid residues at protein-protein interfaces, allowing for a more holistic and interpretable characterization of interface properties.

2. By incorporating multiple data modalities (views) into the learning process, the model can capture the similarity between correlated interface substructures to enhance the quality and robustness of the learned representations.

3. Contrastive learning techniques enable the model to disentangle informative features from noisy data, facilitating the extraction of discriminative and biologically relevant interface features.

4. The edge message passing layer captures the interdependence among various interactions involving a residue and its sequentially or spatially adjacent residues.

5. The transformer architecture, known for its effectiveness in capturing long-range dependencies, can provide a robust framework for learning representations from protein sequence data and capturing contextual information within protein-protein interfaces.

The comprehensive dataset, encompassing over half a million samples, provided a robust foundation for training and evaluating the model. This extensive dataset enabled the model to generalize effectively across a wide range of protein interfaces, thereby enhancing its applicability and reliability.

The results of various evaluations, including ablation studies, t-SNE, and UMAP analysis, demonstrate the model's ability to produce highly accurate and interpretable representations. These representations not only capture the essential structural characteristics of protein interfaces but also reveal biologically relevant insights, thereby facilitating a deeper understanding of protein-protein interactions.

Incorporating multiple data modalities and contrastive learning techniques proved instrumental in disentangling informative features from noisy data. This, combined with the edge message passing layer and the transformer architecture, allowed the model to capture both local and global dependencies within protein structures, resulting in a more holistic representation of protein-protein interfaces.

In conclusion, the multiview contrastive autoencoder-transformer approach represents a promising and innovative solution to the challenges of protein-protein interface representation learning. By providing a more accurate and interpretable representation of protein-protein interfaces, this method paves the way for future research.

# Chapter 4

# EXPERIMENTAL ANALYSIS

## *4.1 Protein-Protein Interface Validation*

This chapter explains the methodology and framework developed for validating protein-protein interfaces. Validating these interfaces is crucial for understanding the biological relevance of protein interactions, which are fundamental to numerous cellular processes, including signaling pathways, metabolic networks, and structural assembly. The chapter begins by detailing the datasets used for training and evaluating the validation model. These datasets include a variety of protein-protein interfaces from diverse sources, providing a comprehensive foundation for assessing the model's performance.

Next, the chapter describes the architecture of the validation model. A Graph Neural Network (GNN) framework is introduced, leveraging graph-based representations of protein-protein interfaces. The model architecture is designed to effectively distinguish between biologically relevant interfaces and those that are merely artifacts of crystallization or other experimental conditions. The core of this chapter focuses on the validation model training process. Following the training methodology, the chapter presents the evaluation metrics and experimental results. Additionally, the chapter discusses the results in detail, interpreting the implications of the findings and their relevance to biological research. Finally, the chapter concludes with a summary of the key findings and their significance.

### *4.1.1 Data*

*DeepInterface Data Set*

The DeepInterface data set [Balci et al., 2019] served as a crucial component for the validation of protein-protein interactions. This data set was obtained from four sources: DOCKGROUND [Liu et al., 2008], PPI4DOCK [Yu and Guerois, 2016], PIFACE [Cukuroglu et al., 2014], and PDB [Sussman et al., 1998].

Positive interfaces were primarily obtained from PIFACE, which contains interface structures extracted from PDB files published until 2012. Additional positive interfaces were extracted from PDB files deposited after 2012 to avoid redundancy. Negative interfaces were derived from DOCKGROUND and PPI4DOCK decoy sets, using CAPRI's classification [Janin et al., 2003]. In this work, a positive interface refers to the interface found in a protein-protein complex that exists in the Protein Data Bank (PDB)—assuming it is biologically relevant. Conversely, a negative interface corresponds to the interface observed in a docked protein-protein complex with an unacceptable Critical Assessment of Predicted Interactions (CAPRI) score—meaning it has a fraction of native contacts (Fnat) value lower than 0.1, ligand root mean square deviation (LRMSD) greater than 10Å, and interface root mean square deviation (iRMSD) greater than 4Å. The PDB and chain IDs of positive samples as well as model names of negative decoys and their division as training, validation, and test, are available through the GitHub repository at https://github.com/ku-cosbi/ProInterVal.

Table 4.1 showcases the quantitative distribution of samples from each source, emphasizing their contribution to the data set and the partitioning scheme into distinct subsets for training, validation, and test purposes.

To ensure a fair comparison, the model was evaluated alongside two graph-based docking model evaluation models, namely GNN-DOVE[Wang et al., 2021] and DeepRank-GNN[Réau et al., 2023]. The evaluation was performed on three data sets: The deepInterface data set, the GNN-DOVE data set, and the DeepRank-GNN data set.

Table 4.1: Distribution of positive and negative protein interface data from diverse sources within the DeepInterface data set, illustrating the partitioning scheme into distinct subsets for training, validation, and test purposes.

| | PIFACE | PPI4DOCK | DOCKGROUND | PDB2012 | TOTAL |
|---|---|---|---|---|---|
| Training | 127,790 | 127,790 | - | - | 255,580 |
| Validation | - | - | 4,632 | 4,632 | 9,264 |
| Test | - | - | 4,632 | 4,632 | 9,264 |
| TOTAL | 127,790 | 127,790 | 9,264 | 9,264 | 274,108 |

*GNN-DOVE Data Set*

GNN-DOVE utilizes a combination of the Dockground data set[Liu et al., 2008] and ZDOCK [Hwang et al., 2010], as well as a CAPRI scoring data set [Lensink and Wodak, 2014]. To eliminate redundancy, the complexes were grouped based on sequence alignment and TM-align [Zhang and Skolnick, 2004]. Specifically, two complexes were assigned to the same group if any pair of proteins from the two complexes exhibited a TM-score greater than 0.5 and a sequence identity of 30% or higher. Dockground and ZRANK are used for training and validation, while the CAPRI scoring data set is designated for testing. Dockground, ZRANK, and CAPRI datasets are downloadable from http://dockground.compbio.ku.edu/downloads/unbound/decoy /decoys1.0.zip, https://zlab.umassmed.edu/zdock/decoys_bm4_zd3.0.2_6deg.tar.gz, and http://cb.iri.univ-lille1.fr/Users/lensink/Score_set respectively.

*DeepRank-GNN Data Set*

DeepRank-GNN data set is composed of the Docking Benchmark version 5 (BM5) [Vreven et al., 2015]. The BM5 comprises a non-redundant set of 142 complexes (dimers), excluding antibody-antigen complexes and complexes with more than two chains. A total of 25,300 models per complex were generated using the integrative modeling software HADDOCK [Van Zundert et al., 2016]. The test set consisted of all docking models generated for 15 randomly selected complexes (379,500 models,

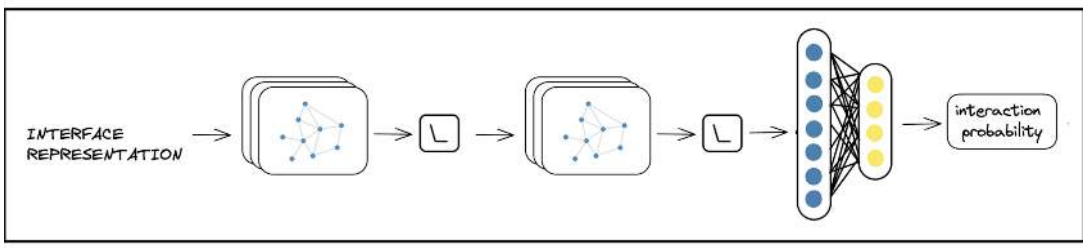


Figure 4.1: Protein-protein interface validation framework. The learned representation is passed through a GNN model to obtain the interaction probability.

10% of the data set), while the remaining 127 complexes were split into a training set (80%, 102 complexes) and a validation set (20%, 25 complexes). DeepRank-GNN dataset is available from https://data.sbgrid.org/dataset/843/ [Kuntz and Privé, 2024].

### *4.1.2 Model Architecture*

The GNN model consists of a graph convolution layer (GCL), a ReLU activation, a max pooling layer, and a fully-connected layer. The cross-entropy loss was used as the loss function to train the model, and the Adam optimizer was used for parameter optimization. The model was trained for 20 epochs with a batch size of 32. The model is visually depicted in Figure 4.1. The learned representation is passed through the model, and the interaction probability is outputted.

The input and output sizes of each layer in the architecture are summarized in 4.2.

### *4.1.3 Results and Discussion*

A comprehensive performance comparison of the model was performed with the two state-of-the-art methods, GNN-DOVE and DeepRank-GNN, for validating protein-protein interactions. To ensure a fair evaluation, all models were trained from scratch on three distinct training sets and evaluated on their respective test sets. The details of these sets are explained in the data set section. The results are evaluated based

on the following metrics:

$$\begin{aligned} Accuracy &= (TP + TN)/(TP + TN + FP + FN) \\ Sensitivity(Recall) &= TP/(TP + FN) \\ Specificity(Spec.) &= TN/(TN + FP) \\ Precision &= TP/(TP + FP) \\ MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \end{aligned} \tag{4.1}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Notably, the analysis revealed that the proposed method consistently outperformed the other models across all three data sets. Figure 4.5 visually depicts the performance of ProInterVal compared to GNN-DOVE and DeepRank-GNN on these data sets. These findings highlight the robustness and efficacy of the ProInterVal model in accurately validating PPIs, underscoring its potential as a powerful tool in computational protein studies. Recently a study based on 2D images of proteins has been published [Stebliankin et al., 2023]. Different from the proposed method, it is not a graph-based method, and has an ROC-AUC score of 0.81 on the CAPRI score data set, while the proposed method's ROC-AUC score on the DeepInterface data set, constructed based on CAPRI classification, and the GNN-DOVE data set, sourced from the CAPRI score data set is 0.88.

ProInterVal distinguishes itself from the compared models by leveraging learned representations of protein-protein interfaces, whereas the other models rely on handcrafted features to represent these interfaces. Protein function is governed by a wide range of structural motifs, characterized by intricate spatio-chemical arrangements of atoms and amino acids. Handcrafted features struggle to comprehensively capture these patterns. Comparative modeling faces challenges in detecting these motifs, as the set of sequence/conformational perturbations that preserve function remains unknown. Therefore, this study introduces an approach that utilizes learned interface representations for validating protein-protein interfaces. Furthermore, as outlined

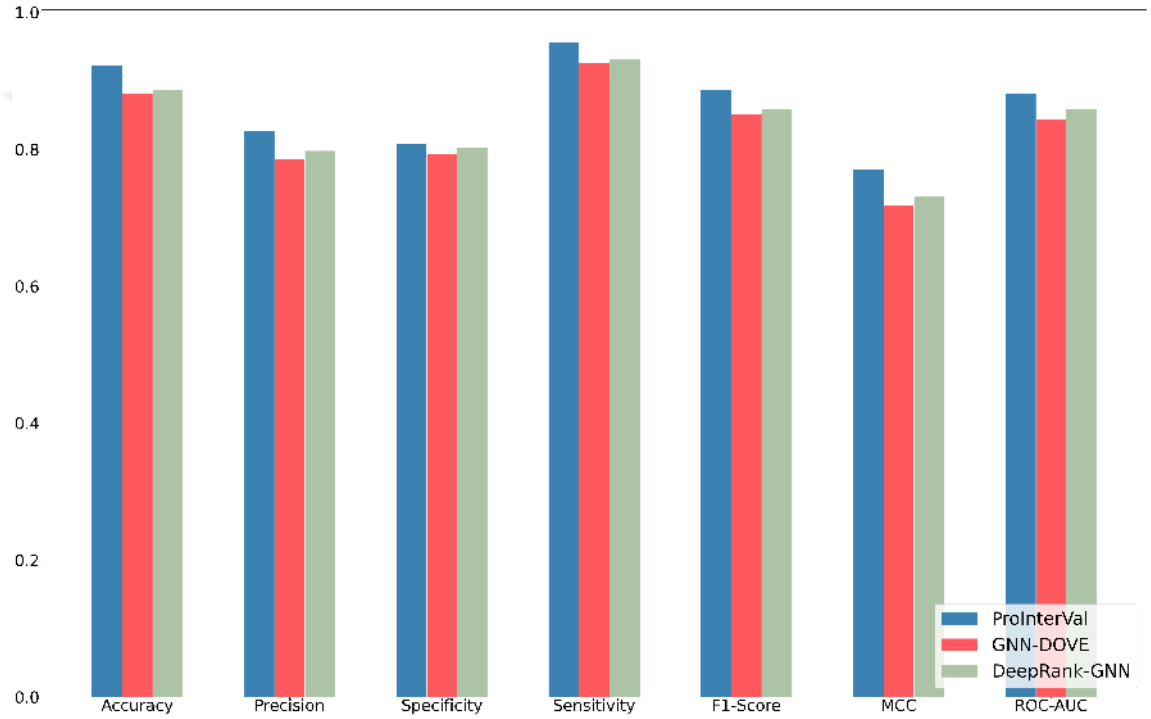


Figure 4.2: Accuracy, precision, specificity, sensitivity, F1-score, MCC, and ROC-AUC values of ProInterVal, GNN-DOVE, and DeepRank-GNN on the DeepInterface test set, after retraining the models from scratch on the training set of DeepInterface dataset.

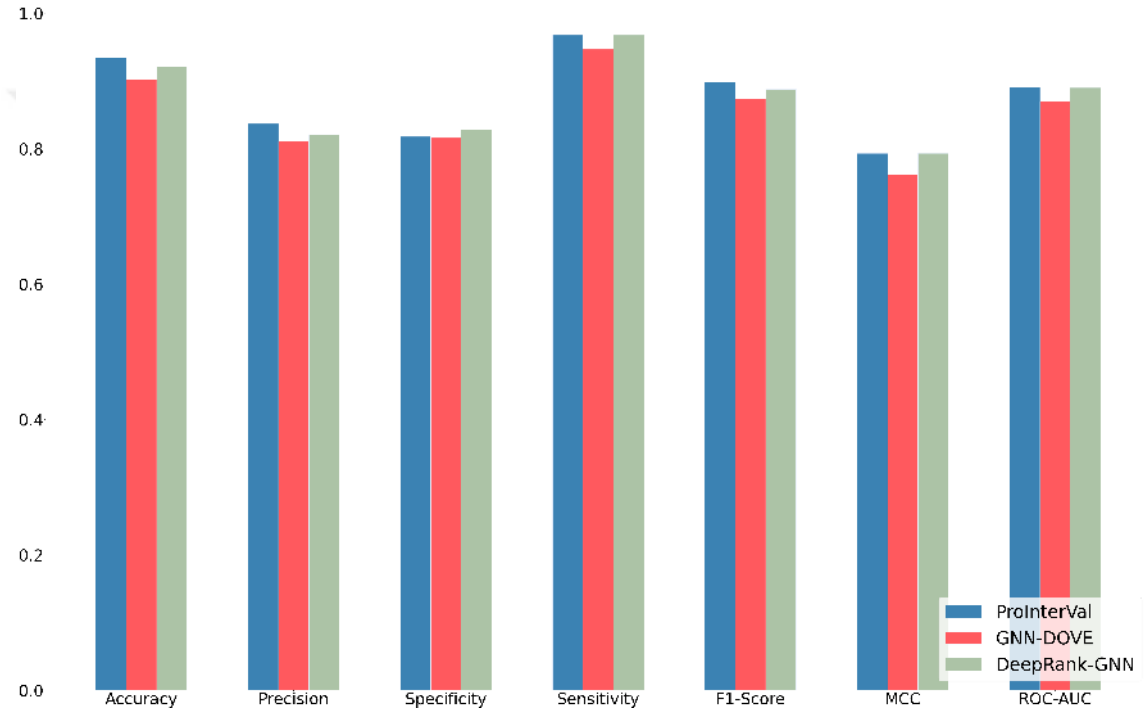


Figure 4.3: Accuracy, precision, specificity, sensitivity, F1-score, MCC, and ROC-AUC values of ProInterVal, GNN-DOVE, and DeepRank-GNN on the DeepRank-GNN test set, after retraining the models from scratch on the training set of DeepRank-GNN dataset.

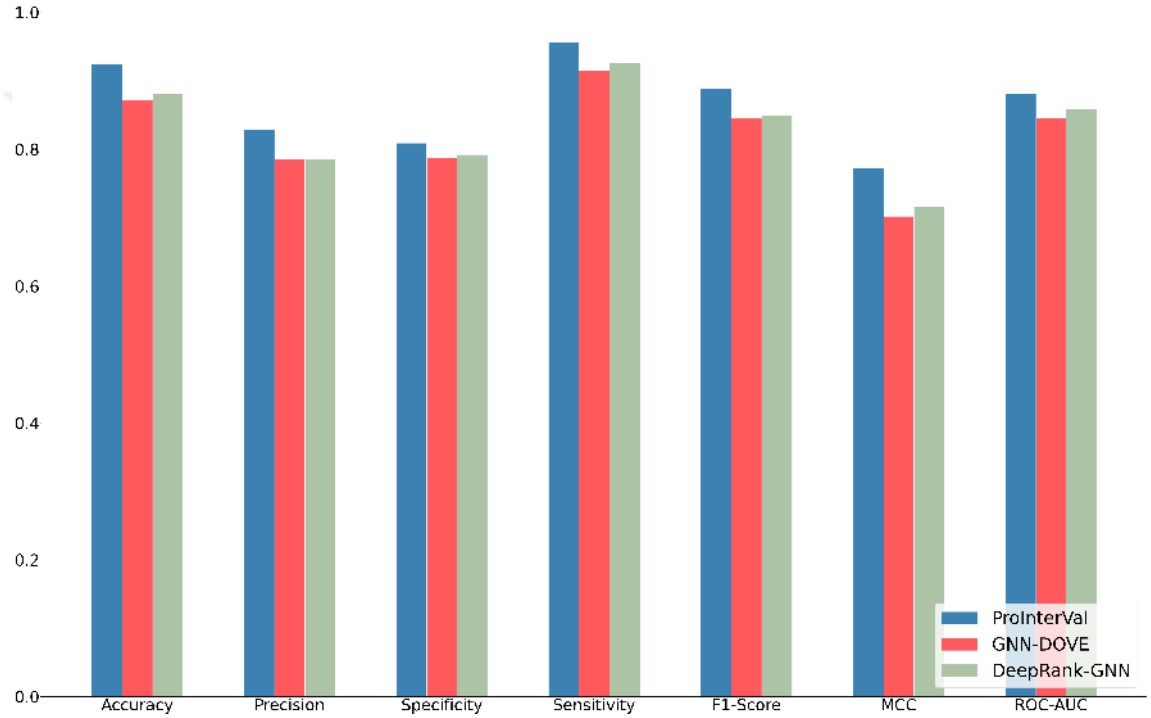


Figure 4.4: Accuracy, precision, specificity, sensitivity, F1-score, MCC, and ROC-AUC values of ProInterVal, GNN-DOVE, and DeepRank-GNN on the GNN-DOVE test set, after retraining the models from scratch on the training set of GNN-DOVE dataset.

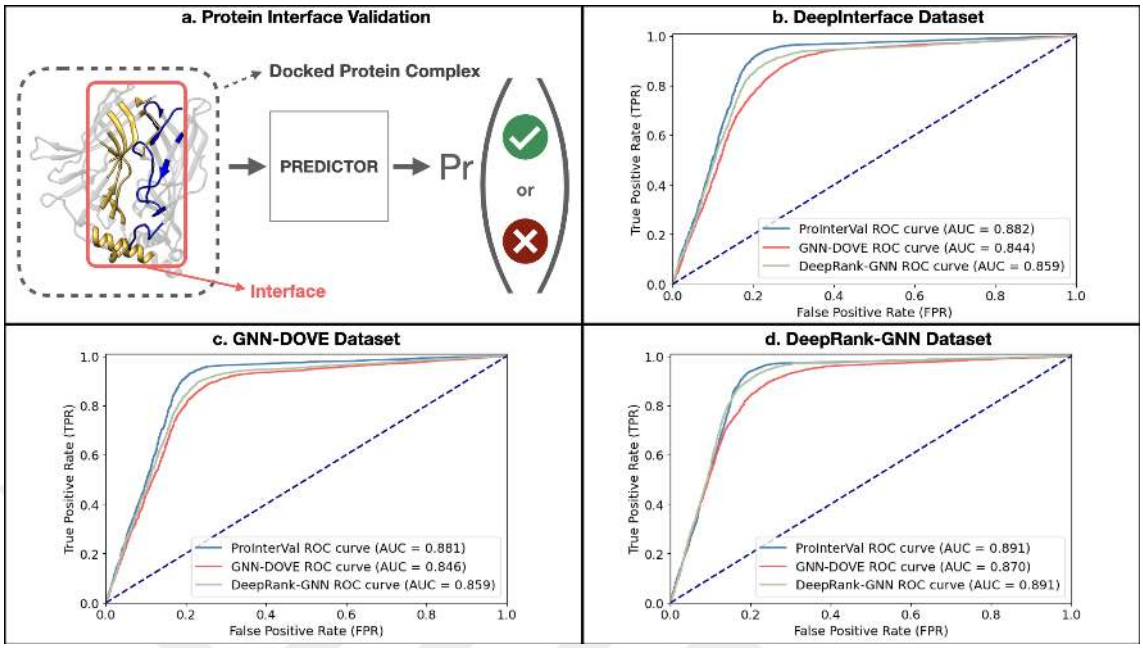


Figure 4.5: Protein-protein interface validation and performance comparison of ProInterVal, GNN-DOVE, and DeepRank-GNN on multiple data sets. (a) Illustration depicting the concept of protein-protein interface validation, where the predictor generates a probability for the docked complex interface to be a biological complex. (b) Performance evaluation results of ProInterVal, GNN-DOVE, and DeepRank-GNN on the DeepInterface test set, after retraining the models from scratch on the training set of DeepInterface data set. (c) Performance evaluation results of ProInterVal, GNN-DOVE, and DeepRank-GNN on the GNN-DOVE test set, after retraining the models from scratch on the training set of GNN-DOVE data set. (d) Performance evaluation results of ProInterVal, GNN-DOVE, and DeepRank-GNN on the DeepRank-GNN test set, after retraining the models from scratch on the training set of DeepRank-GNN data set.

in the methodology section, the model is trained on a data set of over 500,000 interfaces, enabling it to capture diverse information and exhibit generalizability.

A ranking analysis was conducted to assess its performance in distinguishing a single native interface (positive sample) from a set of approximately 100 incorrect docking models (negative samples). This evaluation involved comparing the performance of the proposed model with GNN-DOVE and DeepRank-GNN, as well as the docking models' scoring function ZRANK [Pierce and Weng, 2007].

100 positive protein-protein complexes were randomly selected from the DeepIn-

terface test data set to carry out this analysis. For each of these 100 complexes, 300 decoys were generated using the GRAMM-X protein docking tool [Tovchigrechko and Vakser, 2006]. From the pool of generated decoys, 100 incorrect decoys per complex were carefully selected based on the CAPRI criteria. As a result, a data set consisting of 100 incorrect decoys and one native structure per complex was obtained.

Next, ProInterVal, ZRANK, GNN-DOVE, and DeepRank-GNN were applied to score and rank these structures. Specifically, the performance of these scoring tools was assessed by examining the position at which the native PDB structure (positive interface) was ranked among the 100 incorrect decoys (negative interface). When referred to as the "top 10%," it signifies that the near-native interface is positioned within the uppermost 10% of its respective ranked list. Figure 4.6 indicates the ranking results. The model achieved a 63% success rate in the top 1% and a 90% success rate in the top 10% and outperformed the others. In practical terms, this indicates that the proposed model placed the native structure among the top 1% for 63 out of the selected 100 complexes and within the top 10% for 90 out of these 100 complexes.

The remarkable performance of the proposed model in interface ranking highlights the potential of integrating it into template-based protein-protein interaction (PPI) predictors to improve their overall predictive capabilities.

### *4.1.4 Concluding Remarks*

This chapter encapsulates the pivotal strides made in validating protein-protein interactions through advanced machine-learning techniques. The proposed model represents a significant enhancement over existing methods such as GNN-DOVE and DeepRank-GNN. By leveraging learned representations of protein interfaces rather than relying on handcrafted features, ProInterVal demonstrates a superior ability to capture the intricate structural motifs and spatial-chemical arrangements essential for accurate validation.

The comprehensive evaluation of the model across diverse datasets, including the

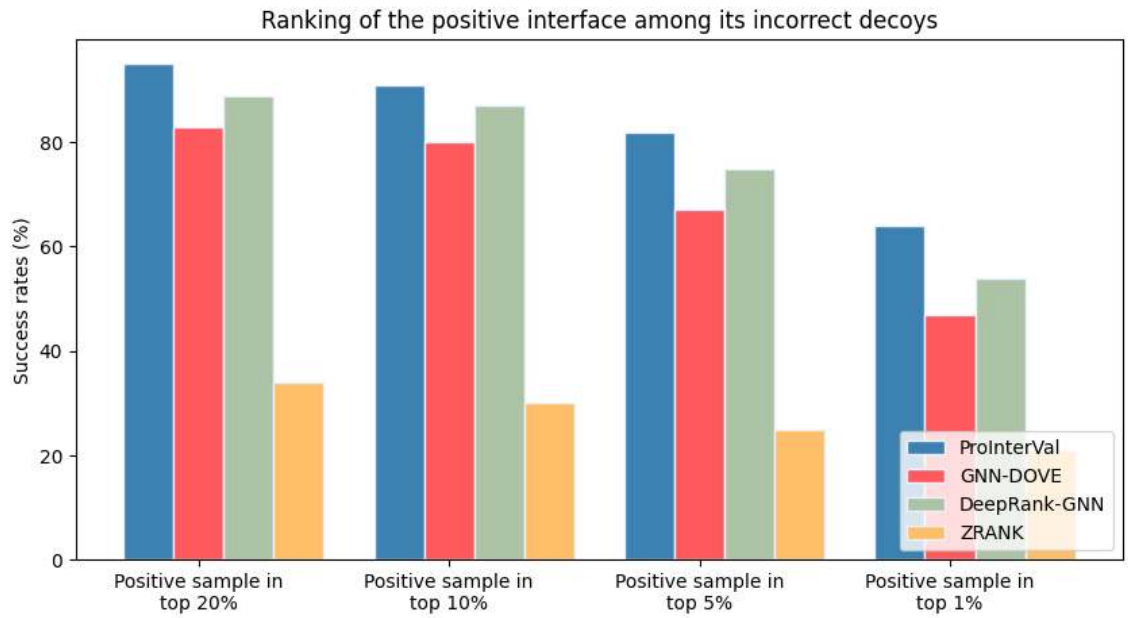


Figure 4.6: Ranking of the positive interface among the incorrect decoys of the corresponding complex.

DeepInterface, GNN-DOVE, and DeepRank-GNN datasets, underscores its robustness and efficacy. The model consistently outperforms its counterparts, as evidenced by higher accuracy, precision, specificity, sensitivity, F1-score, MCC, and ROC-AUC values. Notably, the ranking analysis, which positions native interfaces among incorrect docking models, highlights the model's remarkable performance, achieving a 63% success rate in the top 1% and a 90% success rate in the top 10% of rankings.

The detailed analysis and comparison with existing scoring functions, such as ZRANK, further validate the superiority of the proposed model in distinguishing biologically relevant protein interfaces from decoy models. The methodological rigor applied, including retraining models from scratch on various training sets and evaluating on respective test sets, ensures a fair and transparent comparison.

In conclusion, the advancements presented in this chapter significantly enhance the landscape of protein-protein interface validation. The success of the proposed model not only validates its immediate application in computational protein studies but also sets a new benchmark for future research in the domain. This model's

ability to provide accurate, scalable, and interpretable validation of protein-protein interactions holds immense potential for advancing our understanding of molecular interactions, with far-reaching implications in biomedical research and drug discovery.

## *4.2 Biological vs. Crystal Classification*

This chapter explains the methodology and framework for distinguishing between biological and crystallographic interfaces within protein complexes. Discrimination between biologically significant interfaces and crystallographic contacts within protein complexes is paramount for establishing associations between cellular functions and the corresponding macromolecular assemblies. X-ray crystallography is one of the most widely used experimental techniques for determining the 3D structure of protein complexes. While most X-ray structures are reliable and of good quality, crystallographic PPIs may be formed as a byproduct of the crystal formation process. Therefore, accurately identifying these interfaces is essential for interpreting protein structures determined through X-ray crystallography. Understanding the differences helps identify biologically relevant interactions, enhancing our knowledge of protein functions and interactions.

The chapter begins by describing the dataset, which includes a balanced selection of protein complexes featuring both biological and crystallographic interfaces. Next, it outlines the model architecture. Following this, evaluation metrics and experimental results are presented, assessing the model's performance using accuracy, precision, recall, and F1-score. Comparative analyses with existing methods highlight the advantages of the proposed approach. Finally, the chapter concludes with a summary of key findings and their broader implications, suggesting directions for future research.

### *4.2.1 Data and Methods*

In this study, the MANY/DC benchmark data set comprising 5739 dimers sourced from the MANY data set [Baskaran et al., 2014] and 80 biological and 81 crystal

interfaces, characterized by similar interface areas, from DC data set [Duarte et al., 2012] were employed, which are available from https://data.sbgrid.org/dataset/843/ [Kuntz and Privé, 2024]. The training set encompassed 80% of the dimers from MANY data set, with the remaining 20% constituting the validation set. Subsequently, the model was tested using the DC data set. The MANY and DC datasets feature a balanced distribution of biological and crystal dimers (∼50%/50%), the latter stemming from crystal packing. Distinguishing crystal dimers from biological ones is challenging without specific knowledge about the complex, as they appear similar. Although biological interfaces typically exhibit larger surface areas than crystal dimers across various datasets, the DC dataset has been meticulously curated to include biological and crystal dimers with comparable interface areas.

Another benchmark data set of 1677 homodimer protein crystal structures [Schweke et al., 2023], which included a balanced mix of physiological and non-physiological complexes, is used for further evaluation. It contains 836 structures classified as physiological dimers, identified using two complementary techniques. These techniques involve maintaining the three-dimensional structure of the interface across homologs, as determined by QSalign [Dey et al., 2018, Dey et al., 2022], and across crystal forms in the PDB, as described by ProtCID [Xu and Dunbrack Jr, 2020]. The non-physiological set was constructed using two methods. First, QSalign was used to find homodimers with two unique and conserved interfaces, one of which was not observed in homologs. This initial step produced 141 assemblies with an interface area distribution similar to that of physiological dimers. This additional criterion ensured the creation of a non-physiological set that is challenging to distinguish from physiological dimers based on interface area alone. Due to the lower number of non-physiological dimers (141 total) compared to the physiological set, a second technique was employed to balance the quantities. This involved identifying non-physiological dimers among structures with sufficient crystal forms. Specifically, interfaces distinct from all others across crystal forms, as indicated by the ProtCID database, were selected [Schweke et al., 2023]. The selection was biased towards dimers with interface areas larger than or comparable to their physiological counter-

parts. This process added 700 non-physiological interfaces, resulting in a total of 841 non-physiological interfaces, achieving a balanced dataset with physiological interactions. The data is accessible from https://github.com/vibbits/Elixir-3DBioInfo-Benchmark-Protein-Interfaces [Schweke et al., 2023].

The same GNN architecture explained in Chapter 4.1 was used. The model is trained using the MANY data set described above for 50 epochs, and the model with the lowest loss in the validation set was selected as the final model.

### *4.2.2 Web Server*

This section introduces a web server designed to predict class scores for protein-protein interface structures, effectively distinguishing between biologically relevant and crystallographic interfaces. The server is accessible without registration at `https://interactome.ku.edu.tr/prointerval-bioxtal/`.

The tool addresses the fundamental challenge in structural biology by leveraging protein representation learning and advanced graph-based deep learning techniques to analyze critical features that distinguish genuine biological interfaces from those formed in protein crystals.

The web server accepts input in the form of a PDB file containing the 3D structure of a dimer. Upon receiving the input, the server extracts the interface of the complex and constructs an interface graph. This graph is subsequently fed into a novel graph-based contrastive model to generate interface representations. Finally, the interface is classified by a graph neural network model whose input is the learned interface representation (Figure 4.7.a).

Users can enter a 4-letter PDB ID to fetch protein-protein complex structure from PDB or upload their local file in PDB format. In case the structure file contains a multimer with more than two chains, users need to specify two chain identifiers of which they want to classify the interface.

After specifying the required inputs, the user clicks the submit button to navigate to the result page. The result page displays two predicted probabilities for the biological and crystal classes. The model usually takes a few seconds or less to make

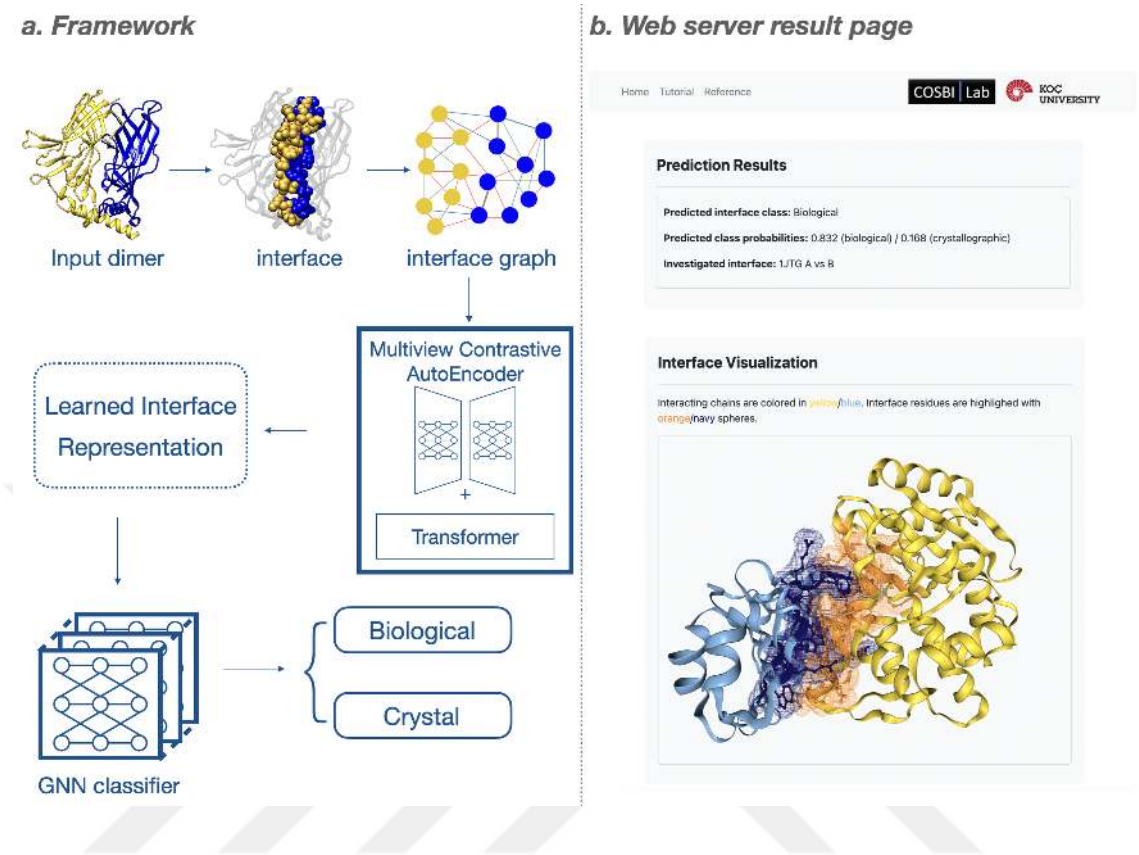


Figure 4.7: (a) Overall framework. A dimer input complex undergoes interface extraction, followed by interface graph construction. The resulting graph is inputted into the proposed novel graph-based contrastive model to produce interface representations, subsequently classified by a graph neural network model. (b) Result page. The upper section displays the predicted class and corresponding class probabilities for the input protein-protein interface. In the lower segment, the complex structure is visualized, with the interface residues highlighted.

predictions. The input interface is labeled with the class with a higher probability value. At the bottom of the page is a visualization section showing the structure of the input protein-protein complex with two partners with a cartoon representation. Partners of the complex are colored in blue and yellow. The interface residues of the blue partner are highlighted with navy spheres, while the interface residues of the yellow partner are highlighted with orange spheres (Figure 4.7.b).

### *4.2.3 Results and Discussion*

The evaluation of the proposed method was conducted using the DC dataset, as described in Section 4.2.1. The performance of the method was compared against several existing classifiers, including DeepRank-GNN, PRODIGY-CRYSTAL, EP-PIC 3, PISA, and QSAlign. The proposed method achieved approximately 88% accuracy, 88% precision, and an 85% F1 score on the test set, outperforming the compared methods. The detailed performance results are presented in Table 4.3.

Additionally, the proposed method was evaluated on another benchmark data set [Schweke et al., 2023]. This study reported that among the existing methods, DeepRank-GNN performs the best with a 0.85 AUC score. Our method achieved 83% accuracy and 0.91 AUC score on this benchmark set, surpassing the performance of DeepRank-GNN. Although Schweke et al. showed that a cross-validated Random Forest (RF) classifier can achieve an AUC score of 0.94, this method requires calculating at least 20 features using various external tools. Furthermore, its performance could not be assessed on the DC test due to the unavailability of feature calculation and classifier codes.

The superior performance of the proposed method underscores its potential as a leading approach in the field of protein-protein interface classification. By combining speed and accuracy, the method ensures that users can achieve reliable results promptly. The user-friendly web server further enhances its accessibility, providing a valuable resource for researchers to make accurate predictions and gain deeper insights into the complex interactions between biological and crystallographic interfaces in protein structures.

The integration of speed and accuracy is crucial for efficient analysis of protein-protein interface structures. Traditional classifiers often struggle with a trade-off between these two aspects, but the proposed approach successfully balances them. This balance enables extensive and detailed analyses without compromising on time, facilitating a more streamlined workflow and faster progression in research.

The web server's user-friendly interface ensures that researchers of varying expertise levels can effectively utilize the tool, broadening its applicability and impact

within the scientific community. By offering reliable predictions, the method significantly contributes to understanding the intricate interplay between biological and crystallographic interfaces in protein structures. Accurate classification of these interfaces is vital for numerous applications, including drug discovery, molecular biology, and structural bioinformatics.

By offering reliable predictions, the proposed method significantly contributes to our understanding of the intricate interplay between biological and crystallographic interfaces in protein structures. Accurate classification of these interfaces is vital for numerous applications, including drug discovery, molecular biology, and structural bioinformatics.

In summary, the proposed method outperforms existing classifiers by effectively combining speed and accuracy. Its user-friendly web server offers a valuable resource for reliable predictions and enhanced insights into the complex interactions between biological and crystallographic interfaces in protein structures. This method not only streamlines the analysis process but also contributes to significant advancements in understanding protein-protein interactions, ultimately supporting the broader goals of research and discovery in structural biology and related disciplines.

### *4.2.4 Concluding Remarks*

The distinction between biological and crystallographic interfaces within protein complexes is a critical aspect of structural biology, with significant implications for understanding cellular functions and molecular interactions. The research presented in this chapter introduces an advanced method for accurately classifying these interfaces, addressing a fundamental challenge in the field.

The utilization of the MANY/DC benchmark datasets, featuring a balanced distribution of biological and crystal dimers, provided a robust basis for training and evaluating the proposed method. The meticulous curation of these datasets to include dimers with comparable interface areas highlights the rigor of the experimental design and the reliability of the results obtained.

The development of a novel graph-based contrastive model, combined with a

graph neural network for classification, represents a significant advancement over existing methods. This approach effectively captures the unique characteristics of protein-protein interfaces, leveraging advanced deep-learning techniques to enhance classification accuracy. The model's performance, achieving approximately 88% accuracy, 88% precision, and an 85% F1 score on the test set, surpasses several established classifiers, including DeepRank-GNN, PRODIGY-CRYSTAL, EPPIC 3, PISA, and QSAlign.

The integration of a user-friendly web server further extends the accessibility and applicability of this method. By providing a platform for researchers to input protein-protein complex structures and obtain rapid, reliable predictions, the web server facilitates broader adoption and practical use of the classification tool. The detailed visualization of predicted interfaces enhances interpretability and supports deeper insights into protein interactions.

In summary, the advancements presented in this chapter significantly enhance the capability to distinguish between biological and crystallographic interfaces in protein complexes. The proposed method's superior performance and accessibility via the web server underscore its potential as a valuable tool in structural biology, with broad implications for drug discovery, molecular biology, and bioinformatics. The successful application of this method paves the way for further research and development, ultimately contributing to a deeper understanding of protein-protein interactions and their role in cellular functions.

## *4.3 Gene Ontology Term Prediction*

This chapter explains the comprehensive analysis conducted to evaluate the effectiveness of the proposed approach in predicting gene ontology (GO) terms using a well-established dataset. Gene ontology annotations provide crucial insights into the molecular functions, biological processes, and cellular components associated with proteins, thereby enhancing our understanding of their roles within biological systems.

This study focuses on specific classes of proteins chosen to represent distinct

functional categories. The primary objectives of this study were to predict various facets of protein functionality. To benchmark the performance of our model, it is compared with the existing state-of-the-art methods in the field.

### *4.3.1 Data and Methods*

The Structural Classification of Proteins—extended (SCOPe) dataset [Fox et al., 2014] is a comprehensive resource used for the classification and analysis of protein structures. It extends the original SCOP database [Murzin et al., 1995], providing a detailed and hierarchical organization of protein structural data. The SCOPe dataset categorizes proteins based on their structural and evolutionary relationships, offering valuable insights into the commonalities and differences among various protein families.

The SCOPe dataset [Fox et al., 2014] is organized into several levels of classification, including:

- Class: This is the highest level of classification, grouping proteins based on their overall structural architecture. Classes include all-alpha proteins, all-beta proteins, alpha/beta proteins, and others.

- Fold: Within each class, proteins are further categorized into folds based on their major structural features and the arrangement of secondary structures.

- Superfamily: This level groups proteins that have low sequence similarity but share a common structural framework and are believed to have a common evolutionary origin.

- Family: Proteins within a superfamily are grouped into families if they have a high degree of sequence similarity and often share similar functions.

The SCOPe dataset is widely used in structural bioinformatics and computational biology for various applications, including protein structure prediction, functional annotation, and evolutionary studies. Its hierarchical organization facilitates

a detailed understanding of protein structures and their relationships, making it an invaluable tool for researchers in the field.

In this chapter, we utilize the SCOPe dataset to evaluate the performance of our approach in predicting gene ontology (GO) terms. We randomly selected 1000 proteins from the dataset corresponding to the equal number of proteins for the following classes: ATP binding (GO:0005524) and Heme binding (GO:0020037). Among these, 600 proteins were allocated for training, 200 for validation, and 200 for testing.

By selecting proteins from specific functional classes and dividing them into training, validation, and testing subsets, we ensure a thorough and balanced assessment of our model. The SCOPe dataset's rich and well-annotated structural information provides a robust foundation for our analysis, enabling us to accurately predict molecular functions, biological processes, cellular components, and structural classifications.

The GNN architecture detailed in Chapter 4.1 was employed. The model underwent training with the aforementioned data for 50 epochs, and the version that achieved the lowest loss on the validation set was chosen as the final model.

### *4.3.2 Results and Discussion*

The objective of this study was to predict various aspects of protein functionality, including molecular function, biological process, cellular components, and fold classification. To thoroughly assess the performance of the proposed model, a comparative analysis against the geometric structure pretraining models developed by Zhang et al. [Zhang et al., 2022] was conducted. This comparison allowed us to benchmark the proposed model's effectiveness and reliability. The results of this evaluation, which are detailed in Table 4.4, highlight the accuracy the proposed model achieved on the test set, showcasing its capability in accurately predicting the targeted protein characteristics.

The results presented in Table 6.1 demonstrate the superiority of the proposed representation learning model over the geometric structure pretraining models devel-

oped by Zhang et al. [Zhang et al., 2022]. The proposed model consistently achieved higher prediction accuracies across all categories, namely molecular function, biological process, cellular component, and fold classification. This improvement can be attributed to the extensive dataset used for training, which included a larger and more diverse set of protein structures than those utilized by Zhang et al.[Zhang et al., 2022]. This comprehensive training set enabled the model to capture a broader range of protein interactions and structural variations, thereby enhancing its predictive power and generalizability. However, it is important to note that despite the improvements, the accuracies achieved are still relatively low, indicating that there is significant room for further development and refinement of the model.

### *4.3.3 Concluding Remarks*

In this chapter, the effectiveness of the proposed representation learning model is evaluated on the gene ontology term prediction task. The results, detailed in Table 4.4, clearly demonstrate the superior performance of the proposed model in predicting molecular function, biological process, cellular component, and fold classification compared to the geometric structure pretraining models by Zhang et al. This success can largely be attributed to the extensive and diverse dataset used for training, which allowed the proposed model to capture a wide range of protein interactions and structural variations.

However, despite these notable improvements, the accuracies achieved by the proposed model are still relatively low, indicating that there is significant room for further development and refinement. This underscores the complexity of protein function prediction and the necessity for continuous advancements in data quality, model architecture, and training methodologies. Enhancing these aspects will be crucial to fully harness the potential of deep learning models in accurately predicting gene ontology terms and other protein-related features.

The implications of the findings of this study are manifold. Firstly, the superior performance of the proposed model highlights the value of using larger and more diverse datasets in training, which can lead to more robust and generalizable models.

Secondly, the need for further improvement suggests that integrating additional biological knowledge, optimizing model parameters, and exploring novel architectures could significantly enhance predictive accuracy. Thirdly, the potential applications of the proposed model are vast, ranging from drug discovery to understanding disease mechanisms, where accurate protein function prediction is pivotal.

Moving forward, future research should focus on several key areas. Enhancing the dataset with more comprehensive and high-quality annotations will be vital. Incorporating multi-modal data, such as combining structural, sequence, and functional data, could provide a more holistic view of protein interactions. Additionally, developing more sophisticated model architectures that can better capture the complexities of protein structures and their interactions will be essential.

In conclusion, while the proposed representation learning model marks a significant advancement in the field of protein function prediction, it also opens up new avenues for research and development. The current limitations highlighted by this analysis provide a clear direction for future work, aiming to improve the accuracy and applicability of these models in various biological and medical contexts. By continuing to refine and expand upon the proposed approach, it is possible to contribute to a deeper understanding of protein functionalities and interactions, ultimately advancing the field of computational biology and bioinformatics.

Table 4.2: Summary of the inputs, hyperparameters and outputs of each architecture component.

| **Layer** | **Input Size** | **Output Size** | **# Parameters** |
|---|---|---|---|
| ***GCN Autoencoder*** | | | |
| Input (Node size) | 30 | - | - |
| GCN Layer | 30 | 64 | 64*30+64+64 |
| GCN Layer | 64 | 32 | 64*32+32+32 |
| GCN Layer | 32 | 16 | 32*16+16+16 |
| GCN Layer | 16 | 128 | 16*128+128+128 |
| Multiview Contrastive | 128 | 30 | 128*30+30 |
| Edge Message Passing | 30 | 30 | 30*30+30 |
| GCN Layer | 128 | 16 | 128*16+16+16 |
| GCN Layer | 16 | 32 | 16*32+32+32 |
| GCN Layer | 32 | 64 | 32*64+64+64 |
| GCN Layer | 64 | 30 | 64*30+30+30 |
| ***Tranformer*** | | | |
| Input | 30 | - | - |
| Embedding Layer | 30 | 128 | 30*128+128+128 |
| Multi-Head Self-Attention | 128 | 128 | 4*(128*128+128+128) |
| Feed-Forward Network | 128 | 128 | 2*(128*128+128+128) |
| ***GNN*** | | | |
| Input | 128 | - | - |
| GCL Layer | 128 | 64 | 128*64+64 |
| ReLU | 64 | 64 | - |
| Max pooling | 64 | 32 | - |
| FC Layer | 32 | 2 | 32*2+2 |

Table 4.3: Performance comparison of the proposed method, DeepRank-GNN, PRODIGY-CRYSTAL, EPPIC 3, PISA, and QSAlign.

| Method | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| PISA | 0.78 | 0.81 | 0.77 | 0.79 |
| EPPIC 3 | 0.73 | 0.71 | 0.76 | 0.73 |
| PRODIGY-CRSYTAL | 0.74 | 0.78 | 0.76 | 0.77 |
| QSAlign | 0.79 | 0.76 | 0.82 | 0.79 |
| DeepRank-GNN | 0.81 | 0.82 | 0.80 | 0.81 |
| **Proposed method** | **0.88** | **0.88** | **0.83** | **0.85** |

Table 4.4: Gene Ontology (GO) Term Prediction Accuracy. MF: Molecular Function, BP: Biological Process, CC: Celular Component, FC: Fold Classification.

| Method | MF | BP | CC | FC |
|---|---|---|---|---|
| GearNet | 0.356 | 0.503 | 0.414 | 0.284 |
| GearNet-IEConv 3 | 0.381 | 0.563 | 0.422 | 0.423 |
| GearNet-Edge | 0.403 | 0.580 | 0.450 | 0.440 |
| GearNet-Edge-IEConv | 0.400 | 0.581 | 0.430 | 0.483 |
| GearNet-Contrastive | 0.490 | 0.654 | 0.488 | 0.541 |
| **Proposed method** | **0.541** | **0.698** | **0.520** | **0.632** |

# Chapter 5

# UNVEILING INTERFACES WITH EXPLAINABLE AI

This chapter delves into the integration of explainable AI techniques, specifically LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), within the proposed multiview-contrastive GCL autoencoder model. These techniques are employed to analyze node features, providing a comprehensive understanding of the learned graph representations.

To demonstrate the efficacy of the approach, a detailed case study is presented, applying GNNExplainer, highlighting the model's ability to identify evolutionarily conserved hot spot residues within protein-protein interactions.

## *5.1 Input Features Importance Analysis*

Feature importance analysis is useful for identifying and ranking the significance of different features in a predictive model. This analysis helps understand which features contribute most to the prediction accuracy of the model. In this study, residues are represented by various features as explained in Section 3.2. Their interaction with each other is also represented using three different edge types (Section 3.2). The proposed method, which uses these features, has achieved successful predictions on various tasks. Understanding which features drive the predictions helps explain what the model has learned and ensures transparency.

To analyze feature importance, the SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017] method is integrated into the architecture. SHAP values are a unified measure of feature importance based on cooperative game theory. They provide a way to explain the output of any machine learning model by fairly distributing the contribution of each feature to the prediction.

Originating from cooperative game theory, the Shapley value assigns a fair payout

to players (features) based on their contribution to the total payout (prediction). Each feature's contribution is calculated by considering all possible combinations of features and measuring how much the feature contributes on average [Lundberg and Lee, 2017].

SHAP provides both local and global interpretability to any machine learning model: SHAP values explain individual predictions by showing how each feature contributes to a particular prediction and aggregating SHAP values across the dataset provides insights into the overall importance of each feature in the model.

Local Interpretable Model-agnostic Explanations (LIME) method is also used for the analysis. LIME is designed to interpret the predictions of any machine learning model. It provides explanations by approximating the model locally with an interpretable model, allowing for an understanding of individual predictions [Ribeiro et al., 2016]. LIME operates under the principle that simpler, interpretable models can locally approximate complex models.

LIME generates a set of perturbed instances around the prediction of interest. These perturbations are slight variations of the original instance, created by randomly altering the values of its features. The complex model is used to predict the outcomes for these perturbed instances. This step generates a set of predictions corresponding to the perturbed data. Each perturbed instance is assigned a weight based on its proximity to the original instance. Instances closer to the original instance are given higher weights, emphasizing their relevance in the local approximation. An interpretable model, typically a linear model, is trained on the perturbed instances and their associated predictions from the complex model. The weights assigned in the previous step are used to emphasize the importance of instances close to the original. The coefficients of the trained interpretable model provide insights into the importance of each feature in the prediction of the original instance. These coefficients are used to generate an explanation that highlights the contribution of each feature to the model's prediction [Ribeiro et al., 2016].

The analysis of the contributions of input features to the reconstruction loss in the representation learning model explained in Section 3.3 was achieved through

the application of SHAP and LIME. Due to the variable number of nodes in each input PPI graph, node-level features were aggregated to create fixed-length graph-level features. Mean pooling was employed to aggregate the features, ensuring each graph was represented by a consistent feature vector. A surrogate model was trained to predict the reconstruction loss based on the aggregated graph-level features. A Support Vector Regressor (SVR) was selected for this purpose. A sample set of 2000 PPIs was selected from the PPI data set. The data was split into training and testing sets to evaluate the surrogate model's performance.

First, SHAP was utilized to interpret the surrogate model's predictions. The SHAP explainer was initialized with the training data and used to compute SHAP values for the test data. These SHAP values were converted to standard Python data types to ensure compatibility with visualization tools. The SHAP force plot was employed to visualize the contributions of individual features to the reconstruction loss, providing insights into which features had the most significant impact.

The SHAP summary plot, Figure 5.1 provides a comprehensive overview of feature importance and the nature of their impact on the model's predictions. The features are listed on the y-axis, ranked from most important (top) to least important (bottom). Feature 27 is the most important feature, followed by Features 25, 4, 16, and so on. The x-axis represents the SHAP values, which quantify the impact of each feature on the model's output. Positive SHAP values increase the model's prediction, while negative values decrease it. The color of each point represents the original value of the feature: red for high values and blue for low values. This indicates how the feature value correlates with the SHAP value and, consequently, with the model's prediction.

As seen in Figure 5.1, Feature 27 is the most important feature, having the highest impact on the model's predictions. This feature significantly influences the output, both positively and negatively, as shown by the wide spread of SHAP values. Feature 26 is also important, but its impact is much smaller compared to Feature 27. For Feature 27, high values (red) tend to increase the model's prediction, as indicated by the positive SHAP values. Conversely, low values (blue) decrease the

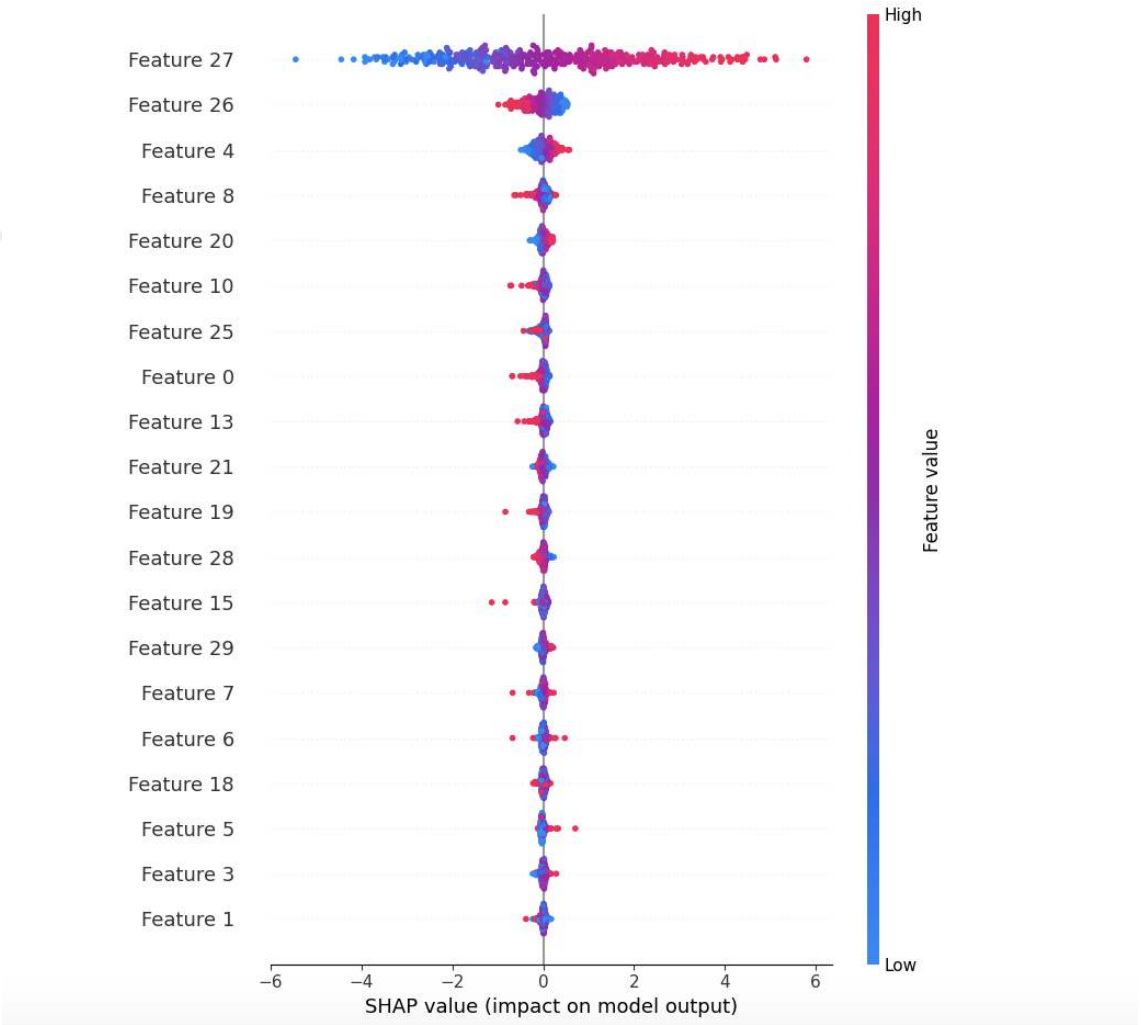


Figure 5.1: Feature importance analysis using SHAP values. SHAP summary plot displays the impact of individual features on the model's predictions, with colors representing feature values (red for high, blue for low).

prediction. For Feature 26, the pattern is less clear, but it still shows that variations in this feature have a noticeable impact on the model's output. Features lower in the ranking, such as Feature 4, 8, 20, and others, have minimal impact on the model's predictions. Their mean absolute SHAP values are close to zero, indicating they contribute little to the model's decisions.

LIME was also applied to provide local interpretability for the surrogate model's predictions. The LIME explainer was initialized with the training data, and explanations were generated for specific instances. These explanations highlighted the contributions of input features to the predicted reconstruction loss, aiding in the understanding of feature importance at a granular level.

Figure 5.2 provides an interpretability analysis of a specific prediction made by the model. It explains which features contributed to the model's output and how they affected the prediction. Features are listed with their contribution to the prediction, split into positive (orange) and negative (blue) contributions. Features pushing the prediction higher are on the right (positive), while those pushing it lower are on the left (negative). Each feature's actual value is shown alongside the range of values considered. For example, "$feature\_27 \leq ...$" indicates the specific condition of Feature 27 that contributed to the prediction. The size of the bars represents the magnitude of the contribution. Longer bars indicate more significant contributions to the model's prediction.

The LIME analysis effectively breaks down the prediction into contributions from individual features, highlighting the importance and direction of their impact. Feature 27 is identified as the most critical, with a substantial positive contribution, while other features like Feature 21 and Feature 26 also positively influence the prediction. Features like Feature 25 and Feature 19 have negative impacts, albeit to a lesser extent.

RandomForestRegressor was also applied as a surrogate model to ensure that the features found to be important were independent of the chosen model. The most important feature that affects the model decision is again Feature 27, as depicted in Figure 5.3.

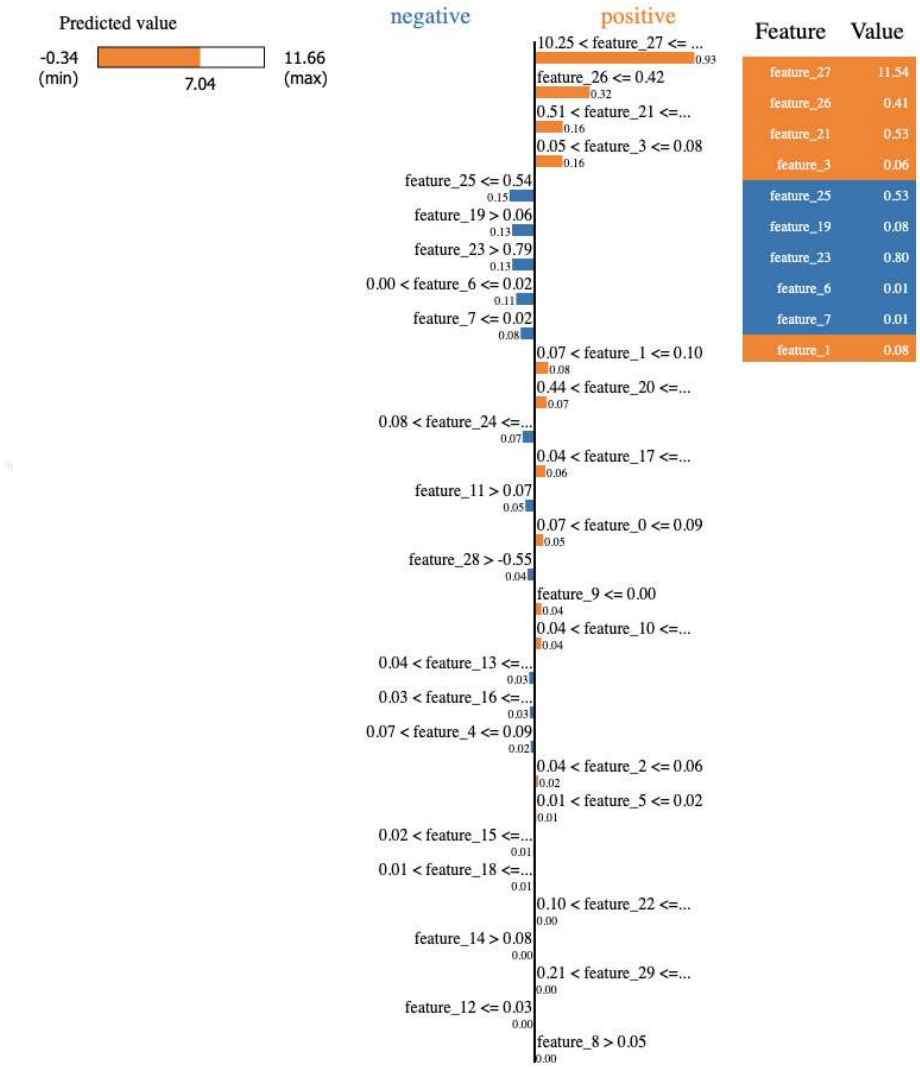


Figure 5.2: Feature importance analysis using LIME. Features are listed with their contributions to the prediction, where positive contributions (orange) increase the prediction, and negative contributions (blue) decrease it.

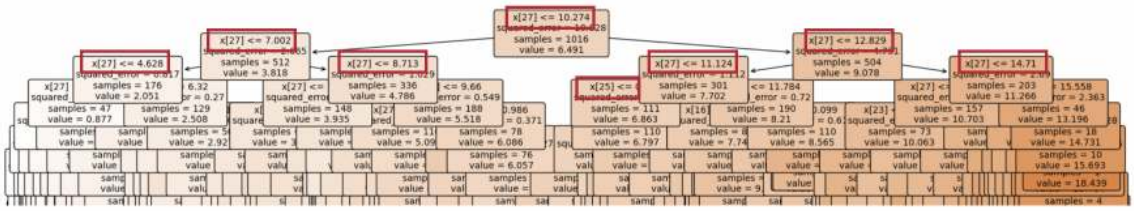

Figure 5.3: Decision tree of the Random Forest Regressor

Feature descriptions are given in Table 3.1. Feature 27 represents knowledge-based pair potential, while Feature 25 and 26 represent relative accessible surface area in monomer and complex form, respectively. Both SHAP and LIME analyses have shown that the most influential feature of the representation learning model is pair potential (Feature 27).

The application of SHAP and LIME provided comprehensive insights into the contributions of input features to the reconstruction loss in the representation learning model. The SHAP plot revealed the overall impact of each feature, while LIME offered detailed, instance-specific explanations. Together, these methods facilitated a deeper understanding of the model's behavior and the significance of individual input features.

## *5.2 Identification of Important Residues*

This section describes how to use GNNExplainer [Ying et al., 2019] to identify the significant residues (nodes) that have the most impact on the model prediction.

GNNExplainer [Ying et al., 2019] is a model-agnostic method specifically designed to provide interpretable explanations for predictions made by graph neural networks. It identifies a compact subgraph and a subset of node features that are most relevant for the model's prediction. The primary goal of GNNExplainer is to explain which parts of the graph structure and which node features contribute most significantly to the model's decision-making process.

GNNExplainer works by optimizing a mask over the graph's structure and node features to maximize the mutual information between the mask and the model's prediction. It initializes two masks on an input graph: one for the edges (edge_mask) and one for the node features (node_feat_mask). These masks are initially set to have values between 0 and 1, indicating the relevance of each edge and feature. The objective of GNNExplainer is to find masks that maximize the mutual information between the subgraph (induced by the masks) and the model's prediction. This is achieved by optimizing the masks so that the prediction on the masked graph is as close as possible to the original prediction. GNNExplainer uses gradient-based

optimization to learn the values of the edge_mask and node_feat_mask. The optimization process adjusts the masks to highlight the most important edges and node features. After optimization, the learned masks indicate the importance of each edge and node feature. A threshold can be applied to the masks to extract a subgraph that contains the most important nodes and edges for the prediction [Ying et al., 2019].

In this study, we first identified significant nodes that contribute to the model's prediction more than the other nodes using GNNExplainer. Then, we compare these residues (nodes) with the experimentally known hot spot residues.

Hot spot residues play a crucial role in protein-protein interactions by contributing significantly to the binding free energy of the interaction. These residues are typically found at the interface of interacting proteins and are essential for the stability and specificity of the protein complex. Understanding these hot spots is vital for understanding molecular mechanisms underlying protein interactions and therapeutic targeting [Tuncbag et al., 2009b].

The Structural Kinetic and Energetic Database of Mutant Protein Interactions (SKEMPI) dataset is a curated collection of data on the changes in protein-protein binding energy, kinetics, and thermodynamics resulting from mutations [Jankauskaitė et al., 2019]. The interface residues whose observed binding free energy changes are $\geq$ 2.0 kcal/mol upon mutation to another residue are considered as hot spots [Tuncbag et al., 2009b].

The SKEMPI data set contains 158 unique protein-protein complexes and 376 experimentally identified hot spot residues. Among these hot spot residues, 277 of them are identified as significant residues, corresponding to nearly 74%.

### *5.2.1 Case Study*

SHV-1 is a type of beta-lactamase enzyme that is responsible for bacterial resistance to beta-lactam antibiotics such as penicillin and cephalosporins. These enzymes hydrolyze the Beta-lactam ring present in these antibiotics, rendering them ineffective. BLIP is a protein that binds to Beta-lactamase enzymes and inhibits their activ-

ity. This inhibition can restore the effectiveness of Beta-lactam antibiotics against resistant bacterial strains [Reynolds et al., 2006]. The PDB entry 2G2U corresponds to the crystal structure of the SHV-1 Beta-lactamase complexed with the Beta-lactamase inhibitor protein (BLIP).

Here, the Beta-lactamase SHV-1 (Chain A) complexed with the Beta-lactamase inhibitor protein (Chain B) is selected as a case study to analyze which residues (nodes in the graph representation) are found to be important for the prediction model explained in Section 4.1.2, which is finetuned and used for a variety of downstream tasks including protein-protein interface validation (Chapter 4.1), biological vs. crystal classification (Chapter 4.2), and gene ontology term prediction (Chapter 4.3).

The structure of Beta-lactamase SHV-1 and BLIP complex (PDB: 2G2U_A_B) has a biologically relevant interface (labeled as positive in DeepInterface dataset [Balci et al., 2019]) classified as biological in the DC dataset [Duarte et al., 2012]. Some of the hot spot residues on this interface are also experimentally determined and reported in the SKEMPI v2 data set [Jankauskaitė et al., 2019].

Figure 5.4 represents the structure of the Beta-lactamase SHV-1 complexed with the Beta-lactamase inhibitor protein (BLIP). GNNExplainer has identified residues 99, 105, 167, 215, 235, and 244 from BLIP protein (Chain A) and residues 36, 41, 43, 50, 53, and 55 from Beta-lactamase SHV-1 (Chain B). These important residues are highlighted in purple and red, respectively.

Among the important residues identified by GNNExplainer, $\Delta\Delta G$ of two interface residues on the B chain (Residue 36 and Residue 53) upon mutation to alanine are reported as 2.76 kcal/mol and 2.30 kcal/mol, respectively, in SKEMPI v2 [Jankauskaitė et al., 2019]. Therefore, these residues are hot spot residues. Additionally, residues 99, 105, 235, and 244 from chain A and residues 41 and 50 from chain B are predicted as hot spots by a computational hot spot prediction server called HotPoint [Tuncbag et al., 2010].

The case study involving SHV-1 and BLIP demonstrates that the model can accurately identify and highlight the hot spot residues that are crucial for the inter-

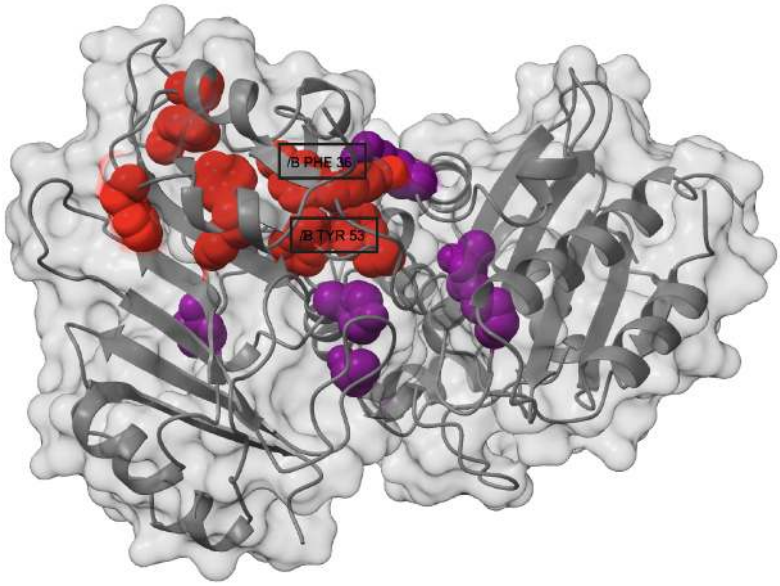


Figure 5.4: 3D structure of the Beta-lactamase SHV-1 complexed with the Beta-lactamase inhibitor protein (PDB ID: 2G2U). Important nodes identified by GNNExplainer are highlighted: key residues of the BLIP protein (Chain A) are shown in purple, while significant residues of Beta-lactamase SHV-1 (Chain B) are marked in red.

action. This not only validates the model's predictions but also provides a deeper understanding of the protein-protein interaction at a molecular level. Furthermore, the findings of this case study overlap with the findings of SHAP and LIME analysis, which identifies PP and relASA as the most important features. These two features have been proven to be effective in predicting hot spots in computational studies [Tuncbag et al., 2010].

## *5.3 Concluding Remarks*

In summary, the integration of explainable AI techniques, specifically LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), into the multiview-contrastive GCL autoencoder model has provided profound insights into the model's decision-making process by elucidating the contributions of various node features to the learned graph representations.

The case study presented in this chapter underscores the efficacy of this approach in identifying hot spot residues within protein-protein interactions. The robustness of the model, combined with the clarity provided by the explainable AI techniques, highlights the potential to uncover critical biological insights.

Through the application of SHAP and LIME, the chapter has demonstrated the importance of feature importance analysis in enhancing the transparency and interpretability of complex graph-based models.

The successful identification of important residues in the Beta-lactamase SHV-1 and BLIP protein complex showcases the practical utility of combining explainable AI with advanced graph-based learning models. This integration not only advances our understanding of protein interactions but also paves the way for developing more interpretable and trustworthy models in computational biology.

In conclusion, the chapter illustrates the transformative impact of incorporating explainable AI techniques into protein-protein interface studies. This integration has the potential to significantly enhance the interpretability and applicability of models, thereby contributing to a deeper understanding of biological processes.

# Chapter 6

# CONCLUSION

This dissertation presents a novel deep-learning approach combining a graph-based multiview contrastive autoencoder with a transformer architecture to address the limitations of existing methods in representing protein-protein interfaces. The proposed model leverages the strengths of graph-based representations and transformers to capture complex interaction networks, integrate multiple data modalities, and disentangle informative features from noisy data. This approach was trained on a large, unlabeled dataset of protein-protein interfaces, enabling the learning of robust and interpretable representations. The primary contributions of this work include the development of a comprehensive and interpretable representation learning method, its application to various tasks such as biological relevance prediction, biological vs. crystal classification, and Gene Ontology term prediction, and the integration of explainable AI techniques to enhance the interpretability of the results.

The results demonstrate the effectiveness of the proposed method in capturing the structural and functional characteristics of protein-protein interfaces. The learned representations outperform existing methods in terms of accuracy, scalability, and interpretability. Specifically, the method shows superior performance in validating biological relevance and classifying biological and crystallographic interfaces, highlighting its potential for practical applications in biomedical research and drug discovery.

In the evaluation of the learned representations, the t-SNE and UMAP analyses reveal clear clustering based on protein functions, indicating that the method successfully captures the underlying biological properties of protein-protein interfaces. The ablation studies further confirm the importance of each component of the proposed model, demonstrating that the combination of the graph-based contrastive

autoencoder and the transformer is essential for achieving high-quality representations.

Comprehensive performance comparison of the proposed model against two state-of-the-art methods, GNN-DOVE and DeepRank-GNN, is conducted to validate protein-protein interaction interfaces. To ensure a fair evaluation, all models were trained from scratch using three distinct training sets and subsequently evaluated on their respective test sets. The analysis demonstrated that the proposed model consistently outperformed the other models across all three datasets, achieving an impressive accuracy of 0.91 on the DeepInterface dataset. These findings underscore the robustness and efficacy of the proposed model in accurately validating PPIs, highlighting its potential as a powerful tool in computational protein studies.

Further analysis included a ranking evaluation, where the proposed model achieved a 63% success rate in the top 1% and a 90% success rate in the top 10%, outperforming other models such as GNN-DOVE and DeepRank-GNN. This indicates the potential of the proposed model to improve the predictive capabilities of template-based protein-protein interaction (PPI) predictors.

The success of the proposed method in accurately validating PPIs can be attributed to its innovative use of graph-based and transformer-based techniques. The graph-based approach allows for the detailed representation of protein interfaces, capturing the intricate network of interactions between residues. The transformer architecture further enhances this representation by modeling long-range dependencies and contextual information within protein sequences.

The evaluation of the proposed method on benchmark datasets for biological vs. crystal classification task demonstrated its superior performance compared to existing state-of-the-art methods. The method achieved approximately 88% accuracy, 88% precision, and 85% F1 score on the DC dataset, outperforming other methods such as DeepRank-GNN, PRODIGY-CRYSTAL, EPPIC 3, PISA, and QSAlign. This superior performance highlights the effectiveness of the proposed approach in accurately capturing the structural and functional characteristics of protein-protein interfaces.

The development of a user-friendly web server for predicting class scores for protein-protein interfaces further enhances the accessibility and usability of the proposed method. This web server allows users to upload PDB files and receive predictions on whether the interfaces are biologically relevant or crystallographic. The rapid and accurate predictions provided by the web server make it a valuable resource for researchers.

Our model's performance is thoroughly compared to GearNet pretraining models in order to predict GO terms for proteins based on a variety of functional features. Our approach outperformed compared models with accuracy rates of 54%, 70%, 52%, and 63% on predictions of molecular function, biological process, cellular components, and fold categorization, respectively.

The application of explainable AI techniques, such as SHAP, LIME, and GNNExplainer, provides valuable insights into the key features contributing to the model's predictions. These techniques enhance the interpretability of the results, allowing researchers to understand the molecular basis of protein-protein interactions and identify potential therapeutic targets.

In conclusion, the proposed deep-learning approach represents a significant advancement in the field of protein-protein interface representation, offering improved accuracy, scalability, and interpretability. These advancements pave the way for more effective biomedical research and therapeutic development, ultimately contributing to our ability to combat diseases and understand the intricate workings of biological systems. The integration of computational methods with experimental approaches will continue to drive further breakthroughs in the understanding and manipulation of protein interactions, enhancing our capabilities in disease diagnosis, drug discovery, and the elucidation of cellular mechanisms.

This work not only advances the field of computational biology but also provides a framework for future research in protein-protein interaction interfaces. The novel methodologies and findings presented here will serve as a foundation for developing more sophisticated models and tools, facilitating deeper insights into the molecular basis of cellular processes and the development of targeted therapeutic interventions.

# BIBLIOGRAPHY

[Abali et al., 2022] Abali, Z., Ovek, D., Senyuz, S., Keskin, O., and Gursoy, A. (2022). Protein–protein-binding interfaces. *Protein Interactions: The Molecular Basis of Interactomics*, pages 15–37.

[Abalı, 2021] Abalı, Z. (2021). A data-centric approach for investigation of protein-protein interfaces in protein data bank.

[Agrawal et al., 2022] Agrawal, S., Sisodia, D. S., and Nagwani, N. K. (2022). Function characterization of unknown protein sequences using one hot encoding and convolutional neural network based model. In *International Conference on Machine Intelligence and Signal Processing*, pages 267–277. Springer.

[Akbal-Delibas et al., 2016] Akbal-Delibas, B., Farhoodi, R., Pomplun, M., and Haspel, N. (2016). Accurate refinement of docked protein complexes using evolutionary information and deep learning. *Journal of Bioinformatics and Computational Biology*, 14(03):1642002.

[Alakus and Turkoglu, 2021] Alakus, T. B. and Turkoglu, I. (2021). A novel protein mapping method for predicting the protein interactions in covid-19 disease by deep learning. *Interdisciplinary Sciences: Computational Life Sciences*, 13:44–60.

[Alam et al., 2017] Alam, N., Goldstein, O., Xia, B., Porter, K. A., Kozakov, D., and Schueler-Furman, O. (2017). High-resolution global peptide-protein docking using fragments-based piper-flexpepdock. *PLoS computational biology*, 13(12):e1005905.

[Ali and Imperiali, 2005] Ali, M. H. and Imperiali, B. (2005). Protein oligomerization: how and why. *Bioorganic & medicinal chemistry*, 13(17):5013–5020.

[Andrusier et al., 2008] Andrusier, N., Mashiach, E., Nussinov, R., and Wolfson, H. J. (2008). Principles of flexible protein-protein docking. *Proteins*, 73(2):271–89.

[Anishchenko et al., 2015] Anishchenko, I., Kundrotas, P. J., Tuzikov, A. V., and Vakser, I. A. (2015). Structural templates for comparative protein docking. *Proteins: Structure, Function, and Bioinformatics*, 83(9):1563–1570.

[Asim et al., 2022] Asim, M. N., Ibrahim, M. A., Malik, M. I., Dengel, A., and Ahmed, S. (2022). Adh-ppi: An attention-based deep hybrid model for protein-protein interaction prediction. *Iscience*, 25(10).

[Aybey and Gümüş, 2023] Aybey, E. and Gümüş, Ö. (2023). Sensdeep: an ensemble deep learning method for protein–protein interaction sites prediction. *Interdisciplinary Sciences: Computational Life Sciences*, 15(1):55–87.

[Aytuna et al., 2005] Aytuna, A. S., Gursoy, A., and Keskin, O. (2005). Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21(12):2850–2855.

[Bahadur et al., 2004] Bahadur, R. P., Chakrabarti, P., Rodier, F., and Janin, J. (2004). A dissection of specific and non-specific protein–protein interfaces. *Journal of molecular biology*, 336(4):943–955.

[Balci et al., 2019] Balci, A., Gumeli, C., Hakouz, A., Yuret, D., Keskin, O., and Gursoy, A. (2019). Deepinterface: Protein-protein interface validation using 3d convolutional neural networks. *bioRxiv*, page 617506.

[Bartel and Fields, 1997] Bartel, P. L. and Fields, S. (1997). *The yeast two-hybrid system*. Oxford University Press, USA.

[Baskaran et al., 2014] Baskaran, K., Duarte, J. M., Biyani, N., Bliven, S., and Capitani, G. (2014). A pdb-wide, evolution-based assessment of protein-protein interfaces. *BMC Structural Biology*, 14:1–11.

[Bernauer et al., 2008] Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J., and Poupon, A. (2008). Dimovo: a voronoi tessellation-based method for discriminating crystallographic and biological protein–protein interactions. *Bioinformatics*, 24(5):652–658.

[Bliven et al., 2018] Bliven, S., Lafita, A., Parker, A., Capitani, G., and Duarte, J. M. (2018). Automated evaluation of quaternary structures from protein crystals. *PLoS computational biology*, 14(4):e1006104.

[Bock and Gough, 2001] Bock, J. R. and Gough, D. A. (2001). Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–60.

[Capel et al., 2022] Capel, H., Feenstra, K. A., and Abeln, S. (2022). Multi-task learning to leverage partially annotated data for ppi interface prediction. *Scientific Reports*, 12(1).

[Capitani et al., 2016] Capitani, G., Duarte, J. M., Baskaran, K., Bliven, S., and Somody, J. C. (2016). Understanding the fabric of protein crystals: computational classification of biological interfaces and crystal contacts. *Bioinformatics*, 32(4):481–489.

[Chakraborty et al., 2014] Chakraborty, C., Priya, D., Chen, L., Zhu, H., et al. (2014). Evaluating protein-protein interaction (ppi) networks for diseases pathway, target discovery, and drug-design using 'in silico pharmacology'. *Current Protein and Peptide Science*, 15(6):561–571.

[Chen et al., 2019] Chen, M., Ju, C. J., Zhou, G., Chen, X., Zhang, T., Chang, K. W., Zaniolo, C., and Wang, W. (2019). Multifaceted protein-protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, 35(14):i305–i314.

[Chen et al., 2022] Chen, W., Wang, S., Song, T., Li, X., Han, P., and Gao, C. (2022). Dcse: Double-channel-siamese-ensemble model for protein protein interaction prediction. *BMC genomics*, 23(1):555.

[Chen and Jeong, 2009] Chen, X. W. and Jeong, J. C. (2009). Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, 25(5):585–91.

[Chen and Liu, 2005] Chen, X. W. and Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–400.

[Chen et al., 2023] Chen, Z., Liu, N., Huang, Y., Min, X., Zeng, X., Ge, S., Zhang, J., and Xia, N. (2023). Pointde: Protein docking evaluation using 3d point cloud neural network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

[Cui et al., 2021] Cui, F., Zhang, Z., and Zou, Q. (2021). Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. *Briefings in Functional Genomics*, 20(1):61–73.

[Cukuroglu et al., 2014] Cukuroglu, E., Gursoy, A., Nussinov, R., and Keskin, O. (2014). Non-redundant unique interface structures as templates for modeling protein interactions. *PloS one*, 9(1):e86738.

[Czibula et al., 2021] Czibula, G., Albu, A.-I., Bocicor, M. I., and Chira, C. (2021). Autoppi: An ensemble of deep autoencoders for protein–protein interaction prediction. *Entropy*, 23(6):643.

[Da Silva et al., 2015] Da Silva, F., Desaphy, J., Bret, G., and Rognan, D. (2015). Ichempic: a random forest classifier of biological and crystallographic protein–protein interfaces. *Journal of chemical information and modeling*, 55(9):2005–2014.

[de Vries et al., 2010] de Vries, S. J., van Dijk, M., and Bonvin, A. M. (2010). The haddock web server for data-driven biomolecular docking. *Nat Protoc*, 5(5):883–97.

[Degiacomi, 2019] Degiacomi, M. T. (2019). Coupling molecular dynamics and deep learning to mine protein conformational space. *Structure*, 27(6):1034–1040.

[Deng et al., 2019] Deng, L., Sui, Y., and Zhang, J. (2019). Xgbprh: Prediction of binding hot spots at protein(-)rna interfaces utilizing extreme gradient boosting. *Genes (Basel)*, 10(3).

[Derry and Altman, 2023] Derry, A. and Altman, R. B. (2023). Collapse: A representation learning framework for identification and characterization of protein structural sites. *Protein Science*, 32(2):e4541.

[Dey et al., 2022] Dey, S., Prilusky, J., and Levy, E. D. (2022). Qsalignweb: A server to predict and analyze protein quaternary structure. *Frontiers in molecular biosciences*, 8:787510.

[Dey et al., 2018] Dey, S., Ritchie, D. W., and Levy, E. D. (2018). Pdb-wide identification of biological assemblies from conserved quaternary structure geometry. *Nature methods*, 15(1):67–72.

[Ding and Arnold, 2006] Ding, J. and Arnold, E. (2006). Naccess.

[Ding and Kihara, 2018] Ding, Z. and Kihara, D. (2018). Computational methods for predicting protein-protein interactions using various protein features. *Curr Protoc Protein Sci*, 93(1):e62.

[Dong et al., 2021] Dong, T. N., Brogden, G., Gerold, G., and Khosla, M. (2021). A multitask transfer learning framework for the prediction of virus-human protein–protein interactions. *BMC Bioinformatics*, 22(1).

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[Duarte et al., 2012] Duarte, J. M., Srebniak, A., Schärer, M. A., and Capitani, G. (2012). Protein interface classification by evolutionary analysis. *BMC bioinformatics*, 13:1–16.

[Elez et al., 2020] Elez, K., Bonvin, A. M., and Vangone, A. (2020). Biological vs. crystallographic protein interfaces: An overview of computational approaches for their classification. *Crystals*, 10(2):114.

[Esquivel-Rodríguez and Kihara, 2012] Esquivel-Rodríguez, J. and Kihara, D. (2012). Fitting multimeric protein complexes into electron microscopy maps using 3d zernike descriptors. *The journal of physical chemistry B*, 116(23):6854–6861.

[Esquivel-Rodríguez et al., 2012] Esquivel-Rodríguez, J., Yang, Y. D., and Kihara, D. (2012). Multi-lzerd: multiple protein docking for asymmetric complexes. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1818–1833.

[Fink et al., 2011] Fink, F., Hochrein, J., Wolowski, V., Merkl, R., and Gronwald, W. (2011). Procos: Computational analysis of protein–protein complexes. *Journal of computational chemistry*, 32(12):2575–2586.

[Fischer et al., 1995] Fischer, D., Lin, S. L., Wolfson, H. L., and Nussinov, R. (1995). A geometry-based suite of moleculardocking processes. *Journal of Molecular Biology*, 248(2):459–477.

[Fox et al., 2014] Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2014). Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309.

[Fukasawa and Tomii, 2019] Fukasawa, Y. and Tomii, K. (2019). Accurate classification of biological and non-biological interfaces in protein crystal structures using subtle covariation signals. *Scientific Reports*, 9(1):12603.

[Gainza et al., 2020a] Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M., and Correia, B. (2020a). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192.

[Gainza et al., 2020b] Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. M., and Correia, B. E. (2020b). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192.

[Gao et al., 2023] Gao, H., Chen, C., Li, S., Wang, C., Zhou, W., and Yu, B. (2023). Prediction of protein-protein interactions based on ensemble residual convolutional neural network. *Computers in Biology and Medicine*, 152:106471.

[Gavin et al., 2002] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.

[Georgiev and Maksimenko, 2014] Georgiev, P. and Maksimenko, O. (2014). Mechanisms and proteins involved in long-distance interactions. *Frontiers in genetics*, 5:71865.

[Ghosh and Mitra, 2024] Ghosh, S. and Mitra, P. (2024). Matpip: A deep-learning architecture with explainable ai for sequence-driven, feature mixed protein-protein interaction prediction. *Computer Methods and Programs in Biomedicine*, 244:107955.

[Gong et al., 2005] Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J., Bolser, D. M., Oh, D., Kim, D.-S., and Bhak, J. (2005). A protein domain interaction interface database: Interpare. *Bmc Bioinformatics*, 6:1–8.

[Graef et al., 2023] Graef, J., Ehrt, C., and Rarey, M. (2023). Binding site detection remastered: enabling fast, robust, and reliable binding site detection and descriptor calculation with dogsite3. *Journal of Chemical Information and Modeling*, 63(10):3128–3137.

[Gray et al., 2003] Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. (2003). Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, 331(1):281–299.

[Guo et al., 2008] Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res*, 36(9):3025–30.

[Hashemifar et al., 2018] Hashemifar, S., Neyshabur, B., Khan, A. A., and Xu, J. (2018). Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics*, 34(17):i802–i810.

[Hasibi and Michoel, 2021] Hasibi, R. and Michoel, T. (2021). A graph feature auto-encoder for the prediction of unobserved node features on biological networks. *BMC bioinformatics*, 22:1–17.

[Hu et al., 2018] Hu, J., Liu, H.-F., Sun, J., Wang, J., and Liu, R. (2018). Integrating co-evolutionary signals and other properties of residue pairs to distinguish biological interfaces from crystal contacts. *Protein Science*, 27(9):1723–1735.

[Hu et al., 2022a] Hu, X., Feng, C., Ling, T., and Chen, M. (2022a). Deep learning frameworks for protein-protein interaction prediction. *Comput Struct Biotechnol J*, 20:3223–3233.

[Hu et al., 2021] Hu, X., Feng, C., Zhou, Y., Harrison, A., and Chen, M. (2021). Deeptrio: a ternary prediction system for protein-protein interaction using mask multiple parallel convolutional neural networks. *Bioinformatics*, 38(3):694–702.

[Hu et al., 2022b] Hu, X., Feng, C., Zhou, Y., Harrison, A., and Chen, M. (2022b). Deeptrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks. *Bioinformatics*, 38(3):694–702.

[Huang and Zou, 2008] Huang, S.-Y. and Zou, X. (2008). An iterative knowledge-based scoring function for protein–protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 72(2):557–579.

[Huang et al., 2004] Huang, T. W., Tien, A. C., Huang, W. S., Lee, Y. C., Peng, C. L., Tseng, H. H., Kao, C. Y., and Huang, C. Y. (2004). Point: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, 20(17):3273–6.

[Hwang et al., 2010] Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein–protein docking benchmark version 4.0. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3111–3114.

[Janin et al., 2003] Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I., and Wodak, S. J. (2003). Capri: a critical assessment of predicted interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1):2–9.

[Jankauskaitė et al., 2019] Jankauskaitė, J., Jiménez-García, B., Dapkūnas, J., Fernández-Recio, J., and Moal, I. H. (2019). Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469.

[Jha et al., 2023] Jha, K., Karmakar, S., and Saha, S. (2023). Graph-bert and language model-based framework for protein–protein interaction identification. *Scientific Reports*, 13(1):5663.

[Jimenez et al., 2017] Jimenez, J., Doerr, S., Martinez-Rosell, G., Rose, A. S., and De Fabritiis, G. (2017). Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042.

[Jiménez et al., 2017] Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., and De Fabritiis, G. (2017). Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042.

[Jiménez-García et al., 2019] Jiménez-García, B., Elez, K., Koukos, P. I., Bonvin, A. M., and Vangone, A. (2019). Prodigy-crystal: a web-tool for classification of biological interfaces in protein complexes. *Bioinformatics*, 35(22):4821–4823.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

[Katchalski-Katzir et al., 1992] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992). Molecular surface recognition:

determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences*, 89(6):2195–2199.

[Keskin et al., 1998] Keskin, O., Bahar, I., Jernigan, R., Badretdinov, A., and Ptitsyn, O. (1998). Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Science*, 7(12):2578–2586.

[Keskin et al., 2008a] Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2008a). Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem Rev*, 108(4):1225–44.

[Keskin et al., 2008b] Keskin, O., Tuncbag, N., and Gursoy, A. (2008b). Characterization and prediction of protein interfaces to infer protein-protein interaction networks. *Curr Pharm Biotechnol*, 9(2):67–76.

[Keskin et al., 2016a] Keskin, O., Tuncbag, N., and Gursoy, A. (2016a). Predicting protein–protein interactions from the molecular to the proteome level. *Chemical reviews*, 116(8):4884–4909.

[Keskin et al., 2016b] Keskin, O., Tuncbag, N., and Gursoy, A. (2016b). Predicting protein-protein interactions from the molecular to the proteome level. *Chem Rev*, 116(8):4884–909.

[Khan et al., 2022] Khan, S. H., Tayara, H., and Chong, K. T. (2022). Prob-site: Protein binding site prediction using local features. *Cells*, 11(13).

[Kingsley et al., 2016] Kingsley, L. J., Esquivel-Rodríguez, J., Yang, Y., Kihara, D., and Lill, M. A. (2016). Ranking protein–protein docking results using steered molecular dynamics and potential of mean force calculations. *Journal of computational chemistry*, 37(20):1861–1865.

[Kozakov et al., 2017] Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., and Vajda, S. (2017). The cluspro web server for protein–protein docking. *Nature protocols*, 12(2):255–278.

[Krapp et al., 2023] Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F., and Dal Peraro, M. (2023). Pesto: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature communications*, 14(1):2175.

[Krissinel and Henrick, 2007] Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3):774–797.

[Kuntz and Privé, 2024] Kuntz, D. and Privé, G. (2024). X-ray diffraction data for: Nras q61k crystal. pdb code 8vm2.

[Kurcinski et al., 2020] Kurcinski, M., Badaczewska-Dawid, A., Kolinski, M., Kolinski, A., and Kmiecik, S. (2020). Flexible docking of peptides to proteins using cabs-dock. *Protein Science*, 29(1):211–222.

[Kurcinski et al., 2015] Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A., and Kmiecik, S. (2015). Cabs-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic acids research*, 43(W1):W419–W424.

[Kuzmanic et al., 2020] Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J., and Gervasio, F. L. (2020). Investigating cryptic binding sites by molecular dynamics simulations. *Acc Chem Res*, 53(3):654–661.

[Lee et al., 2008] Lee, S. A., Chan, C. H., Tsai, C. H., Lai, J. M., Wang, F. S., Kao, C. Y., and Huang, C. Y. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*, 9 Suppl 12(Suppl 12):S11.

[Lensink et al., 2018] Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C., and Wodak, S. J. (2018). The challenge of modeling protein assemblies: the casp12-capri experiment. *Proteins: Structure, Function, and Bioinformatics*, 86:257–273.

[Lensink et al., 2016] Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S.-Y., Schneidman-Duhovny, D., Sali, A., Segura, J., Fernandez-Fuentes, N., Viswanath, S., Elber, R., et al. (2016). Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: a casp-capri experiment. *Proteins: Structure, Function, and Bioinformatics*, 84:323–348.

[Lensink and Wodak, 2014] Lensink, M. F. and Wodak, S. J. (2014). Score_set: a capri benchmark for scoring protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 82(11):3163–3169.

[Li et al., 2012] Li, B. Q., Feng, K. Y., Chen, L., Huang, T., and Cai, Y. D. (2012). Prediction of protein-protein interaction sites by random forest algorithm with mrmr and ifs. *PLoS One*, 7(8):e43927.

[Li et al., 2021a] Li, Y., Golding, G. B., and Ilie, L. (2021a). Delphi: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics*, 37(7):896–904.

[Li et al., 2023] Li, Y., Lu, S., Nan, X., Zhang, S., and Zhou, Q. (2023). Mspbrsp: Multi-scale protein binding residues prediction using language model. *bioRxiv*, pages 2023–02.

[Li et al., 2021b] Li, Y., Wang, Z., Li, L.-P., You, Z.-H., Huang, W.-Z., Zhan, X.-K., and Wang, Y.-B. (2021b). Robust and accurate prediction of protein–protein interactions by exploiting evolutionary information. *Scientific Reports*, 11(1):16910.

[Liu et al., 2013] Liu, B., Wang, X., Zou, Q., Dong, Q., and Chen, Q. (2013). Protein remote homology detection by combining chou's pseudo amino acid composition and profile-based protein representation. *Molecular Informatics*, 32(9-10):775–782.

[Liu et al., 2020a] Liu, L., Zhu, X., Ma, Y., Piao, H., Yang, Y., Hao, X., Fu, Y., Wang, L., and Peng, J. (2020a). Combining sequence and network information to enhance protein–protein interaction prediction. *BMC bioinformatics*, 21(16):1–13.

[Liu et al., 2020b] Liu, L., Zhu, X., Ma, Y., Piao, H., Yang, Y., Hao, X., Fu, Y., Wang, L., and Peng, J. (2020b). Combining sequence and network information to enhance protein-protein interaction prediction. *BMC Bioinformatics*, 21(Suppl 16):537.

[Liu et al., 2014] Liu, Q., Li, Z., and Li, J. (2014). Use b-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC bioinformatics*, 15:1–11.

[Liu et al., 2008] Liu, S., Gao, Y., and Vakser, I. A. (2008). Dockground protein–protein docking decoy set. *Bioinformatics*, 24(22):2634–2635.

[Liu et al., 2018] Liu, Y., Ye, Q., Wang, L., and Peng, J. (2018). Learning structural motif representations for efficient protein structure search. *Bioinformatics*, 34(17):i773–i780.

[Lu et al., 2003] Lu, H., Lu, L., and Skolnick, J. (2003). Development of unified statistical potentials describing protein-protein interactions. *Biophysical journal*, 84(3):1895–1901.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

[Luo et al., 2014] Luo, J., Guo, Y., Fu, Y., Wang, Y., Li, W., and Li, M. (2014). Effective discrimination between biologically relevant contacts and crystal packing contacts using new determinants. *Proteins: Structure, Function, and Bioinformatics*, 82(11):3090–3100.

[Mabonga and Kappo, 2019] Mabonga, L. and Kappo, A. P. (2019). Protein-protein interaction modulators: advances, successes and remaining challenges. *Biophys Rev*, 11(4):559–581.

[Mahapatra and Sahu, 2021] Mahapatra, S. and Sahu, S. S. (2021). Improved prediction of protein-protein interaction using a hybrid of functional-link siamese neural network and gradient boosting machines. *Brief Bioinform*, 22(6).

[Matthews et al., 2001] Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, 11(12):2120–6.

[Mayer, 1999] Mayer, B. J. (1999). Protein-protein interactions in signaling cascades. *Molecular biotechnology*, 13:201–213.

[Merwe and Davis, 2003] Merwe, P. A. v. d. and Davis, S. J. (2003). Molecular interactions mediating t cell antigen recognition. *Annual review of immunology*, 21(1):659–684.

[Mitra and Pal, 2011] Mitra, P. and Pal, D. (2011). Combining bayes classification and point group symmetry under boolean framework for enhanced protein quaternary structure inference. *Structure*, 19(3):304–312.

[Mitton et al., 2021] Mitton, J., Senn, H. M., Wynne, K., and Murray-Smith, R. (2021). A graph vae and graph transformer approach to generating molecular graphs. *arXiv preprint arXiv:2104.04345*.

[Moal and Bates, 2010] Moal, I. H. and Bates, P. A. (2010). Swarmdock and the use of normal modes in protein-protein docking. *International journal of molecular sciences*, 11(10):3623–3648.

[Moal et al., 2013] Moal, I. H., Torchala, M., Bates, P. A., and Fernández-Recio, J. (2013). The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC bioinformatics*, 14:1–15.

[Moreira et al., 2017] Moreira, I. S., Koukos, P. I., Melo, R., Almeida, J. G., Preto, A. J., Schaarschmidt, J., Trellet, M., Gumus, Z. H., Costa, J., and Bonvin, A. (2017). Spoton: High accuracy identification of protein-protein interface hotspots. *Sci Rep*, 7(1):8007.

[Murzin et al., 1995] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540.

[Nadaradjane et al., 2018] Nadaradjane, A. A., Guerois, R., and Andreani, J. (2018). Protein-protein docking using evolutionary information. *Protein Complex Assembly: Methods and Protocols*, pages 429–447.

[Nambiar et al., 2023] Nambiar, A., Liu, S., Heflin, M., Forsyth, J. M., Maslov, S., Hopkins, M., and Ritz, A. (2023). Transformer neural networks for protein family and interaction prediction tasks. *Journal of Computational Biology*, 30(1):95–111.

[Oliwa and Shen, 2015] Oliwa, T. and Shen, Y. (2015). cnma: a framework of encounter complex-based normal mode analysis to model conformational changes in protein interactions. *Bioinformatics*, 31(12):i151–i160.

[Padhorny et al., 2016] Padhorny, D., Kazennov, A., Zerbe, B. S., Porter, K. A., Xia, B., Mottarella, S. E., Kholodov, Y., Ritchie, D. W., Vajda, S., and Kozakov, D. (2016). Protein–protein docking by fast generalized fourier transforms on 5d rotational manifolds. *Proceedings of the National Academy of Sciences*, 113(30):E4286–E4293.

[Pan et al., 2022] Pan, J., You, Z.-H., Li, L.-P., Huang, W.-Z., Guo, J.-X., Yu, C.-Q., Wang, L.-P., and Zhao, Z.-Y. (2022). Dwppi: A deep learning approach for predicting protein–protein interactions in plants based on multi-source information with a large-scale biological network. *Frontiers in Bioengineering and Biotechnology*, 10.

[Peterson et al., 2017] Peterson, L. X., Roy, A., Christoffer, C., Terashi, G., and Kihara, D. (2017). Modeling disordered protein interactions from biophysical principles. *PLoS computational biology*, 13(4):e1005485.

[Peterson et al., 2018a] Peterson, L. X., Shin, W.-H., Kim, H., and Kihara, D. (2018a). Improved performance in capri round 37 using lzerd docking and template-based modeling with combined scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 86:311–320.

[Peterson et al., 2018b] Peterson, L. X., Togawa, Y., Esquivel-Rodriguez, J., Terashi, G., Christoffer, C., Roy, A., Shin, W.-H., and Kihara, D. (2018b). Modeling the assembly order of multimeric heteroprotein complexes. *PLoS computational biology*, 14(1):e1005937.

[Phizicky and Fields, 1995] Phizicky, E. M. and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, 59(1):94–123.

[Pierce and Weng, 2007] Pierce, B. and Weng, Z. (2007). Zrank: reranking protein docking predictions with an optimized energy function. *Proteins: Structure, Function, and Bioinformatics*, 67(4):1078–1086.

[Pierce et al., 2011] Pierce, B. G., Hourai, Y., and Weng, Z. (2011). Accelerating protein docking in zdock using an advanced 3d convolution library. *PloS one*, 6(9):e24657.

[Poluri et al., 2021] Poluri, K. M., Gulati, K., and Sarkar, S. (2021). *Protein-protein interactions*. Springer.

[Ponstingl et al., 2003] Ponstingl, H., Kabir, T., and Thornton, J. M. (2003). Automatic inference of protein quaternary structure from crystals. *Journal of Applied Crystallography*, 36(5):1116–1122.

[Quadrini et al., 2022] Quadrini, M., Daberdaku, S., and Ferrari, C. (2022). Hierarchical representation for ppi sites prediction. *BMC bioinformatics*, 23(1):96.

[Rakers et al., 2015] Rakers, C., Bermudez, M., Keller, B. G., Mortier, J., and Wolber, G. (2015). Computational close up on protein-protein interactions: how to unravel the invisible using molecular dynamics simulations? *Wiley Interdisciplinary Reviews-Computational Molecular Science*, 5(5):345–359.

[Réau et al., 2023] Réau, M., Renaud, N., Xue, L. C., and Bonvin, A. M. (2023). Deeprank-gnn: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, 39(1):btac759.

[Renaud et al., 2021] Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Réau, M. F., Bonvin, A. M., and Xue, L. C. (2021). Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nature communications*, 12(1):7068.

[Reynolds et al., 2006] Reynolds, K. A., Thomson, J. M., Corbett, K. D., Bethel, C. R., Berger, J. M., Kirsch, J. F., Bonomo, R. A., and Handel, T. M. (2006). Structural and computational characterization of the shv-1 $\beta$-lactamase-$\beta$-lactamase inhibitor protein interface. *Journal of biological chemistry*, 281(36):26745–26753.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Rigaut et al., 1999] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, 17(10):1030–1032.

[Ritchie and Grudinin, 2016] Ritchie, D. W. and Grudinin, S. (2016). Spherical polar fourier assembly of protein complexes with arbitrary point group symmetry. *Journal of Applied Crystallography*, 49(1):158–167.

[Rodgers-Melnick et al., 2013] Rodgers-Melnick, E., Culp, M., and DiFazio, S. P. (2013). Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ents. *BMC Genomics*, 14:608.

[Schneidman-Duhovny et al., 2005] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). Geometry-based flexible and symmetric protein docking. *Proteins: Structure, Function, and Bioinformatics*, 60(2):224–231.

[Schweke et al., 2023] Schweke, H., Xu, Q., Tauriello, G., Pantolini, L., Schwede, T., Cazals, F., Lhéritier, A., Fernandez-Recio, J., Rodríguez-Lumbreras, L. A., Schueler-Furman, O., et al. (2023). Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. *Proteomics*, 23(17):2200323.

[Shen et al., 2007] Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*, 104(11):4337–41.

[Soleymani et al., 2023] Soleymani, F., Paquet, E., Viktor, H. L., Michalowski, W., and Spinello, D. (2023). Protinteract: A deep learning framework for predicting protein–protein interactions. *Computational and Structural Biotechnology Journal*, 21:1324–1348.

[Song et al., 2022] Song, B., Luo, X., Luo, X., Liu, Y., Niu, Z., and Zeng, X. (2022). Learning spatial structures of proteins improves protein-protein interaction prediction. *Brief Bioinform*, 23(2).

[Sousa et al., 2024] Sousa, R. T., Silva, S., and Pesquita, C. (2024). Explaining protein–protein interactions with knowledge graph-based semantic similarity. *Computers in Biology and Medicine*, 170:108076.

[Stallcup et al., 2003] Stallcup, M. R., Kim, J. H., Teyssier, C., Lee, Y.-H., Ma, H., and Chen, D. (2003). The roles of protein–protein interactions and protein methylation in transcriptional activation by nuclear receptors and their coactivators. *The Journal of steroid biochemistry and molecular biology*, 85(2-5):139–145.

[Stebliankin et al., 2023] Stebliankin, V., Shirali, A., Baral, P., Shi, J., Chapagain, P., Mathee, K., and Narasimhan, G. (2023). Evaluating protein binding interfaces with transformer networks. *Nature Machine Intelligence*, pages 1–12.

[Sussman et al., 1998] Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., and Abola, E. E. (1998). Protein data bank (pdb): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography*, 54(6):1078–1084.

[Tamir and Cambier, 1998] Tamir, I. and Cambier, J. C. (1998). Antigen receptor signaling: integration of protein tyrosine kinase functions. *Oncogene*, 17(11):1353–1364.

[Tan and Zhang, 2023] Tan, J. and Zhang, Y. (2023). Explainablefold: Understanding alphafold prediction with explainable ai. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2166–2176.

[Tong et al., 2002] Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., et al. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–324.

[Tovchigrechko and Vakser, 2006] Tovchigrechko, A. and Vakser, I. A. (2006). Gramm-x public web server for protein–protein docking. *Nucleic acids research*, 34(suppl_2):W310–W314.

[Tsaban et al., 2022] Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A., and Schueler-Furman, O. (2022). Harnessing protein folding neural networks for peptide-protein docking. *Nat Commun*, 13(1):176.

[Tsuchiya et al., 2006] Tsuchiya, Y., Kinoshita, K., Ito, N., and Nakamura, H. (2006). Prebi: prediction of biological interfaces of proteins in crystals. *Nucleic acids research*, 34(suppl_2):W20–W24.

[Tsuchiya et al., 2008] Tsuchiya, Y., Nakamura, H., and Kinoshita, K. (2008). Discrimination between biological interfaces and crystal-packing contacts. *Advances and Applications in Bioinformatics and Chemistry*, pages 99–113.

[Tubiana et al., 2022] Tubiana, J., Schneidman-Duhovny, D., and Wolfson, H. J. (2022). Scannet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat Methods*, 19(6):730–739.

[Tuncbag et al., 2009a] Tuncbag, N., Gursoy, A., and Keskin, O. (2009a). Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25(12):1513–20.

[Tuncbag et al., 2009b] Tuncbag, N., Gursoy, A., and Keskin, O. (2009b). Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25(12):1513–1520.

[Tuncbag et al., 2011a] Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011a). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism. *Nat Protoc*, 6(9):1341–54.

[Tuncbag et al., 2011b] Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011b). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism. *Nature protocols*, 6(9):1341–1354.

[Tuncbag et al., 2010] Tuncbag, N., Keskin, O., and Gursoy, A. (2010). Hotpoint: hot spot prediction server for protein interfaces. *Nucleic acids research*, 38(suppl_2):W402–W406.

[Valdar and Thornton, 2001] Valdar, W. S. and Thornton, J. M. (2001). Conservation helps to identify biologically relevant crystal contacts. *Journal of molecular biology*, 313(2):399–416.

[Van Kempen et al., 2024] Van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L., Söding, J., and Steinegger, M. (2024). Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246.

[Van Zundert et al., 2016] Van Zundert, G., Rodrigues, J., Trellet, M., Schmitz, C., Kastritis, P., Karaca, E., Melquiond, A., van Dijk, M., De Vries, S., and Bonvin, A. (2016). The haddock2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology*, 428(4):720–725.

[van Zundert et al., 2015] van Zundert, G. C., Melquiond, A. S., and Bonvin, A. M. (2015). Integrative modeling of biomolecular complexes: Haddocking with cryo-electron microscopy data. *Structure*, 23(5):949–960.

[Vassilev et al., 2004] Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., Fotouhi, N., and Liu, E. A. (2004). In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science*, 303(5659):844–8.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Venkatraman et al., 2009] Venkatraman, V., Yang, Y. D., Sael, L., and Kihara, D. (2009). Protein-protein docking using region-based 3d zernike descriptors. *BMC bioinformatics*, 10:1–21.

[Vreven et al., 2015] Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P. A., Fernandez-Recio, J., Bonvin, A. M. J. J., and Weng, Z. (2015). Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, 427(19):3031–3041.

[Wang et al., 2019] Wang, S., Li, M., Guo, L., Cao, Z., and Fei, Y. (2019). Efficient utilization on pssm combining with recurrent neural network for membrane protein types prediction. *Computational Biology and Chemistry*, 81:9–15.

[Wang et al., 2021] Wang, X., Flannery, S. T., and Kihara, D. (2021). Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences*, 8:647915.

[Wang et al., 2020] Wang, X., Terashi, G., Christoffer, C. W., Zhu, M., and Kihara, D. (2020). Protein docking model evaluation by 3d deep convolutional neural networks. *Bioinformatics*, 36(7):2113–2118.

[Wernimont and Edwards, 2009] Wernimont, A. and Edwards, A. (2009). In situ proteolysis to generate crystals for structure determination: an update. *PloS one*, 4(4):e5094.

[Williamson and Sutcliffe, 2010] Williamson, M. P. and Sutcliffe, M. J. (2010). Protein–protein interactions. *Biochemical Society Transactions*, 38(4):875–878.

[Wu et al., 2023a] Wu, T., Guo, Z., and Cheng, J. (2023a). Atomic protein structure refinement using all-atom graph representations and se (3)-equivariant graph transformer. *Bioinformatics*, 39(5):btad298.

[Wu et al., 2023b] Wu, Z., Guo, M., Jin, X., Chen, J., and Liu, B. (2023b). Cfago: cross-fusion of network and attributes based on attention mechanism for protein function prediction. *Bioinformatics*, 39(3):btad123.

[Xia et al., 2022] Xia, C., Feng, S. H., Xia, Y., Pan, X., and Shen, H. B. (2022). Fast protein structure comparison through effective representation learning with contrastive graph neural networks. *PLoS Comput Biol*, 18(3):e1009986.

[Xia et al., 2010] Xia, J. F., Han, K., and Huang, D. S. (2010). Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein Pept Lett*, 17(1):137–45.

[Xia et al., 2021] Xia, Y., Xia, C.-Q., Pan, X., and Shen, H.-B. (2021). Graphbind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9):e51–e51.

[Xu and Dunbrack Jr, 2020] Xu, Q. and Dunbrack Jr, R. L. (2020). Protcid: a data resource for structural information on protein interactions. *Nature communications*, 11(1):711.

[Yang et al., 2020] Yang, F., Fan, K., Song, D., and Lin, H. (2020). Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC Bioinformatics*, 21(1):323.

[Yang et al., 2021a] Yang, X., Yang, S., Lian, X., Wuchty, S., and Zhang, Z. (2021a). Transfer learning via multi-scale convolutional neural layers for human-virus protein-protein interaction prediction. *Bioinformatics*.

[Yang et al., 2021b] Yang, X., Yang, S., Lian, X., Wuchty, S., and Zhang, Z. (2021b). Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction. *Bioinformatics*, 37(24):4771–4778.

[Yao et al., 2019] Yao, Y., Du, X., Diao, Y., and Zhu, H. (2019). An integration of deep learning with feature embedding for protein-protein interaction prediction. *PeerJ*, 7:e7126.

[Ying et al., 2019] Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.

[You et al., 2015] You, Z. H., Chan, K. C., and Hu, P. (2015). Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PLoS One*, 10(5):e0125811.

[Yu and Guerois, 2016] Yu, J. and Guerois, R. (2016). Ppi4dock: large scale assessment of the use of homology models in free docking over more than 1000 realistic targets. *Bioinformatics*, 32(24):3760–3767.

[Yue et al., 2022] Yue, Y., Ye, C., Peng, P.-Y., Zhai, H.-X., Ahmad, I., Xia, C., Wu, Y.-Z., and Zhang, Y.-H. (2022). A deep learning framework for identifying essential proteins based on multiple biological information. *BMC bioinformatics*, 23(1):318.

[Yueh et al., 2017] Yueh, C., Hall, D. R., Xia, B., Padhorny, D., Kozakov, D., and Vajda, S. (2017). Cluspro-dc: Dimer classification by the cluspro server for protein–protein docking. *Journal of molecular biology*, 429(3):372–381.

[Zahiri et al., 2013] Zahiri, J., Yaghoubi, O., Mohammad-Noori, M., Ebrahimpour, R., and Masoudi-Nejad, A. (2013). Ppievo: protein-protein interaction prediction from pssm based evolutionary information. *Genomics*, 102(4):237–42.

[Zhang and Skolnick, 2004] Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710.

[Zhang et al., 2022] Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. (2022). Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*.

[Zhao et al., 2022] Zhao, X., Zhang, Y., and Du, X. (2022). Dfpin: Deep learning-based protein-binding site prediction with feature-based non-redundancy from rna level. *Comput Biol Med*, 142:105216.

[Zhu et al., 2022] Zhu, F., Li, F., Deng, L., Meng, F., and Liang, Z. (2022). Protein interaction network reconstruction with a structural gated attention deep model by incorporating network structure information. *Journal of Chemical Information and Modeling*, 62(2):258–273.

[Zhu et al., 2001] Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. (2001). Global analysis of protein activities using proteome chips. *science*, 293(5537):2101–2105.

[Zhu et al., 2006] Zhu, H., Domingues, F. S., Sommer, I., and Lengauer, T. (2006). Noxclass: prediction of protein-protein interaction types. *BMC bioinformatics*, 7:1–15.

[Zhu et al., 2020] Zhu, X., Liu, L., He, J., Fang, T., Xiong, Y., and Mitchell, J. C. (2020). ipnhot: a knowledge-based approach for identifying protein-nucleic acid interaction hot spots. *BMC Bioinformatics*, 21(1):289.

# Appendix A

# SOURCE CODE AND SCRIPTS

Source code and scripts are available through our GitHub repository at `https://github.com/ku-cosbi/ProInterVal`.

## *A.1 Feature Calculation*

The following values stored in Python dictionaries are used for the feature calculation.

```
VDW_RADII = {'C': 1.76, 'N': 1.65, 'O': 1.40, 'CA': 1.87,
             'H': 1.20, 'S': 1.85, 'CB': 1.87, 'CZ': 1.76, 'NZ': 1.50,
             'CD': 1.81, 'CE': 1.81, 'CG': 1.81, 'C1': 1.80, 'P': 1.90}

VDW_RADII_EXTENDED = {'ALA': {'N': 1.65, 'CA': 1.87, 'C': 1.76,
                              'O': 1.40, 'CB': 1.87, 'OXT': 1.40},
                      'ARG': {'N': 1.65, 'CA': 1.87, 'C': 1.76,
                              'O': 1.40, 'CB': 1.87, 'CG': 1.87, 'CD': 1.87,
                              'NE': 1.65, 'CZ': 1.76, 'NH1': 1.65,
                              'NH2': 1.65, 'OXT': 1.40},
                      'ASP': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                              'CB': 1.87, 'CG': 1.76, 'OD1': 1.40,
                              'OD2': 1.40, 'OXT': 1.40},
                      'ASN': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                              'CB': 1.87, 'CG': 1.76, 'OD1': 1.40,
                              'ND2': 1.65, 'OXT': 1.40},
                      'CYS': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                              'CB': 1.87, 'SG': 1.85, 'OXT': 1.40},
                      'GLU': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                              'CB': 1.87, 'CG': 1.87, 'CD': 1.76,
                              'OE1': 1.40, 'OE2': 1.40, 'OXT': 1.40},
                      'GLN': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                              'CB': 1.87, 'CG': 1.87, 'CD': 1.76,
                              'OE1': 1.40, 'NE2': 1.65, 'OXT': 1.40},
                      'GLY': {'N': 1.65, 'CA': 1.87, 'C': 1.76,
                              'O': 1.40, 'OXT': 1.40},
                      'HIS': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                              'CB': 1.87, 'CG': 1.76, 'ND1': 1.65,
                              'CD2': 1.76, 'CE1': 1.76, 'NE2': 1.65,
                              'OXT': 1.40},
```

```
                        'ILE': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG1': 1.87, 'CG2': 1.87,
                                'CD1': 1.87, 'OXT': 1.40},
                        'LEU': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.87, 'CD1': 1.87,
                                'CD2': 1.87, 'OXT': 1.40},
                        'LYS': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.87, 'CD': 1.87,
                                'CE': 1.87, 'NZ': 1.50, 'OXT': 1.40},
                        'MET': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.87, 'SD': 1.85,
                                'CE': 1.87, 'OXT': 1.40},
                        'PHE': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.76, 'CD1': 1.76,
                                'CD2': 1.76, 'CE1': 1.76, 'CE2': 1.76,
                                'CZ': 1.76, 'OXT': 1.40},
                        'PRO': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.87, 'CD': 1.87,
                                'OXT': 1.40},
                        'SER': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'OG': 1.40, 'OXT': 1.40},
                        'THR': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'OG1': 1.40, 'CG2': 1.87,
                                'OXT': 1.40},
                        'TRP': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.76, 'CD1': 1.76,
                                'CD2': 1.76, 'NE1': 1.65, 'CE2': 1.76,
                                'CE3': 1.76, 'CZ2': 1.76, 'CZ3': 1.76,
                                'CH2': 1.76, 'OXT': 1.40},
                        'TYR': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.76, 'CD1': 1.76,
                                'CD2': 1.76, 'CE1': 1.76, 'CE2': 1.76,
                                'CZ': 1.76, 'OH': 1.40, 'OXT': 1.40},
                        'VAL': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG1': 1.87, 'CG2': 1.87,
                                'OXT': 1.40},
                        'ASX': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.76, 'AD1': 1.50, 'AD2': 1.50},
                        'GLX': {'N': 1.65, 'CA': 1.87, 'C': 1.76, 'O': 1.40,
                                'CB': 1.87, 'CG': 1.76, 'CD': 1.87,
                                'AE1': 1.50, 'AE2': 1.50, 'OXT': 1.40},
                        'ACE': {'C': 1.76, 'O': 1.40, 'CA': 1.87},
                        'PCA': VDW_RADII, 'UNK': VDW_RADII}

DICT_ATOM = {'ACE': ['CA'], 'ALA': ['CB'], 'GLY': ['CA'], 'SER': ['OG'],
```

```
            'ASN': ['OD1', 'ND2'], 'ASP': ['OD1', 'OD2'],
            'CYS': ['SG'], 'PRO': ['CB', 'CG', 'CD'], 'THR': ['OG1'],
            'VAL': ['CG1', 'CG2'],
            'ARG': ['NE', 'NH1', 'NH2'], 'GLN': ['OE1', 'NE2'],
            'GLU': ['OE1', 'OE2'],
            'HIS': ['CG', 'ND1', 'CD2', 'CE1', 'NE2'], 'ILE': ['CD1'],
            'LEU': ['CD1', 'CD2'], 'LYS': ['NZ'],
            'MET': ['SD'], 'PHE': ['CG', 'CD1', 'CD2', 'CE1', 'CE2', 'CZ'],
            'TRP': ['CD1', 'CD2', 'NE1', 'CE2', 'CE3', 'CZ2', 'CZ3', 'CH2'],
            'TYR': ['CG', 'CD1', 'CD2', 'CE1', 'CE2', 'CZ', 'OH']}

# standard ASA
MAX_ASA = {'ALA': 107.95, 'CYS': 134.28, 'ASP': 140.39, 'GLU': 172.25,
           'PHE': 199.48, 'GLY': 80.10, 'HIS': 182.88,
           'ILE': 175.12, 'LYS': 200.81, 'LEU': 178.63, 'MET': 194.15,
           'ASN': 143.94, 'PRO': 136.13, 'GLN': 178.50,
           'ARG': 238.76, 'SER': 116.50, 'THR': 139.27, 'VAL': 151.44,
           'TRP': 249.36, 'TYR': 212.76}

# contact potential matrix
CONTACT_POTENTIALS = [
    [-1.77, -2.24, -2.77, -3.5, -3.79, -3.43, -3.02, -3.76, -3.34,
     -3.77, -1.68, -1.81, -1.84, -1.63, -1.58, -3.02,
     -1.32, -2.71, -2.47, -1.01],
    [0, -3.51, -3.86, -4.45, -5.02, -3.99, -4.42, -4.43, -3.96,
     -4.66, -2.49, -2.38, -2.74, -2.06, -2.69, -3.65, -1.91,
     -3.26, -3.29, -1.78],
    [0, 0, -4.86, -5.31, -6.03, -4.82, -5.16,
     -5.33, -4.54, -4.7, -2.95,
     -2.87, -3.28, -2.76, -3.2, -4.29, -2.19, -3.7,
     -4.14, -2.43],
    [0, 0, 0, -5.97, -6.67, -5.53, -5.37, -6.11, -5.39, -5.79,
     -3.47, -3.61, -3.46, -3.42, -3.99, -4.65, -2.88, -4.29,
     -4.55, -3.09],
    [0, 0, 0, 0, -7.16, -6.13, -6.24, -6.65, -5.67, -6.4, -4.12,
     -4.28, -4.16, -3.94, -4.35, -5.36, -3.24, -5.01, -4.85,
     -3.75],
    [0, 0, 0, 0, 0, -7.23, -5.07, -4.81, -4.71, -4.7, -3.41,
     -2.92, -3.56, -1.76, -3.02, -4.37, -2.28, -4.41, -4.79,
     -2.97],
    [0, 0, 0, 0, 0, 0, -5.89, -5.58, -5.01, -5.49, -3.33, -3.33,
     -3.69, -3.05, -3.76, -4.46, -2.36, -4.13, -4.47,
     -3.11],
    [0, 0, 0, 0, 0, 0, 0, -6.45, -5.33, -5.72, -4.01, -3.85,
     -3.52, -3.89, -4.1, -4.89, -2.7, -4.51, -3.82, -3.46],
```

```
    [0, 0, 0, 0, 0, 0, 0, 0, -4.58, -5.12, -3.43, -3.33,
     -3.63, -3.14, -3.62, -4.75, -2.64, -4.32, -3.78, -2.92],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, -5.16, -2.86, -3.48,
     -3.81, -3.31, -3.84, -4.87, -3.12, -4.54, -4.5, -3.96],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -2.14, -2.22, -2.34,
     -2.13, -2.57, -3.06, -1.91, -2.78, -3.12, -1.56],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -2.12, -2.41,
     -2.05, -2.34, -3.44, -1.64, -2.43, -2.73, -1.24],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -2.47,
     -2.5, -2.35, -3.15, -2.47, -3.92, -2.93, -1.24],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1.99,
     -2.43, -3, -1.63, -2.67, -3.05, -1.01],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -2.85,
     -3.45, -2.83, -4.05, -3.15, -1.7],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
     -4.71, -2.56, -4.13, -4.2, -2.67],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
     -1.02, -2.11, -2.14, -0.5],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
     0, -3.98, -3.24, -2.43],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
     0, 0, 0, -3.08, -1.8],
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
     0, 0, 0, 0, 0, 0, -0.83]]

# indices of residues
PP_indices = {'GLY': 0, 'ALA': 1, 'VAL': 2, 'ILE': 3, 'LEU': 4,
              'CYS': 5, 'MET': 6, 'PHE': 7, 'TYR': 8, 'TRP': 9,
              'SER': 10, 'THR': 11, 'ASP': 12, 'ASN': 13,
              'GLU': 14, 'GLN': 15, 'LYS': 16, 'ARG': 17, 'HIS': 18,
              'PRO': 19}

# polarity (polar: 0, non-polar: 1)
POLARITY = {'ALA': 1, 'CYS': 0, 'ASP': 0, 'GLU': 0,
            'PHE': 1, 'GLY': 1, 'HIS': 0,
            'ILE': 1, 'LYS': 0, 'LEU': 1, 'MET': 1,
            'ASN': 0, 'PRO': 1, 'GLN': 0,
            'ARG': 0, 'SER': 0, 'THR': 0, 'VAL': 1,
            'TRP': 1, 'TYR': 0}

# charge (positive: 1, neutral: 0, negative: -1)
CHARGE = {'ALA': 0, 'CYS': 0, 'ASP': -1, 'GLU': -1,
          'PHE': 0, 'GLY': 0, 'HIS': 1,
          'ILE': 0, 'LYS': 1, 'LEU': 0, 'MET': 0,
          'ASN': 0, 'PRO': 0, 'GLN': 0,
```

```
          'ARG': 1, 'SER': 0, 'THR': 0, 'VAL': 0,
          'TRP': 0, 'TYR': 0}

hydrophobicity_scale = {
        'ALA': 1.8, 'CYS': 2.5, 'ASP': -3.5,
        'GLU': -3.5, 'PHE': 2.8, 'GLY': -0.4,
        'HIS': -3.2, 'ILE': 4.5, 'LYS': -3.9,
        'LEU': 3.8, 'MET': 1.9, 'ASN': -3.5,
        'PRO': -1.6, 'GLN': -3.5, 'ARG': -4.5,
        'SER': -0.8, 'THR': -0.7, 'VAL': 4.2,
        'TRP': -0.9, 'TYR': -1.3
    }
```

### *A.1.1 NACCESS*

Naccess is an external tool that is used to calculate the accessible surface area of proteins [Ding and Arnold, 2006]. A solvent probe is rolled on the target molecule as the basic method. The user can select the solvent's radius; nevertheless, 1.4 Å is the default value. The accessible surface area is given by the path taken by the probe's center. PDB format files are accepted as input by Naccess. In addition to the accessible surface area, each residue's relative accessible area is included in the Naccess output file. The percentage of a residue's accessibility in relation to its accessibility in the tripeptide ALA-XALA is known as its relative accessibility. This residue is typically classified as surface residue if its value is greater than 5% [Gong et al., 2005, Aytuna et al., 2005]. In this study, Naccess was used with default values to calculate proteins' ASA values in the monomer and complex states.

## *A.2 Models*

Codes are written in Python. PyTorch library is used to build and train deep learning models. The biopython package is utilized to parse and handle protein structure files in the pdb format. Scientific libraries, including NumPy, SciPy, Pandas, and Scikit-learn, are employed for numerical operations, scientific computations, data manipulation and analysis, and machine learning utilities and metrics. Trained models are deposited to Hugging Face and available at `https://huggingface.co/cosbi-ku/ProInterVal`. We have added a Python script (run.py) to our GitHub repository, which allows users to provide a new dataset as input for our final models and retrieve predictions. We have also added a "USAGE" section to our README that explains how to use these models to make predictions.