İZMİR
KÂTİP ÇELEBİ
ÜNİVERSİTESİ
2010
GRADUATE SCHOOL OF NATURAL
AND APPLIED SCIENCES

Applications of Deep Learning for Binding Site Predictions in Proteins

Submitted to the Graduate School of Natural and Applied Sciences in partial fulfillment of the requirements for the degree of

Master of Science in Biomedical

Engineering

by

Nihat Tolga Filoğlu

ORCID 0009-0001-8094-9763

Advisor: Assist. Prof. Dr. Onan Güren

January, 2024

This is to certify that we have read the thesis **Applications of Deep Learning for Binding Site Predictions in Proteins** submitted by **Nihat Tolga Filoğlu**, and it has been judged to be successful, in scope and in quality, at the defense exam and accepted by our jury as a MASTER'S THESIS.

APPROVED BY:

Advisor:	**Assist. Prof. Dr. Onan Güren** İzmir Kâtip Çelebi University
Co-advisor:	**Assist. Prof. Dr. Arzu Uyar** İzmir Yüksek Teknoloji Enstitüsü
Committee Members:	
	Assist. Prof. Dr. Onan Güren İzmir Kâtip Çelebi University
	Assist. Prof. Dr. Arzu Uyar İzmir Yüksek Teknoloji Enstitüsü
	Assoc. Prof. Dr. Çağatay CEYLAN İzmir Yüksek Teknoloji Enstitüsü
	Prof. Dr. Aytuğ Onan İzmir Kâtip Çelebi University
	Assoc. Prof. Dr. Didem Şen Karaman İzmir Kâtip Çelebi University

Date of Defense: December 26, 2024

Declaration of Authorship

I, **Nihat Tolga Filoğlu**, declare that this thesis titled **Applications of Deep Learning for Binding Site Predictions in Proteins** and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for the Master's degree at this university.

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.

- Where I have consulted the published work of others, this is always clearly attributed.

- Where I have quoted from the work of others, the source is always given. This thesis is entirely my own work, with the exception of such quotations.

- I have acknowledged all major sources of assistance.

- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 12.16.2024

Applications of Deep Learning for Binding Site Predictions in Proteins

Abstract

Proteins are fundamental macromolecules that perform a wide range of biological functions, largely determined by their three-dimensional structures and binding interactions. Accurate identification of protein binding sites, particularly orthosteric and allosteric sites, is crucial for drug discovery, biotechnology, and understanding biological processes. Traditional computational methods, such as sequence alignment, homology modeling, and molecular docking, often struggle to handle the dynamic nature and structural complexity of proteins. To address these challenges, this study explores the use of Graph Neural Networks (GNNs) to predict protein binding sites with enhanced accuracy and efficiency.

This research employs three GNN architectures: Graph Attention Networks (GAT), Graph Convolutional Networks (GCN), and Equivariant Graph Neural Networks (EGNN). The models are trained on a combined dataset from CasBench (2023) and BioLiP, comprising approximately 300 protein structures with annotated orthosteric and allosteric binding sites. Node features such as amino acid identity, dihedral angles, solvent-accessible surface area (SASA), and secondary structure elements (SSE) are integrated with edge features like Euclidean distances and cosine angles to construct informative protein graphs.

Results demonstrate that the EGNN model excels in predicting binding sites for proteins with stable and well-defined structures due to its ability to maintain geometric consistency. The GAT and GCN models effectively capture local structural patterns and relational features within protein graphs, with GAT leveraging dynamic attention mechanisms to highlight key residues. Comparative analysis with existing methods highlights the advantages of the proposed GNN approach in terms of precision, recall, and AUC-ROC metrics.

Despite these promising results, challenges remain in predicting binding sites for proteins with small ligands or high structural flexibility. Future work will focus on incorporating protein dynamics, ligand-specific features, and hybrid models that combine GNNs with transformer architectures to improve prediction accuracy and generalization.

This study contributes to the evolving field of computational biology by offering a robust, interpretable, and efficient framework for protein binding site prediction, supporting advancements in drug discovery and personalized medicine.

Keywords: Protein Binding Sites, Orthosteric Sites, Allosteric Sites, Graph Neural Network (GNN), Computational Biology, Drug Discovery, Protein Structure, Bioinformatics, Deep Learning

Proteinlerdeki Bağlanma Yeri Tahminleri İçin Derin Öğrenme Uygulamaları

Öz

Proteinler, üç boyutlu yapıları ve bağlanma etkileşimleri tarafından belirlenen, geniş bir biyolojik işlev yelpazesini yerine getiren temel makromoleküllerdir. Özellikle ortosterik ve allosterik bağlanma bölgelerinin doğru bir şekilde tespiti, ilaç keşfi, biyoteknoloji ve biyolojik süreçlerin anlaşılması açısından büyük önem taşımaktadır. Dizi hizalama, homoloji modelleme ve moleküler yerleştirme gibi geleneksel hesaplamalı yöntemler, proteinlerin dinamik doğası ve yapısal karmaşıklığıyla başa çıkmakta yetersiz kalabilmektedir. Bu zorlukların üstesinden gelmek için bu çalışmada, Graf Sinir Ağları (GNN) kullanılarak protein bağlanma bölgelerinin daha yüksek doğruluk ve verimlilikle tahmin edilmesi amaçlanmıştır.

Bu araştırmada üç farklı GNN mimarisi kullanılmıştır: Graf Dikkat Ağları (GAT), Graf Konvolüsyonel Ağları (GCN) ve Eşdeğerlik Temelli Graf Sinir Ağları (EGNN). Modeller, CasBench (2023) ve BioLiP veri setlerinin birleştirilmiş haliyle eğitilmiştir. Bu veri seti, ortosterik ve allosterik bağlanma bölgeleri ile etiketlenmiş yaklaşık 3000 protein yapısını içermektedir. Protein grafikleri oluşturulurken düğüm özellikleri arasında amino asit kimliği, diyhedral açılar, çözücü erişebilir yüzey alanı (SASA) ve ikincil yapı elemanları (SSE) gibi bilgiler; kenar özellikleri arasında ise Öklid uzaklıkları ve kosinüs açıları kullanılmıştır.

Sonuçlar, EGNN modelinin geometrik tutarlılığı koruma yeteneği sayesinde, istikrarlı ve iyi tanımlanmış protein yapılarında bağlanma bölgelerini doğru bir şekilde tahmin ettiğini göstermektedir. GAT ve GCN modelleri ise protein grafikleri içindeki yerel yapısal desenleri ve ilişkisel özellikleri etkili bir şekilde yakalamış; özellikle GAT, önemli kalıntıları vurgulamak için dinamik dikkat mekanizmalarından yararlanmıştır. Mevcut yöntemlerle yapılan karşılaştırmalar, önerilen GNN yaklaşımının doğruluk, geri çağırma ve AUC-ROC ölçütleri açısından avantajlarını ortaya koymaktadır.

Bu umut verici sonuçlara rağmen, küçük ligandlara sahip veya yüksek yapısal esnekliğe sahip proteinlerde bağlanma bölgesi tahmininde bazı zorluklar devam etmektedir. Gelecekteki çalışmalar, protein dinamiklerini, ligandlara özgü özellikleri ve GNN'leri dönüştürücü (transformer) mimarilerle birleştiren hibrit modelleri entegre ederek tahmin doğruluğunu ve genelleştirilebilirliği artırmaya odaklanacaktır.

Bu çalışma, protein bağlanma bölgesi tahmininde sağlam, yorumlanabilir ve verimli bir çerçeve sunarak, ilaç keşfi ve kişiselleştirilmiş tıp alanlarındaki gelişmelere katkıda bulunmayı hedeflemektedir.

Anahtar Kelimeler: Protein Bağlanma Bölgeleri, Ortosterik Bölgeler, Allosterik Bölgeler, Graf Sinir Ağı (GNN), Hesaplamalı Biyoloji, İlaç Keşfi, Protein Yapısı, Biyoinformatik, Derin Öğrenme

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
Cα	Carbon Alpha Atom
DCA	Distance Correlation Coefficient
DCC	Distance Covariance
DSSP	Dictionary of Secondary Structure of Proteins
EGNN	Equivariant Graph Neural Network
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
PDB	Protein Data Bank
RNN	Recurrent Neural Network
SASA	Solvent Accessible Surface Area
SSE	Secondary Structure Elements
SVM	Support Vector Machine

List of Symbols

α (alpha)	Used to represent the angle in dihedral angles (e.g., phi, psi, omega) and the LeakyReLU activation function parameter.
τ (tau)	Dihedral angle in protein structures.
θ (theta)	Dihedral angle in protein structures.
φ (phi)	Dihedral angle in protein structures.
ψ (psi)	Dihedral angle in protein structures.
ω (omega)	Dihedral angle in protein structures.
e	Base of the natural logarithm, approximately equal to 2.71828.
σ (sigma)	Standard deviation, often used in statistical contexts.
$\hat{y}$	Predicted logit (output of the model before applying the sigmoid function).
y	Ground truth label (0 or 1).
$\sum$ (Sigma)	Summation symbol, used to denote the sum of a series of terms.
$\prod$ (Pi)	Product symbol, used to denote the product of a series of terms.
$\|\|x\|\|$	Norm of a vector x, representing its magnitude.

Chapter 1

Introduction

1.1 Background

Proteins are vital macromolecules that play crucial roles in nearly every biological process. Proteins are composed of amino acids joined by peptide bonds, creating a polypeptide chain that folds into a unique three-dimensional structure, which ultimately determines the protein's function. Studying protein structures is essential for understanding basic biological processes and has important implications in areas such as pharmaceuticals, medical treatment, and biotechnology.

The structure of proteins can be described at four levels:

- **Primary Structure:** The linear sequence of amino acids in the polypeptide chain, as determined by the encoding gene.

- **Secondary Structure:** Local folding of the polypeptide chain forming shapes like α-helices and β-sheets, stabilized by hydrogen bonds.

- **Tertiary Structure:** The overall three-dimensional shape of a single polypeptide chain, formed through interactions between the amino acids' side chains (R-groups).

- **Quaternary Structure:** The structure formed when two or more polypeptide chains (subunits) assemble into a functional protein complex.

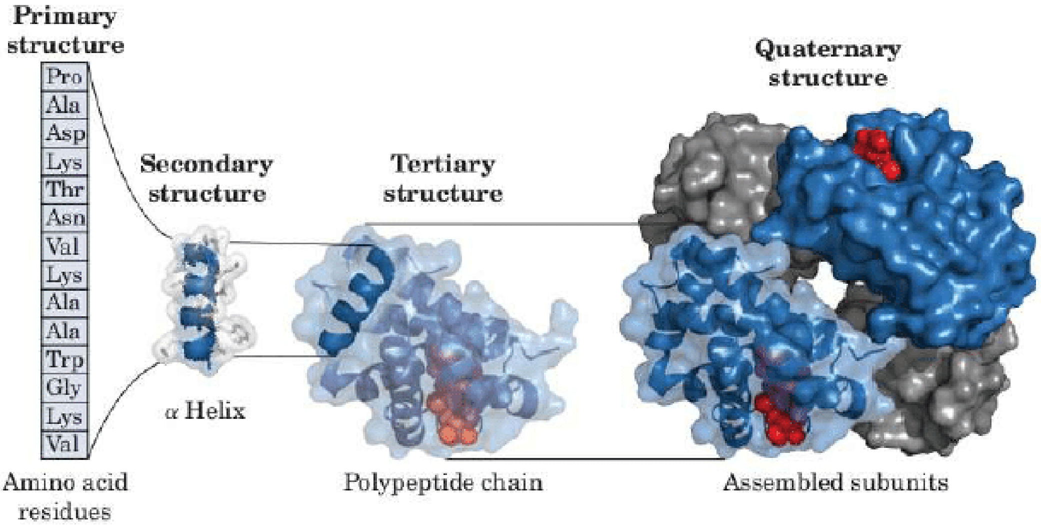

Figure 1.1: The four levels of protein structure (Rashid et al., 2015)

Understanding these structures is essential in understanding the functions of proteins, including enzymatic reactions, providing support, transporting molecules, signaling within cells, and defending the body (Jones & Smith, 2023; Gordon & Hahn, 2022). Furthermore, with the advancement of computational biology, the precise forecasting of protein structures and binding locations is increasingly vital. The rise of high-throughput methods and large biological datasets has driven the merging of computational methods with experimental biology, providing unique chances to speed up drug discovery and development.

1.2 Protein Binding Sites

In most general terms, binding sites on proteins are specific regions of the protein to which a myriad of other molecules may bind, such as ligands or substrates, or to which additional proteins may bind. There are two broad categories of binding sites:

- **Orthosteric Binding Sites:** The primary active sites where endogenous ligands bind directly, facilitating the protein's biological function (Cherezov et al., 2007).

- **Allosteric Binding Sites:** Distinct sites from the orthosteric site that, when bound by a molecule, modulate the protein's activity, often enhancing or inhibiting its function (Olsen & Sieghart, 2008).

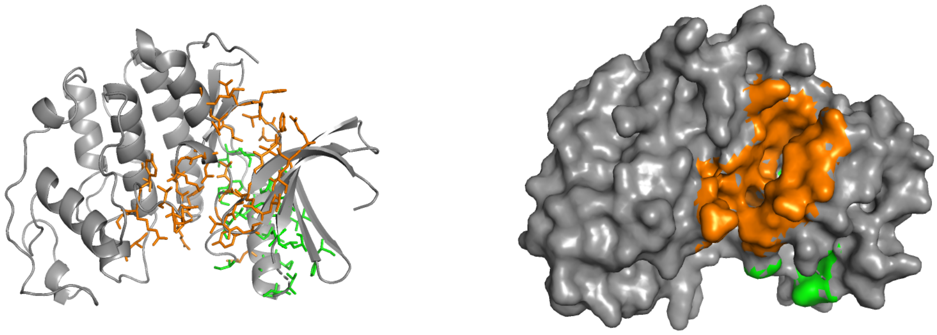

Figure 1.2: Representation of orthosteric (orange) and allosteric (green) binding sites in the protein structure 1KE5

Recognizing and explaining binding sites is essential in the development of new medications, as they provide potential points for medical intervention (Edfeldt et al., 2011). Recent progress in computational biology has greatly improved our capacity to anticipate these binding locations, simplifying the creation of drugs that are more selective and efficient, ultimately decreasing the chances of adverse side effects.

1.3 Challenges in Predicting Binding Sites

Prediction of the binding site is a complex challenge due to many reasons, despite advances in computational biology:

- **Structural Complexity:** Proteins are dynamic molecules with complex three-dimensional structures that can obscure binding sites, making them difficult to predict (Karplus & McCammon, 2002).

- **Limitations of Traditional Methods:** Methods like sequence alignment, homology modeling, and molecular docking frequently lack accuracy, especially when handling flexible proteins or distant evolutionary connections (Baker & Sali, 2001; Meng et al., 2011).

- **Diversity of Binding Sites:** The variety of binding sites in different proteins poses a challenge for developing a prediction method that can be universally applied (Keskin et al., 2008).

Precise identification of binding sites is crucial in pharmaceutical studies, in comprehending protein function, and in furthering biomedical research. DiMasi et al. (2016) introduce a new scenario by using artificial intelligence (AI) and machine learning (ML), more precisely, Graph Neural Networks GNNs have recently developed as effective tools for modeling intricate relationships in the structure of proteins, hence offering a promising approach to enhancing accuracy and efficiency in the identification of binding sites.

1.4 Objectives of the Study

This study aims to develop a computational method using Graph Neural Networks (GNNs) to predict orthosteric and allosteric binding sites in proteins. The specific objectives include:

- **Developing a GNN Model:** Creating a model capable of accurately predicting binding sites based on structural and physicochemical properties of proteins.

- **Implementing Advanced Techniques:** Use state-of-the-art machine learning methodologies and graph theory-based methods to enhance the prediction accuracy.

- **Evaluating Model Performance:** Assessing the model using metrics such as accuracy, precision, recall, F1-score, and AUC (Chicco & Jurman, 2020).

- **Comparing with Existing Techniques:** Conducting a comparative analysis with traditional and similar deep learning prediction methods to highlight the advantages of the GNN approach and demonstrate the differences with similar studies (Ehrt et al., 2018).

1.5 Significance of the Study

The significance of this study lies in its potential to contribute to the broader trends in computational biology and drug discovery. Much more accurate and effective ways of predicting the site of protein binding can be developed with GNN technology to attend to the most pressing need. These are of particular importance for the prevailing drift toward personalized medicine and targeted drugs, wherein knowledge is paramount on exact interactions between the drugs and their protein targets.

On the other hand, another step toward integrating GNNs with traditional methods represents a novel phase in the course of the continuous evolution of computational approaches to biological problems. The current study did not only focus on predictive performance improvement but has given valuable insights into underlying mechanisms of protein-ligand interactions that can further help in the design of more efficient therapeutic strategies.

Chapter 2

Literature Review

2.1 Approaches to Protein Binding Site Prediction

Understanding protein-ligand interactions is foundational to drug discovery and biotechnology. Early methods relied on sequence conservation, structural alignments, and molecular docking, providing valuable but sometimes limited insights into binding site locations (Brylinski & Skolnick, 2008). However, these approaches often missed crucial dynamic and structural properties of proteins, especially in predicting non-canonical binding sites like allosteric regions.

In recent years, structure-based methods have become more prominent, integrating 3D features of protein structures. For example, algorithms like DeepSite use 3D convolutional neural networks (CNNs) to predict potential binding pockets based on protein surface characteristics (Jiménez et al., 2017). In addition, models such as DeltaVina XGB have successfully combined docking techniques with feature re-ranking for greater prediction accuracy (Trott & Olson, 2010).

2.2 Machine Learning Approaches for Binding Site Prediction

Machine learning (ML) revolutionized the field by enabling the integration of large-scale biological data into predictive models. These approaches often outperformed classical methods by incorporating evolutionary, sequence-based, and sometimes structural information. Models like DeepBind and iDeep made groundbreaking strides in RNA-protein interaction prediction using recurrent and convolutional neural networks (Alipanahi et al., 2015).

For protein binding sites, ML methods enhanced accuracy by leveraging vast experimental datasets. However, they often lacked the spatial context of the protein structure, necessitating the evolution towards more advanced models like graph-based learning.

2.3 Deep Learning Approaches for Binding Site Prediction

Deep learning techniques such as CNNs, Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and E(n)-Equivariant GNNs have increasingly been applied to protein-ligand binding predictions (Liu et al., 2020). CNNs are particularly effective in capturing local structural motifs, as evidenced in DeepSite (Jiménez et al., 2017). However, their limitation lies in the handling of non-local relationships.

Graph-based methods, which model proteins as 3D graphs where nodes represent amino acids and edges reflect spatial proximity or atomic interaction, address this limitation. GCNs capture the overall protein topology, while GATs use attention mechanisms to highlight crucial binding residues, significantly improving prediction of both orthosteric and allosteric sites (Velickovic et al., 2018).

Additionally, newer architectures like E(n)-Equivariant GNNs integrate geometric transformations while preserving physical symmetries, leading to more accurate modeling of protein structures (Bronstein et al., 2017). These models have demonstrated high performance, especially in datasets where structural flexibility plays a key role in binding site formation (Rives et al., 2021).

2.4 Allosteric and Orthosteric Site Prediction

Orthosteric sites, typically the main active sites of proteins, are often more conserved and easier to predict using both sequence- and structure-based approaches. Models leveraging GNNs or CNNs excel in identifying these regions with high accuracy.

Allosteric sites, however, present a unique challenge due to their dynamic nature and lack of strong conservation patterns. These sites are often distant from orthosteric regions but play crucial roles in modulating protein function through long-range conformational changes (Süel et al., 2003). Models like AlloPred combine structural and sequence data to predict these elusive binding sites, demonstrating success in capturing the allosteric communication between distant residues in protein networks (Greener et al., 2015). GNN-based models are particularly well-suited for this task, as they model the protein as a network of interacting residues, making it possible to capture the indirect communication pathways characteristic of allosteric regulation.

2.5 Explainability in Deep Learning Models

While deep learning models achieve high predictive accuracy, their lack of interpretability presents a challenge in fields like drug discovery, where understanding the decision-making process is crucial. Graph Attention Networks

(GATs) provide a partial solution to this by using attention mechanisms to highlight key residues involved in predictions, offering a level of explainability (Velickovic et al., 2018).

Additionally, feature attribution methods such as saliency maps and Grad-CAM have been applied to CNNs and GNNs to visually interpret which regions of the protein contribute most to the prediction (Smilkov et al., 2017). These tools allow researchers to gain insights into the functional importance of specific residues or binding pockets.

Explainability is further enhanced by methods that decompose predictions into biologically meaningful components, helping researchers identify potentially novel binding sites or validate model predictions against known experimental data. For instance, by analyzing which features (sequence motifs, structural properties) contribute most to the prediction, researchers can assess the biological relevance of the predicted binding sites.

2.6 Future Directions and Multi-Modal Integration

Looking forward, integrating multi-modal data, including protein structure, sequence, and evolutionary information, promises to further enhance binding site prediction models. Advances in transformer-based architectures, known for capturing long-range dependencies in sequences, offer new avenues for improving model accuracy and robustness in protein-ligand interaction prediction (Rives et al., 2021).

The application of deep learning models like AlphaFold, which revolutionized the prediction of protein structures, suggests that integrating structure prediction with binding site identification could lead to more accurate and generalized models (Jumper et al., 2021). Future work will likely involve the development of hybrid models that combine graph-based approaches with sequence-based transformers, providing a more holistic view of protein-ligand interactions.

Chapter 3

Methodology

3.2 Dataset

3.2.1 Description of the Dataset

The dataset used in this study combines the CasBench (2023) and BioLiP datasets to enhance the volume and diversity of data, ensuring robust model training. These datasets are well-curated sources, providing comprehensive annotations of orthosteric and allosteric binding sites, essential for accurate binding site prediction. After filtering for redundancy by retaining unique macromolecules, the final dataset consists of approximately 3000 protein structures.

To ensure data quality and relevance, a series of filtering steps were applied. Proteins with redundant macromolecular structures were combined, and only unique structures were retained. Ligands associated with binding sites were filtered based on molecular weight, with a minimum threshold of 150 Da, ensuring the exclusion of very small molecules such as simple ions that are less relevant for binding site prediction. Drug-likeness criteria, such as Lipinski's Rule of Five, were also applied to exclude ligands unlikely to exhibit pharmacological activity. These steps ensured the inclusion of biologically and chemically meaningful binding sites, increasing the dataset's utility for predicting potential drug targets.

The dataset includes a balanced distribution of binding site types, addressing the class imbalance challenges commonly faced in biological datasets. Residues were labeled into three categories: non-binding sites (0), allosteric sites (1), and orthosteric sites (2). Furthermore, to leverage the benefits of semi-supervised learning, only the known allosteric and orthosteric sites were used during model training, while non-binding sites were treated as unknown. This strategy aimed to improve model generalization by encouraging it to infer unknown regions based on learned patterns from annotated binding sites.

This curated and filtered dataset provides a robust foundation for developing predictive models, ensuring a balance between biological diversity and chemical relevance while minimizing noise from irrelevant or redundant data points.

3.2.2 Data Preprocessing

Several crucial steps were undertaken during data preprocessing to guarantee that the dataset was appropriate for both model training and evaluation.

- **Protein Chain Separation**: Proteins were divided into individual chains, considering each chain as an independent graph. This method enables the model to concentrate on the distinct structural and functional characteristics of each chain.

- **Atom Filtering**: Only carbon alpha (Cα) atoms were retained from the protein chains to simplify the graphs and preserve crucial structural details.

- **Node and Edge Feature Calculation**:

  - **Node Features**: Each node (Cα atom) was assessed for various attributes such as amino acid identity (e.g., ALA, ARG, ASN), tau, theta, phi, psi, and omega angles, solvent accessible surface area

    (SASA), and secondary structure components (SSE). These characteristics encompass the geometric and physicochemical attributes of the amino acids.

    - **Edge Features**: Edges were established among nodes (Cα atoms) depending on their close proximity in space. The characteristics of these edges involved the distance and cosine angle between linked atoms, offering insight into the spatial connections within the protein structure (Rao et al., 2020).

- **Labeling**: Nodes were categorized as either orthosteric, allosteric, or non-binding sites using annotations found in the CasBench dataset. Labeling is essential for predicting binding sites in supervised learning.

- **Normalization and Scaling**: To ensure that all features contribute equally to the model training process, the node and edge features were normalized and scaled appropriately.

The raw protein data was processed to create organized graphs with informative features, suitable for inputting into the GAT model. This methodical way of preparing data was crucial for understanding the intricate connections within protein chains, ultimately improving the model's accuracy in predicting binding sites.

3.3 Feature Extraction

3.3.1 Node Features

Node features are critical for representing the individual properties of amino acids within protein structures (Mataeimoghadam et al., 2020). The following features were extracted for each carbon alpha (Cα) atom in the protein chains:

- **Amino Acid Type (One Hot Encoding)**: The amino acid type (e.g., Alanine [ALA], Arginine [ARG]) is encoded as a one-hot vector. For example, if there are 20 standard amino acids, this feature will be a 20-dimensional vector with a single 1 indicating the amino acid type.

- **Dihedral Angles**: The backbone dihedral angles ϕ (phi), ψ (psi), ω (omega), τ (tau), and θ (theta) are calculated for each amino acid (Figure 3.1). These angles describe the geometric configuration of the protein backbone, and they are critical for capturing the 3D folding pattern of the protein.

  The angles are calculated as follows:

  ϕ (phi): The angle between the C'-N-CA-C' atoms.

  ψ (psi): The angle between the N-CA-C'-N atoms.

  ω (omega): The dihedral angle that describes the rotation around the peptide bond between C' and N.

  τ (tau): Torsion angle between the bonds around the Cα atoms.

  θ (theta): Another dihedral angle important for describing secondary structure.

These angles are extracted from the protein's 3D structure using the known Cartesian coordinates of the atoms, applying vector cross-product formulations to compute the angle between planes formed by three consecutive atoms.

Mathematically, dihedral angles can be computed using the following formula:

$$Angle = \arctan 2(b1 \cdot (b2 \times b3), b2 \cdot (b1 \times b3)) \tag{3.1}$$

where b_i are bond vectors.

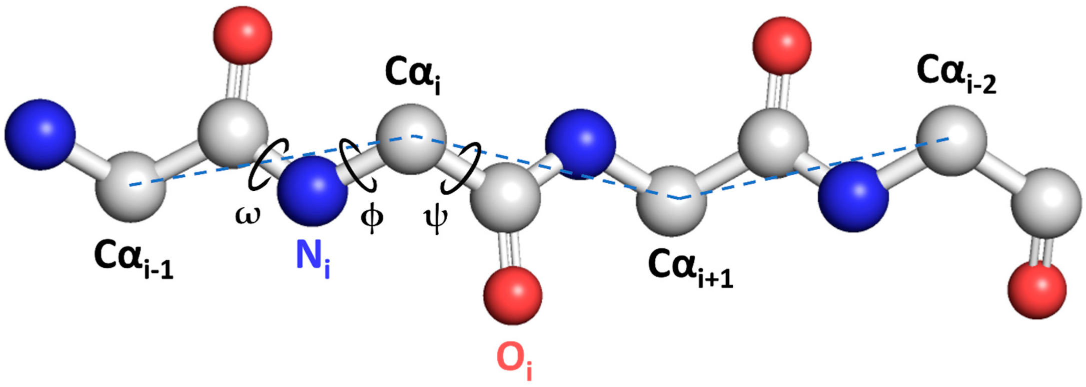

Figure 3.1: Representation of dihedral angles. While φ (phi), ψ (psi) and ω (omega) represent angles between dihedral angles of the peptide backbone; (τ) tau and θ (theta) represent dihedral angle between first 3 and 4 carbon alpha (Cα) atoms (Broz et al., 2023).

- **Solvent Accessible Surface Area (SASA)**: SASA measures how much of the amino acid is exposed to the surrounding solvent. This feature is critical for understanding which residues are likely to be part of binding sites. A high SASA value indicates that the residue is more exposed and likely accessible to ligands. SASA is calculated using a rolling ball algorithm over the molecular surface also called as Shrake–Rupley algorithm (Shrake et al., 1973).

- **Secondary Structure Elements (SSE)**: For each amino acid, its secondary structure element (e.g., α-helix, β-sheet, or coil) is encoded as a one-hot vector. These secondary structures are identified using the DSSP algorithm, which assigns each residue a secondary structure class based on the hydrogen bonding pattern.

- **Coordinates**: The 3D coordinates of the Cα atom are also used as features to capture spatial information. These coordinates are directly extracted from the protein's PDB file.

3.3.2 Edge Features

Edges in the graph represent interactions between amino acids based on their spatial proximity. The following features were extracted for each edge:

- **Euclidean Distance**: The Euclidean distance between two nodes i and j is calculated as:

$$d_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2} \tag{3.2}$$

  where x_i, y_i, z_i and x_j, y_j, z_j the 3D coordinates of the two nodes.

- **Cosine Angle**: The cosine of the angle formed by the vectors connecting the two nodes to their respective neighbors is calculated to capture the spatial orientation between them. This is computed using the dot product:

$$\cos(\theta)=\frac{v_i \cdot v_j}{|v_i| \cdot |v_j|} \tag{3.3}$$

  where v_i and v_j are vectors between the nodes.

3.3.3 Graph Construction

The nodes in the graphs were created by identifying Cα atoms, while edges were determined by spatial proximity:

- **Node Definition**: Every Cα atom within the protein sequence was depicted as a node in the graph, with characteristics obtained as detailed earlier.

- **Edge Definition**: Edges were generated depending on the spatial proximity of Cα atoms. A boundary distance was employed to decide if a connection should be created between two vertices. This specific limit was selected in order to achieve a balance between capturing important connections and not adding too much unnecessary interference.

- **Edge Index**: Calculation of the edge index matrix determined the topology of the graph by listing connected node pairs.

Through systematically extracting these characteristics and building the graphs, we successfully captured both the unique attributes of amino acids and their interactions in the protein structure. This thorough feature extraction procedure is essential for allowing the GAT model to successfully acquire knowledge and forecast protein binding locations.

3.4 Model Architecture

3.4.1 Graph Attention Network (GAT)

The GAT architecture integrates multi-head attention mechanisms to allow the model to focus on the most relevant nodes in the protein graph (Figure 3.2). This approach enhances the model's ability to capture intricate relationships between amino acids and potential binding sites.

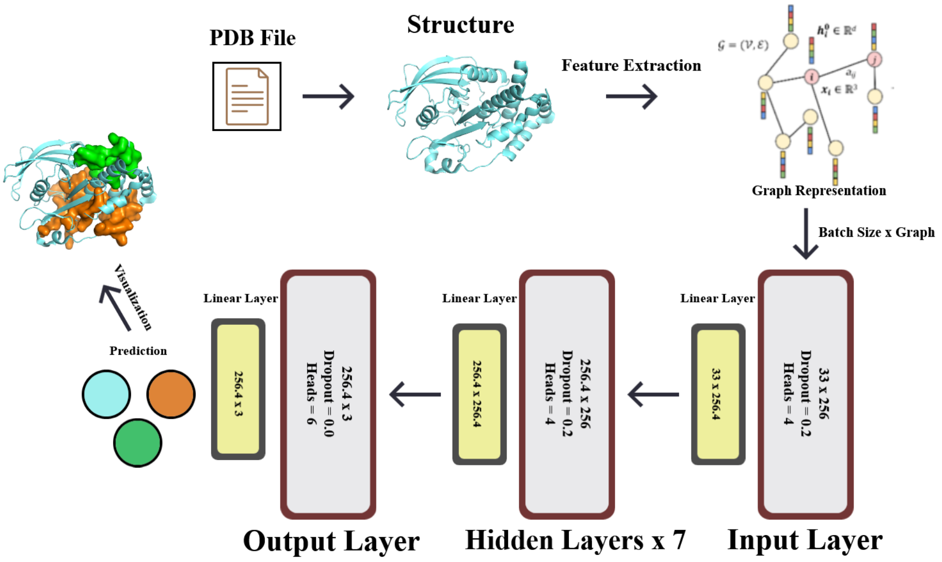

Figure 3.2: This flowchart illustrates the pipeline for protein binding site prediction using a Graph Neural Network (GNN)

- **Input Features**: The input consists of node features, edge indices, and edge attributes. Node features include structural and biochemical properties, while edge features encode distances and angles between nodes.

- **Attention Mechanism**: Each GAT layer computes attention coefficients between connected nodes, enabling the model to weigh the importance of each node's neighbors dynamically. This is achieved using the GATConv layer from PyTorch Geometric, which updates node features based on these attention weights (Figure 3.3). The attention coefficients are computed as:

$$a_{ij} = \frac{\exp(LeakyRelu(a^T[Wh_i]||[Wh_j]))}{\sum_{k \in N(i)} \exp(LeakyRelu(a^T[Wh_i]||[Wh_j]))} \tag{3.4}$$

where h_i is the feature vector of node i, W is a weight matrix, and a is a learnable vector that computes the importance of the edge between nodes i and j.

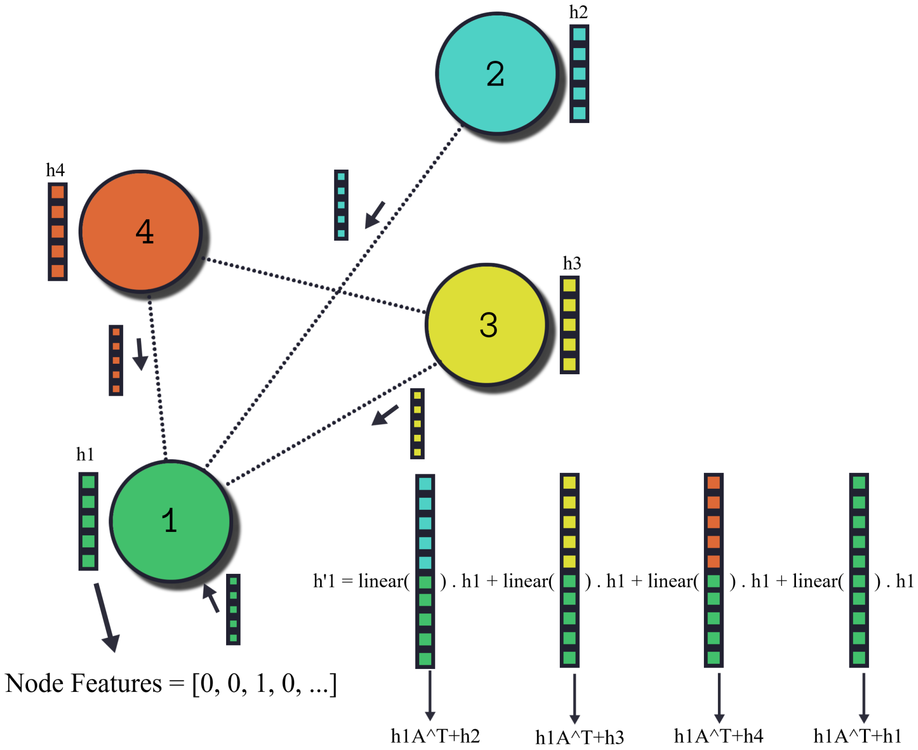
h2
2
h4
4
h3
3
h1
1
Node Features = [0, 0, 1, 0, ...]
h'1 = linear( ) . h1 + linear( ) . h1 + linear( ) . h1 + linear( ) . h1
h1A^T+h2
h1A^T+h3
h1A^T+h4
h1A^T+h1

Figure 3.3: This diagram illustrates the message-passing process in a Graph Attention Network (GAT)

- **Architecture Design**: The model contains multiple attention layers: Five Attention Layers with hidden dimensions of 256, Dropout Regularization to mitigate overfitting and enhance generalization, LeakyReLU Activation (Figure 3.4) to introduce non-linearities and softmax function to produce a probability distribution over the binding site classes on the output layer.

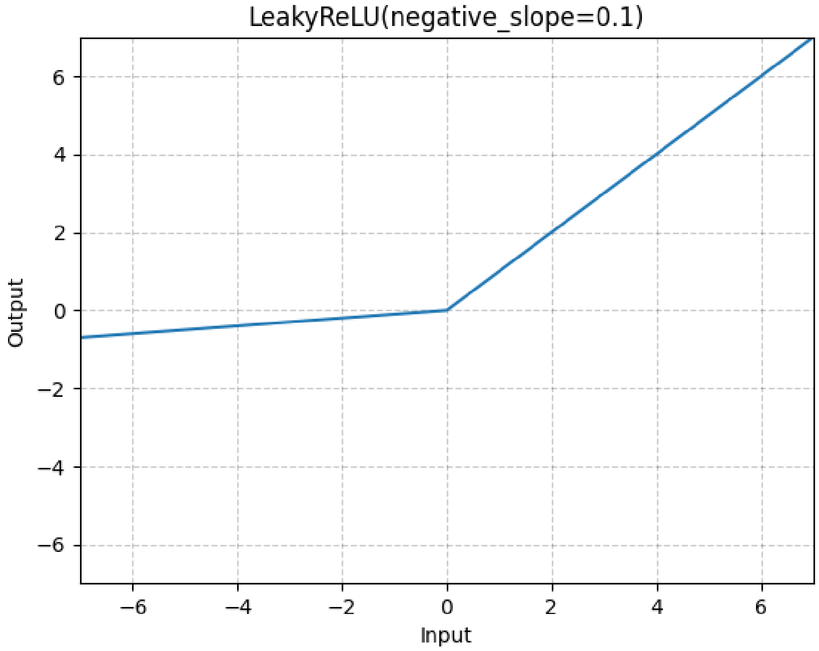

Figure 3.4: The plot illustrates the behavior of the Leaky ReLU activation function with a negative slope of 0.1. Unlike the standard ReLU, which outputs zero for all negative input values, Leaky ReLU allows a small negative output proportional to the input, ensuring a non-zero gradient even when the input is negative. This characteristic helps mitigate issues related to the "dying ReLU" problem, where neurons can become inactive during training.

LeakyReLU defined as:

$$LeakyReLU(x) = max(0, x) + \alpha . min(0, x) \tag{3.5}$$

Here, α (alpha) is a small positive constant such as 0.01, which allows a small gradient when the input x is less than or equal to zero. This helps in keeping the gradient flow alive during the training of deep neural networks, to avoid the dying ReLU problem where neurons permanently output zeros.

- **Output**: The final layer produces probabilities for each node belonging to either allosteric or orthosteric binding sites. The attention weights also provide insights into the model's decision-making process, aiding in explainability

Given a vector z of raw class scores from the final layer of a model, the softmax function is defined for each element of z_i the vector as:

$$Softmax(z_i)=\frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{3.6}$$

Where:

e is the base of the natural logarithm.

z_i is the score for class i.

n is the total number of classes in the vector z.

The denominator is the sum of the exponential scores for all classes, which normalizes the output to be a probability distribution.

3.4.2 Graph Convolutional Network (GCN)

The GCN architecture is designed to learn node representations by aggregating features from neighboring nodes in the protein graph. The model leverages GCN2Conv layers, which improve upon traditional GCNs by incorporating residual connections and enhanced message-passing capabilities.

- **Input Features**: Node features include amino acid properties, while edge indices define the graph's connectivity.

- **Layer Configuration**: Initial Linear Layer projects input node features to a hidden dimension. Multiple GCN2Conv Layers perform message-passing and feature aggregation with residual connections to stabilize learning. LeakyReLU Activation and Dropout are applied between layers to improve robustness.

- **Residual Connections**: Each layer adds the original node features to the transformed features, facilitating deeper network architectures without vanishing gradient issues.

- **Output**: The final linear layer maps the hidden representations to probabilities for binding site classification (allosteric or orthosteric). This design makes the GCN effective for capturing local structural patterns within proteins.

3.4.3 Equivariant Graph Neural Network (EGNN)

The Equivariant Graph Neural Network (EGNN) employed in this study follows a design that maintains E(3) equivariance ensuring that the model's predictions remain consistent under geometric transformations like rotations and translations (Figure 3.5). The architecture comprises a sequence of Equivariant Graph Convolutional Layers (EGCL), which process both the node features and spatial coordinates simultaneously.

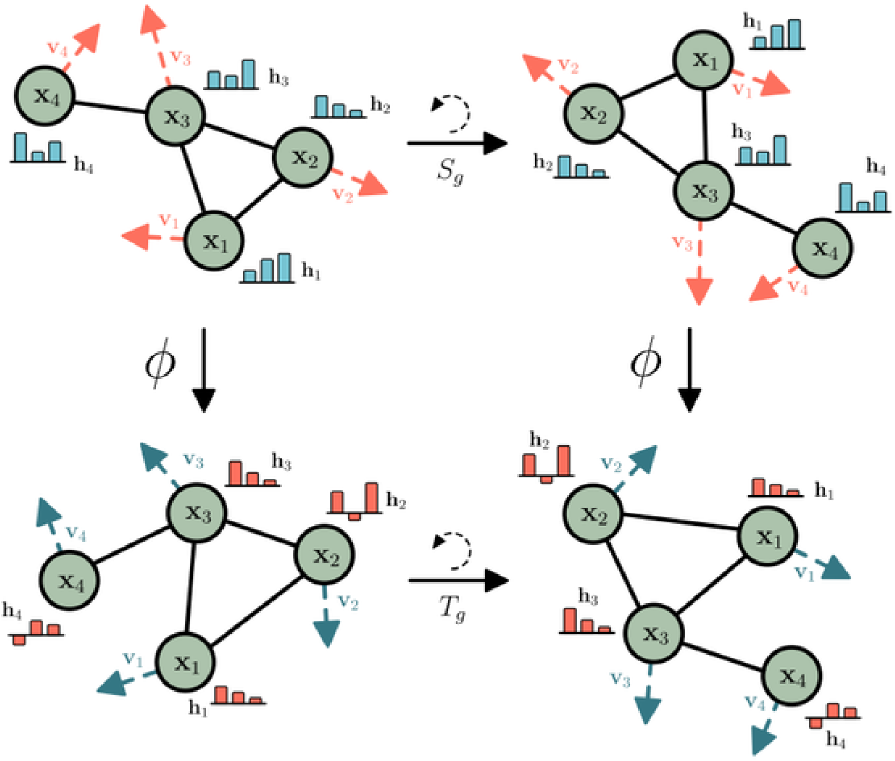

Figure 3.5: This figure illustrates the concept of rotation equivariance in a graph neural network denoted as φ (Satorras et al., 2021).

- **Input Features**: The input includes node features representing amino acid properties and edge features representing the relationships between nodes, such as distances and angles. Additionally, the model takes 3D coordinates of nodes within the protein structure.

- **Layer Configuration**: Each EGCL consists of three main components:

  - An Edge Multilayer Perceptron (MLP) that processes node and edge features to compute interactions.

  - A Node MLP that updates node features based on aggregated edge information.

    - A Coordinate MLP that updates node coordinates using edge features to preserve geometric information.

- **Prediction Outputs**: The EGNN produces:

    - **Updated Node Features**: Indicating potential binding site classifications (allosteric or orthosteric).

    - **Predicted Ligand Coordinates**: Dynamic positions for possible ligand interactions, which can vary in number based on the complexity of the ligand structure. Since the data is not embedded as a ligand-pocket pair, it does not directly give the ligand shape but provides information about the possible ligand region. Mean Squared Error (MSE) is used for calculating loss between coordinates. It defined as:

    $$MSE=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{3.7}$$

    Where:

    N = Number of data points

    y_i = Actual value (ground truth) for the i-th observation

    $\hat{y}_i$ = Predicted value for the i-th observation

This model is particularly effective for handling complex spatial data inherent to protein structures, making it suitable for predicting potential ligand positions rather than strict binding sites.

3.5 Training Process

3.5.1 Data Splitting

The dataset was divided into training, validation, and test sets to ensure robust evaluation of the model's performance as 8:1:1 ratio.

3.5.2 Training Configuration

In order to maintain a balance between computational efficiency and training stability, a batch size of 64 was utilized. The initial learning rate was set to 0.01 and was adjusted using a MultiStepLR learning rate scheduler at 75, 150, and 225 epochs with a gamma of 0.5. The Adam optimizer was employed for its ability to effectively manage sparse gradients and ensure a stable training process. To avoid overfitting, training would stop if the validation F1 score did not improve for 10 consecutive epochs. The model was trained using the binary cross-entropy with logits loss function, which is suitable for multi-class classification tasks (Goodfellow et al., 2018) for 300 epochs, with the best-performing model was saved according to validation performance. This loss function defined as:

$$BCEWithLogitsLoss(y, \hat{y}) = max(\hat{y}, 0) - \hat{y} \cdot y + \log(1 + e^{-|\hat{y}|}) \qquad (3.8)$$

Where y is the ground truth label (0 or 1), $\hat{y}$ is the predicted logit (a real-valued number, before applying the sigmoid function).

When dealing with a batch of data, the loss is typically averaged over all instances in the batch. The formula for the batch loss is:

$$BCEWithLogitsLoss(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N} max(\hat{y}_i, 0) - \hat{y}_i \cdot y_i + \log 1 + e^{-|\hat{y}_i|} \quad (3.9)$$

Where: Y is the vector of ground truth labels, $\hat{Y}$ is the vector of predicted logits, N is the number of instances in the batch. This formulation helps avoid numerical instability issues that can arise when directly computing the binary cross entropy on logits.

3.5.3 Evaluation Metrics

The model's performance was evaluated using several metrics, reflecting different aspects of its predictive power:

- **F1 Score**: Assesses the balance between precision and recall.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3.10)$$

- **Precision**: Measures the accuracy of positive predictions.

$$Precision = \frac{TP}{TP + FP} \quad (3.11)$$

- **Recall**: Evaluates the model's ability to identify all relevant instances.

$$Recall = \frac{TP}{TP + FN} \quad (3.12)$$

- **AUC-ROC**: Analyzes the trade-off between true positive and false positive rates.

- **Brier Score**: Measures the accuracy of probabilistic predictions. It evaluates the mean squared difference between predicted probabilities and the actual outcomes.

$$Brier\ Score = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \tag{3.13}$$

Where:

N = Number of predictions

f_i = Predicted probability for the i-th event (e.g., probability of a positive outcome)

o_i = Actual outcome for the i-th event (1 if the event happened, 0 otherwise)

- **Logarithmic Loss**: Measures the performance of a classification model where the prediction is a probability value between 0 and 1.

$$\log Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{3.14}$$

Where:

N = Number of predictions

y_i = Actual outcome (1 for positive class, 0 for negative class) for the i-th instance

p_i = Predicted probability that the i-th instance belongs to the positive class (must be between 0 and 1)

- **Task Loss**: Specific losses for orthosteric and allosteric site predictions were calculated to assess the model's ability to differentiate between these classes.

3.6 Summary

EGNN focuses on maintaining geometric consistency and is well-suited for predicting potential ligand positions due to its equivariant properties. GAT leverages attention mechanisms to dynamically weigh neighbor contributions, making it ideal for identifying critical nodes in complex graphs. GCN uses efficient message-passing and residual connections to capture structural information, excelling in tasks requiring aggregated local features.

Table 3.1: Summary of models

Model	**Setting**	**Loss Function**	**Objective**
GCN	Supervised	Graph Reconstruction, Contrastive Loss	Learn node embeddings
GCN	Semi-Supervised	Cross-Entropy + Unsupervised Regularization	Predict node labels for known sites
GAT	Supervised	Graph Reconstruction, Contrastive Loss	Capture attention-based embeddings
GAT	Semi-Supervised	Cross-Entropy + Unsupervised Regularization	Predict node labels while leveraging attention
EGNN	Semi-Supervised	Binary Cross-Entropy + L2 (MSE) for Coordinates	Predict allosteric sites and relative ligand coords

These architectures complement each other by addressing different aspects of protein binding site prediction (Table 3.1).

Chapter 4

Results and Discussion

4.1 Overview of Results

This section discusses the outcomes of various graph neural network (GNN) models —GAT, GCN, and EGNN—for predicting allosteric and orthosteric binding sites. The models were evaluated using supervised and semi-supervised settings to determine the most effective approach for accurate binding site detection. The datasets used include a combined version of CasBench and BioLip, comprising approximately 6000 protein samples.

4.2 Performance Metrics and Comparisons

The performance metrics—F1 score, AUC (Area Under the Curve), and Task Loss— were plotted to visualize the learning curves and stability of each model. The test metrics for each model are illustrated in Figure 4.1.

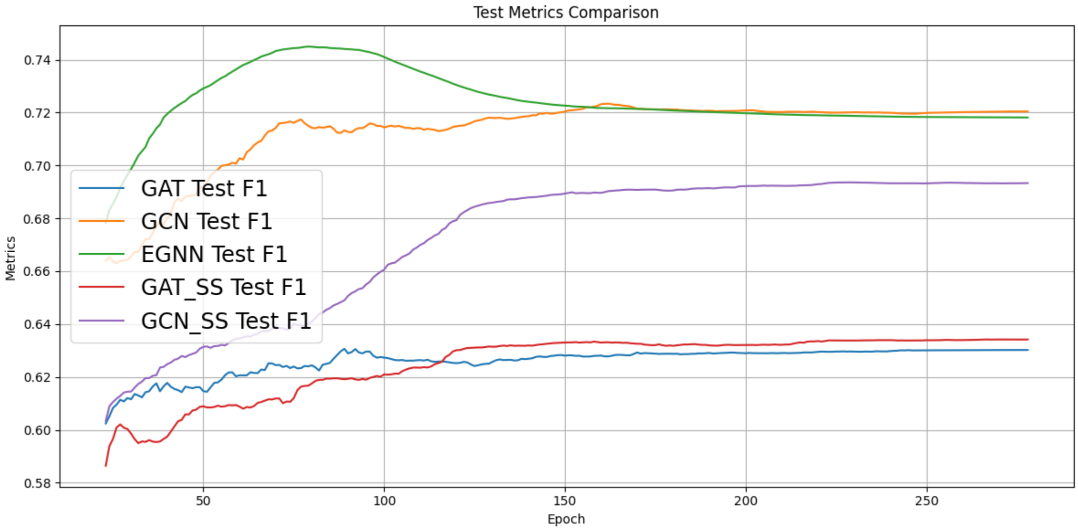

Figure 4.1: F1 metrics comparison for each model

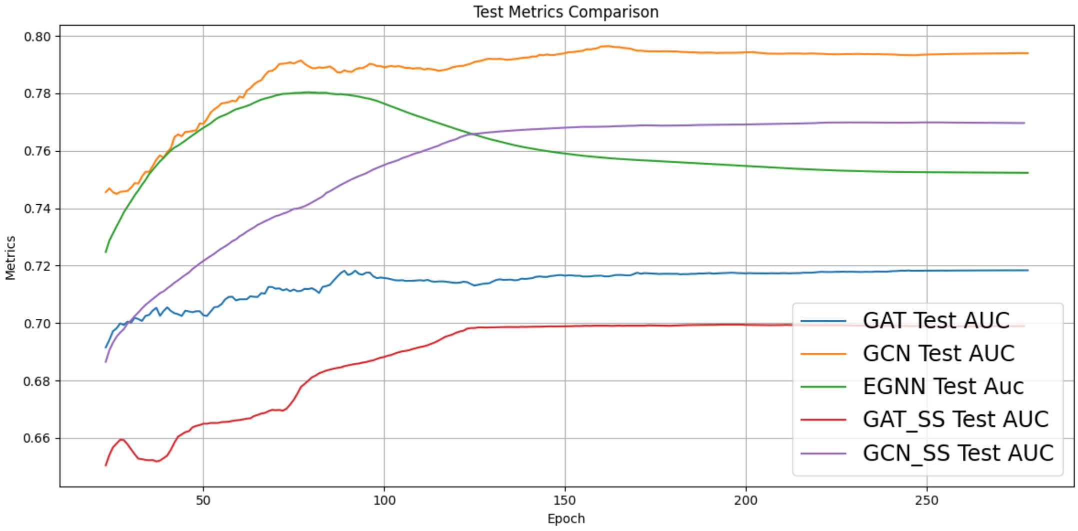

Figure 4.2: AUC - ROC metrics comparison for each model

GAT (Graph Attention Network) models demonstrated stable performance across epochs, achieving relatively high F1 scores and AUC values. In both supervised and semi-supervised settings, GAT exhibited consistent learning and minimal fluctuations. GCN (Graph Convolutional Network) models displayed moderate performance with more variability compared to GAT. The semi-supervised GCN model showed slight improvements in AUC compared to the supervised setting, indicating the utility of leveraging unlabeled data. EGNN (Equivariant Graph Neural Network) showed high performance initially but plateaued rapidly. This behavior could be attributed to the model's reliance on structural information, which limits generalization when ligand molecules are small or less complex. The rapid convergence of Task Loss for the EGNN model suggests that the model is learning to differentiate between binding sites efficiently. However, it does not always generalize well for smaller ligands like ions, which may impact the applicability in certain biological contexts.

4.3 Importance Analysis

4.3.1 EGNN Feature and Edge Importance

The Equivariant Graph Neural Network (EGNN) demonstrates significant insights into the relative importance of different feature groups for predicting binding sites. The analysis is supported by two key visualizations: the overall feature importance (Figure 4.3) and the average saliency map across nodes (Figure 4.4).

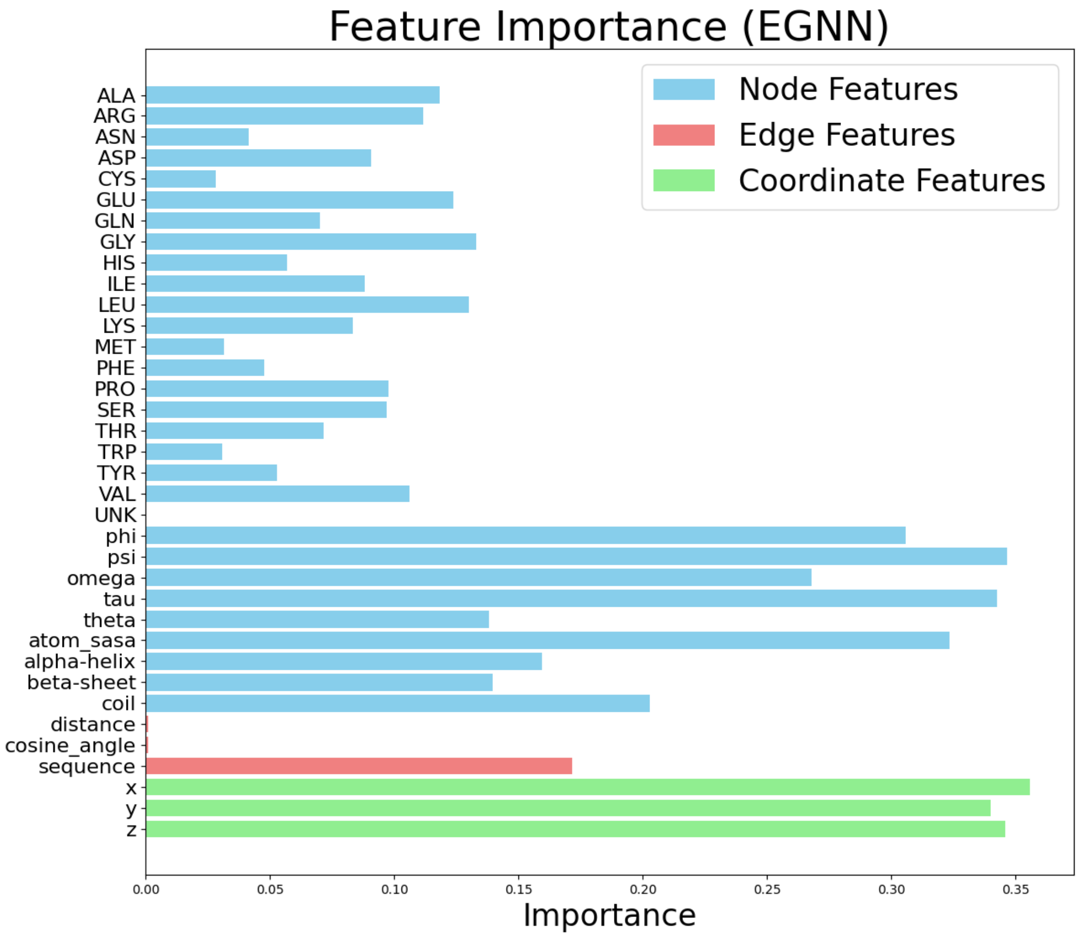

Figure 4.3: Feature importance of EGNN

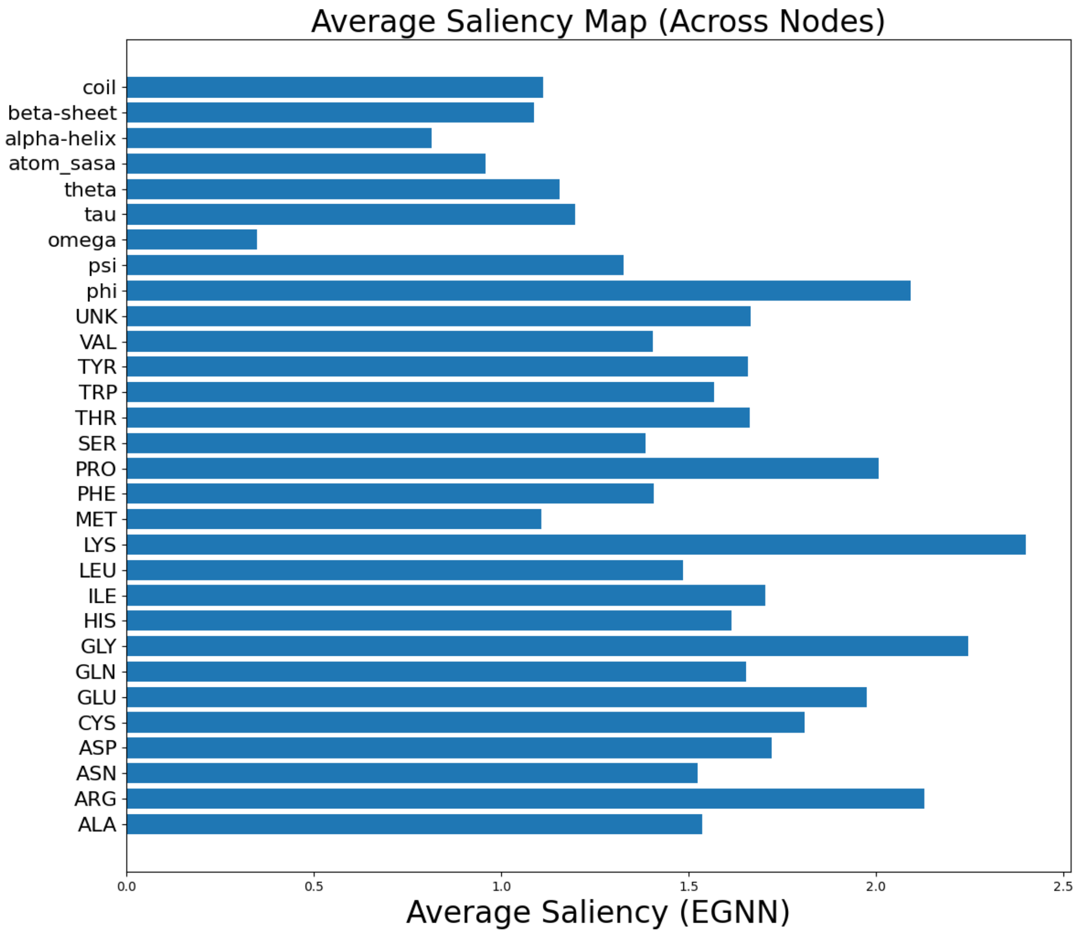

Figure 4.4: Saliency map of EGNN

Backbone dihedral angles such as phi (ϕ), psi (ψ), and omega (ω) emerge as the most critical features among node attributes, showcasing their role in capturing the geometric configuration of the protein backbone. These angles are pivotal in describing secondary structure and folding patterns. Solvent-accessible surface area (SASA) and secondary structure elements (α-helices, β-sheets, and coils) also play a substantial role, as they relate directly to a residue's likelihood of being part of a binding site. Amino acid types like arginine (ARG), lysine (LYS), and histidine (HIS) exhibit higher importance, which aligns with their known involvement in protein-ligand interactions due to their charged or polar side chains. Among the edge attributes, cosine angles contribute the most to binding site prediction, highlighting the importance of relative orientation between connected nodes. This finding reinforces the significance of geometric interactions within the protein structure.

Sequence-based connectivity exhibits moderate importance, reflecting the need to incorporate relational data alongside spatial information. The 3D spatial coordinates (x, y, z) demonstrate a balanced and substantial contribution to model performance, confirming the necessity of geometric awareness for identifying binding pockets.

The saliency analysis (Figure 4.4) reinforces the importance of dihedral angles and SASA, as well as the role of specific amino acids in binding site prediction. Residues such as arginine (ARG), lysine (LYS), and leucine (LEU) consistently exhibit higher saliency, emphasizing their functional relevance in protein-ligand interactions.

Overall, the EGNN model leverages a diverse set of node, edge, and coordinate features to achieve robust binding site predictions. Its ability to integrate geometric consistency with structural attributes makes it particularly well-suited for capturing the complexities of protein-ligand interactions.

4.3.2 GAT and GAT (Semi-Supervised) Feature Importance

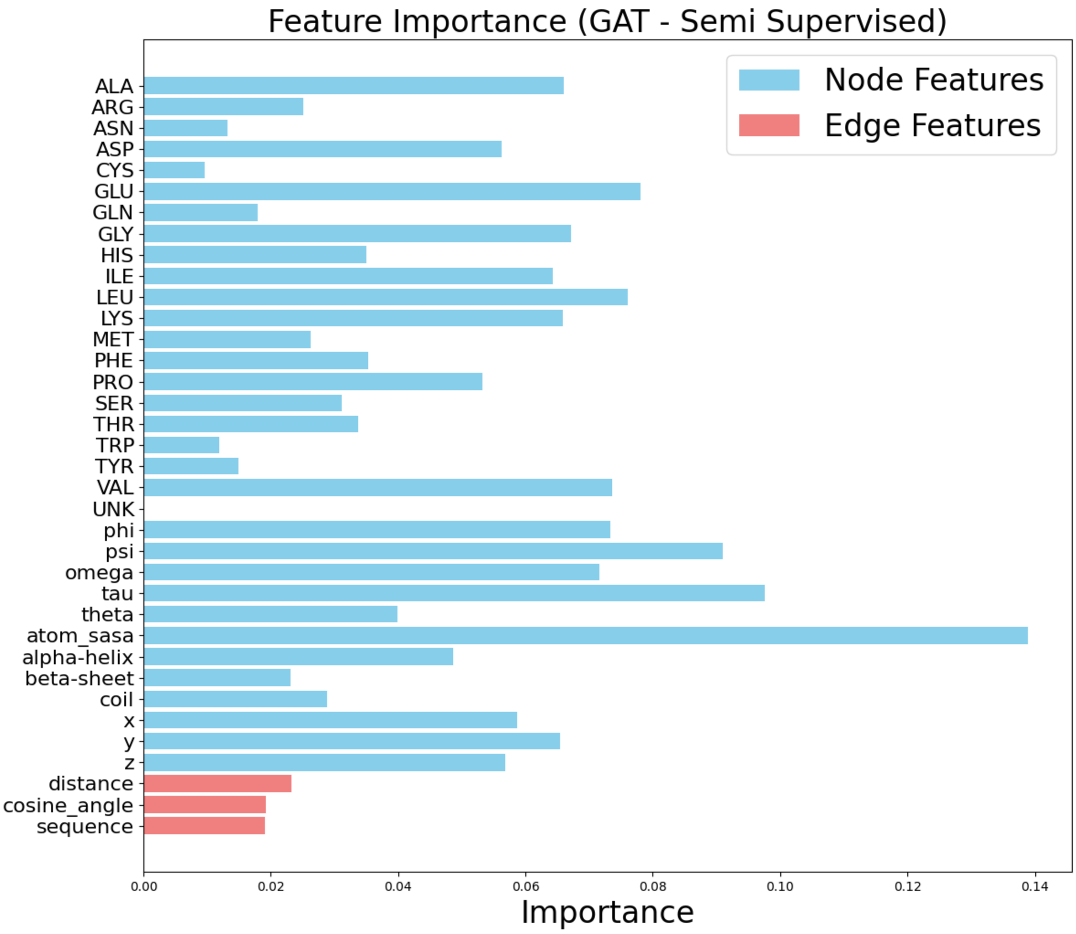

Figure 4.5: Feature importance of GAT (semi - supervised)

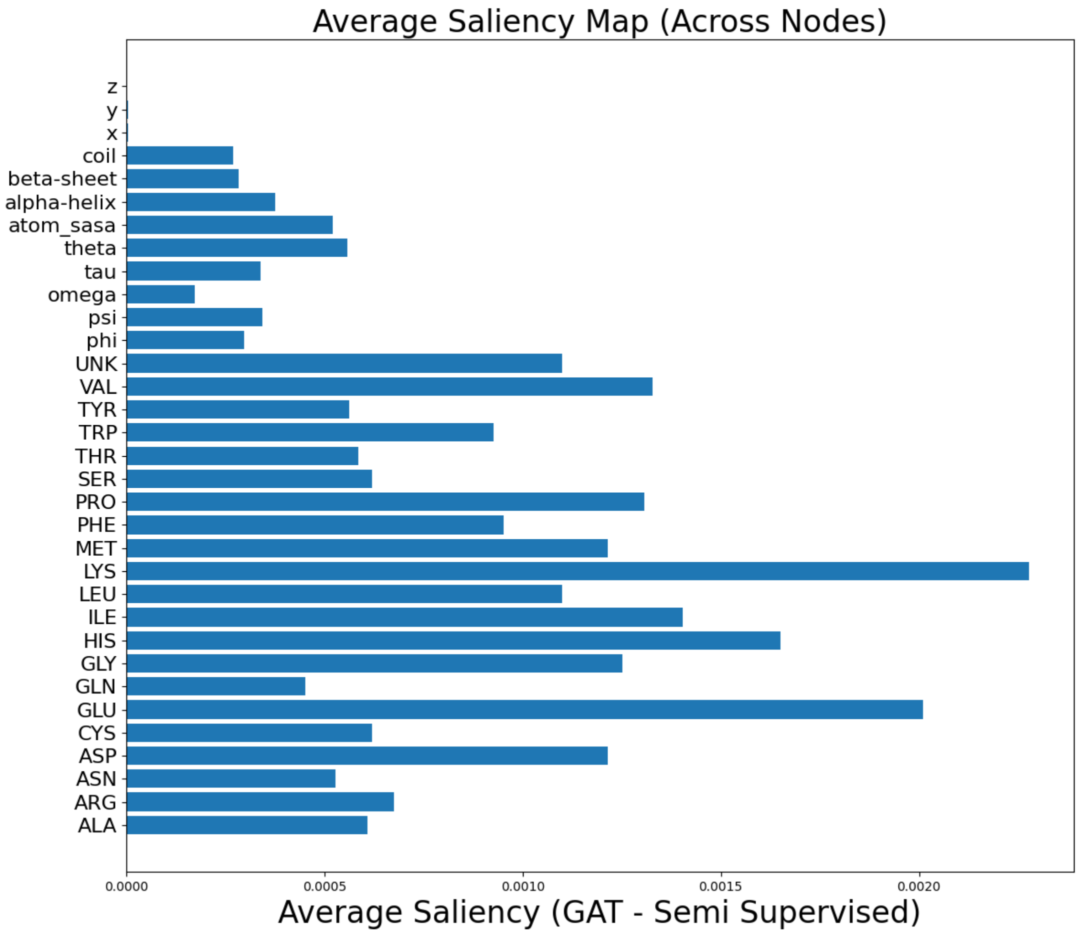

Figure 4.6: Saliency of GAT (semi - supervised)

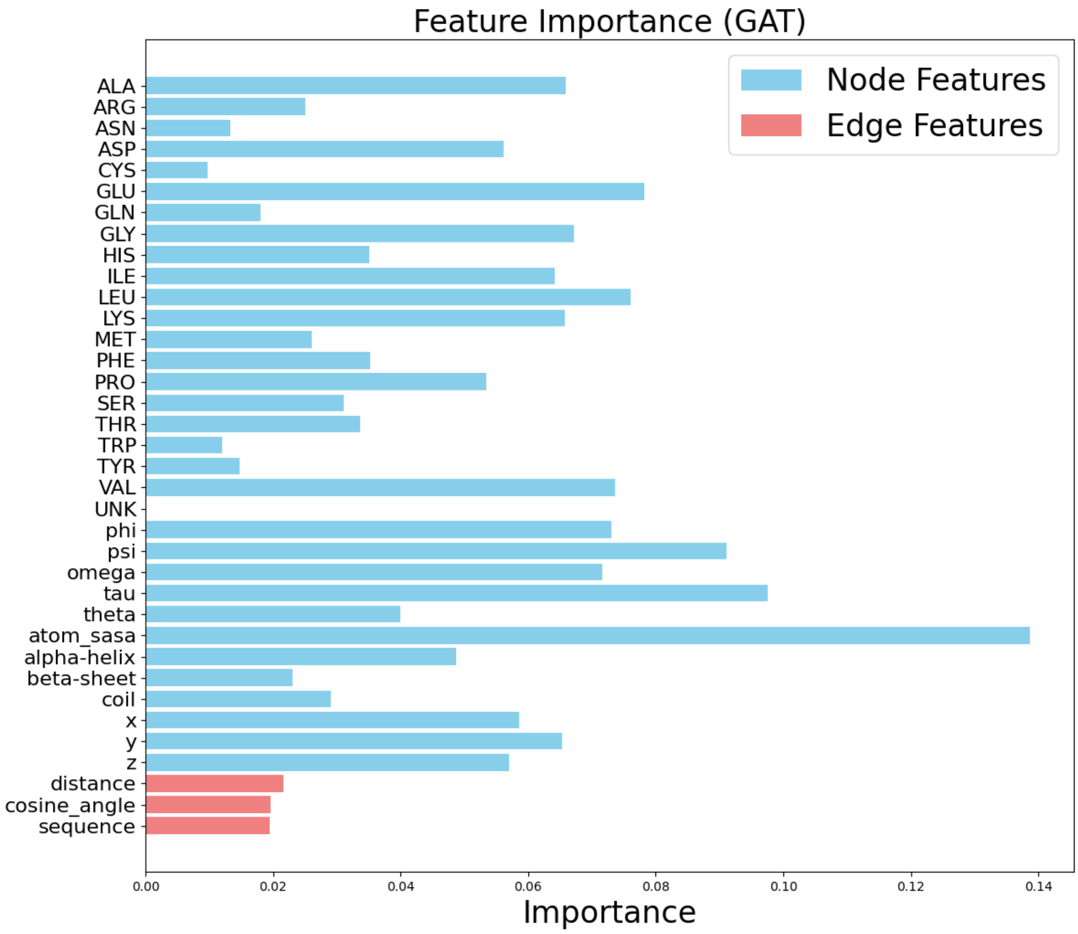

Figure 4.7: Feature importance of GAT

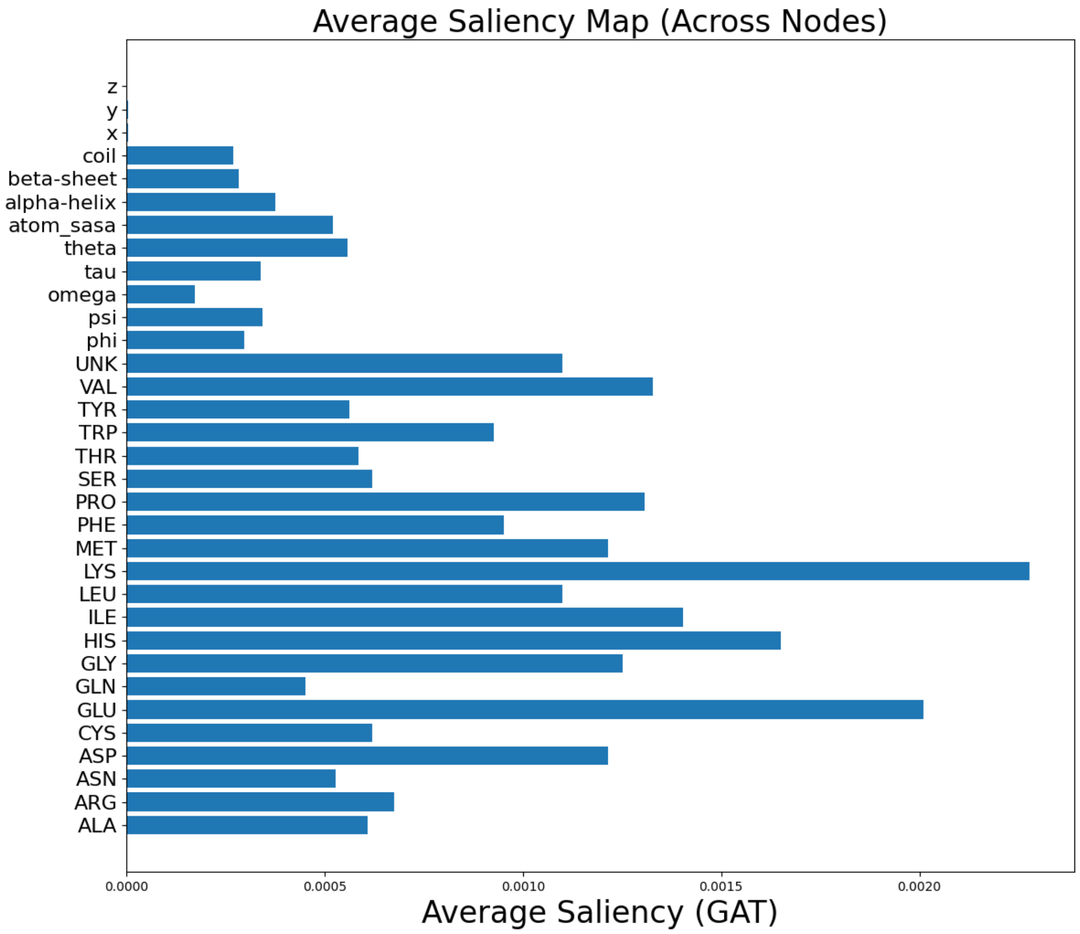

Figure 4.8: Saliency map of GAT

The feature importance analysis for the Graph Attention Network (GAT) model reveals significant insights into how the model utilizes various features to predict binding sites (Figure 4.5, Figure 4.7). Among the node features, lysine (LYS) and leucine (LEU) demonstrate high importance, reflecting their role in stabilizing protein structures and contributing to binding interactions. Backbone dihedral angles such as phi (ϕ), psi (ψ), and tau (τ) emerge as the most critical features, indicating their importance in capturing the geometric folding patterns of proteins, which directly influence the identification of binding sites. Solvent Accessible Surface Area (SASA) and secondary structural elements, particularly alpha-helices and beta-sheets, also show moderate importance, highlighting their relevance in describing residue exposure and structural arrangement.

For the edge features, cosine angle and sequence distance stand out as significant contributors, underscoring the importance of spatial orientation and topological relationships between residues. In contrast, Euclidean distance is found to have comparatively lower importance, suggesting that relative orientation and connectivity are more defining factors for binding site prediction than absolute distances.

The saliency map analysis further provides detailed insights into the average contribution of node features across predictions (Figure 4.6, Figure 4.8). Residues like glycine (GLY), lysine (LYS), and arginine (ARG) are particularly salient, aligning with their biochemical properties, such as flexibility and the ability to form hydrogen bonds or ionic interactions. Hydrophobic residues such as leucine (LEU) and valine (VAL) also show notable saliency, indicating their potential role in forming hydrophobic pockets. Structural features, particularly the backbone dihedral angles phi, psi, and tau, emerge as the most salient, emphasizing the role of protein geometry in binding site identification. Among secondary structural elements, beta-sheets exhibit relatively lower saliency compared to alpha-helices, potentially due to their less dynamic role in forming binding pockets.

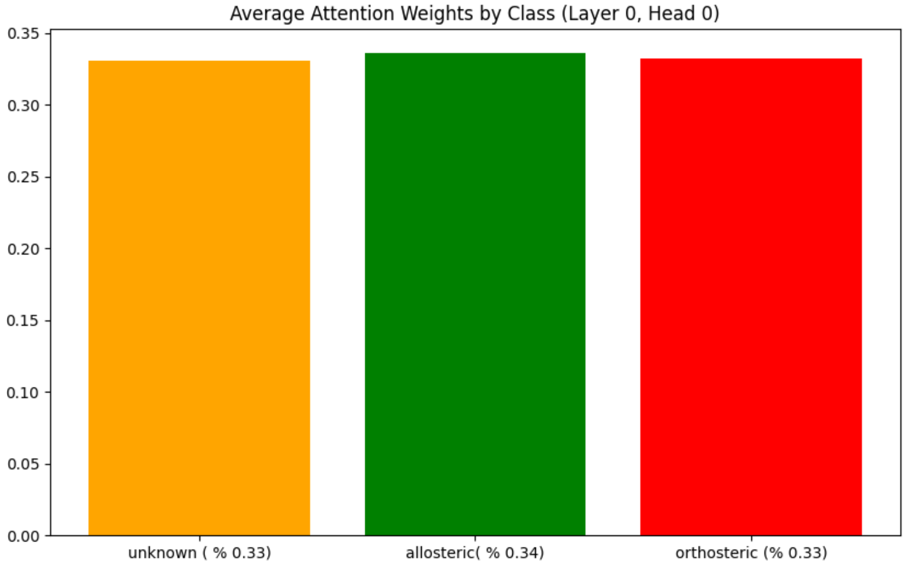

Figure 4.9: Average attention weights for GAT model

This balanced distribution of attention weights (Figure 4.9) indicate the model's ability to capture relevant interactions within the protein graph structure.

Overall, the GAT model effectively captures geometric and structural details through its reliance on node features like dihedral angles and SASA, while edge features like sequence distance and cosine angle underscore the importance of relational data. The balanced attention weights suggest the need for further optimization to enhance the model's ability to differentiate more clearly between orthosteric and allosteric sites.

4.3.3 GCN and GCN (Semi-Supervised) Feature Importance

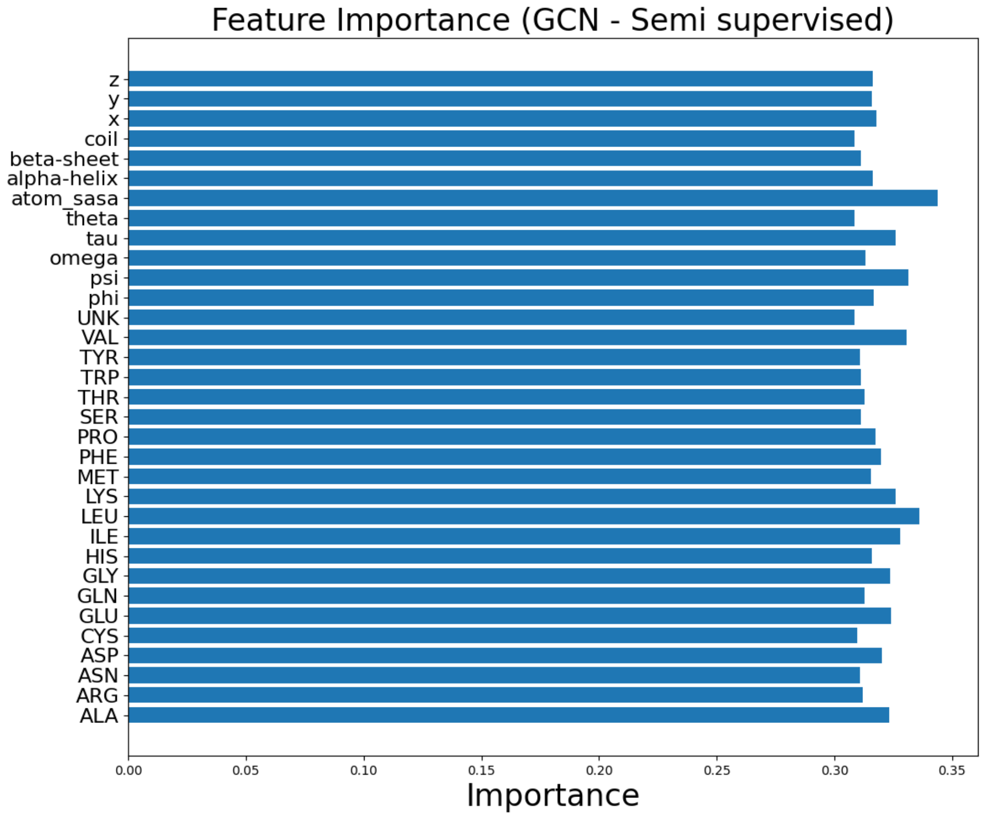

Figure 4.10: Feature importance of GCN (semi - supervised)

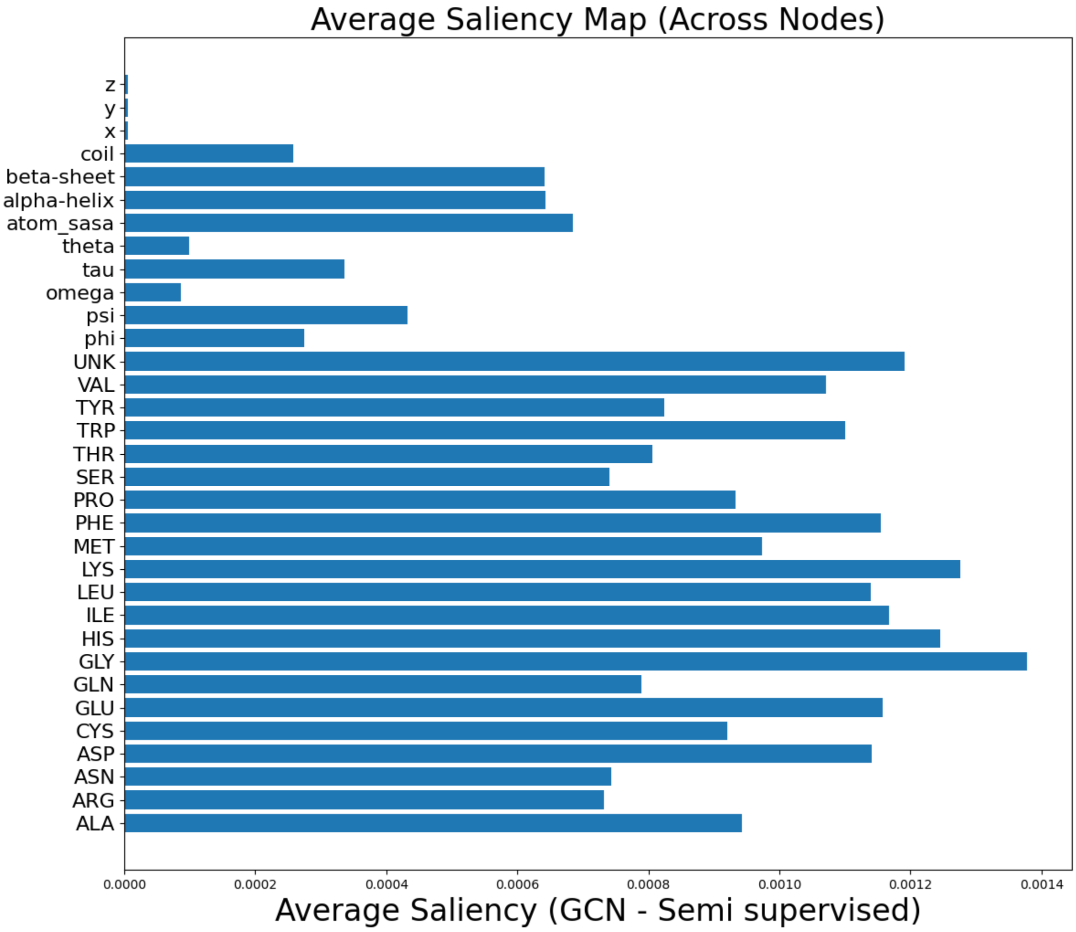

Figure 4.11: Saliency map of GCN (semi - supervised)

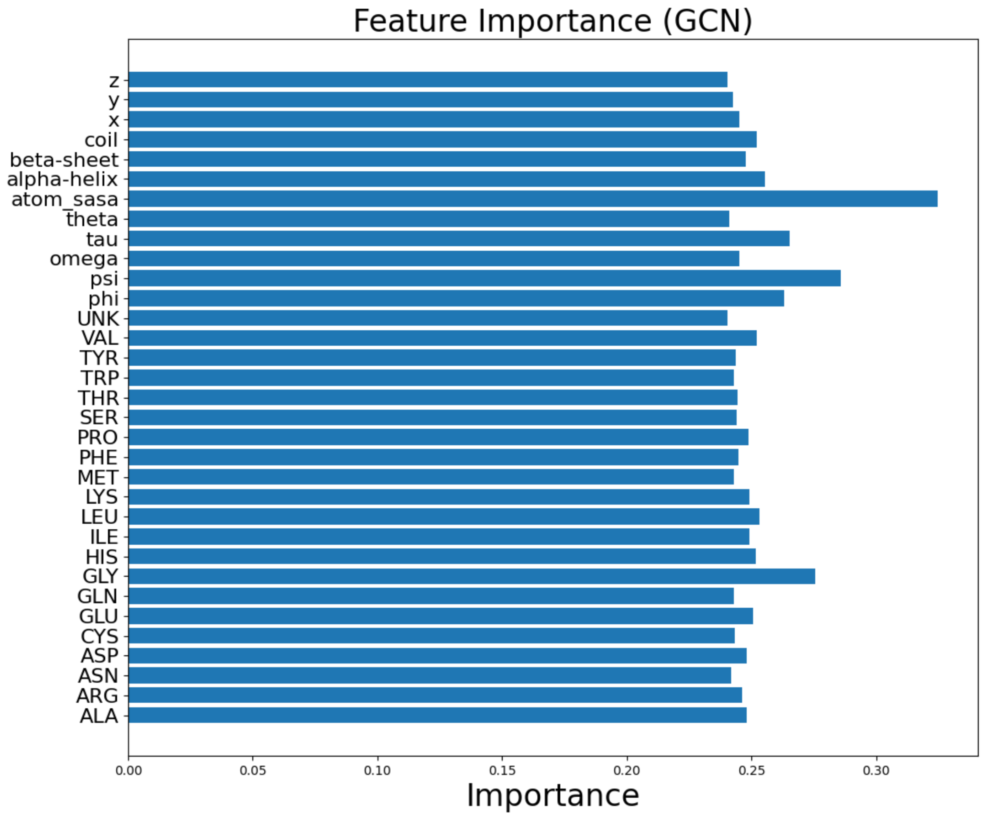

Figure 4.12: Feature importance of GCN

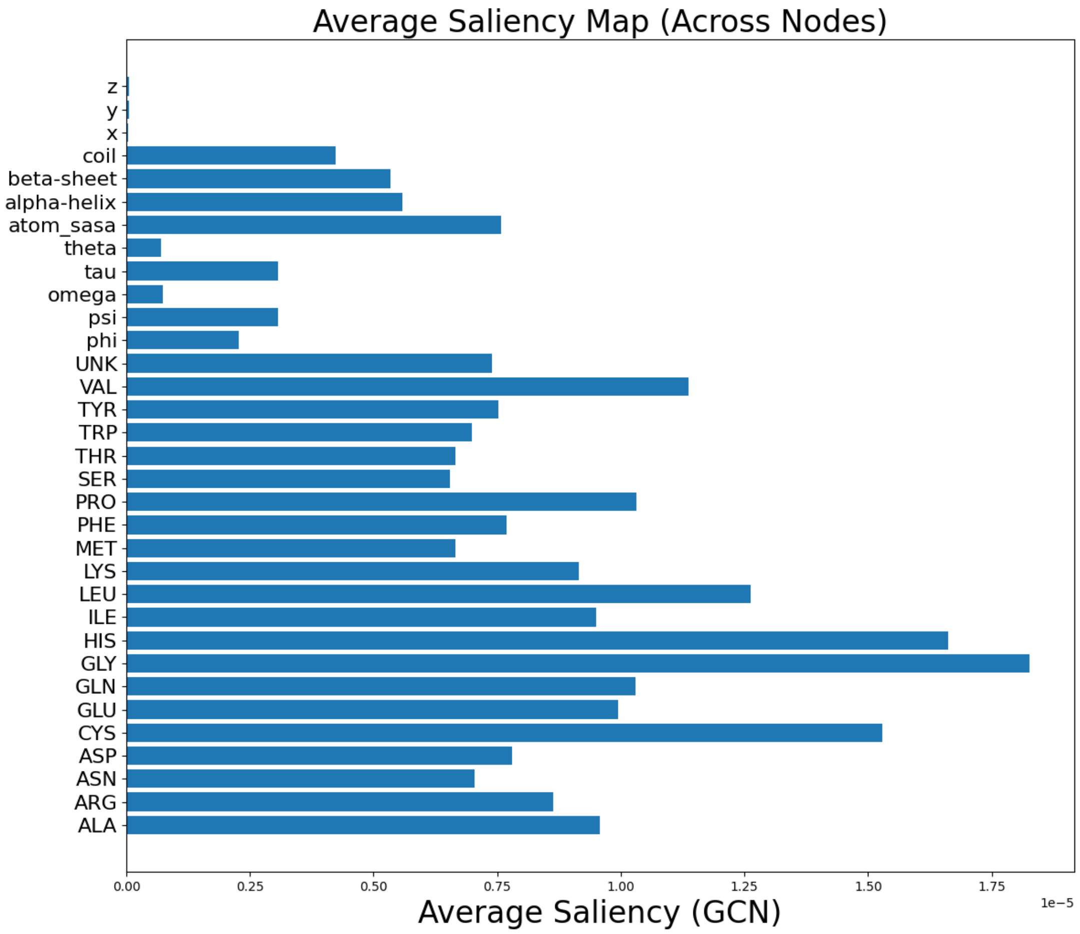

Figure 4.13: Saliency map of GCN

The Graph Convolutional Network (GCN) model demonstrates unique characteristics in its utilization of features for predicting protein binding sites. Node features, particularly those representing structural and physicochemical properties, exhibit balanced importance across various attributes. Unlike models with dynamic mechanisms like attention, GCN relies heavily on the global structural context provided by node and edge features (Figure 4.10, Figure 4.12).

In the context of node features, amino acid-specific properties such as solvent-accessible surface area (SASA), secondary structure elements (e.g., alpha-helix and beta-sheet), and dihedral angles (phi, psi, and omega) are consistently influential. SASA stands out as the most impactful node feature, likely due to its relevance in identifying accessible regions for binding. Among amino acid-specific features, residues like lysine (LYS), histidine (HIS), and glutamate (GLU) demonstrate higher importance. These residues often play critical roles in protein-ligand interactions through electrostatic and hydrogen-bonding mechanisms.

Edge features, particularly those capturing geometric relationships such as cosine angles and distances, are critical for understanding spatial connectivity. The reliance on these features indicates that the GCN model effectively captures the local and global topologies of protein graphs. However, the uniform importance of edge features suggests a limitation in dynamically prioritizing certain interactions over others, unlike attention-based models.

The saliency map analysis provides additional insights, highlighting the significance of specific node features (Figure 4.11, Figure 4.13). Structural features such as SASA and secondary elements continue to dominate, while residues like lysine (LYS), glycine (GLY), and histidine (HIS) emerge as critical for binding site prediction. The saliency values reveal that GCN effectively captures relevant patterns across the protein graph, emphasizing residues with higher reactivity and structural accessibility.

The semi-supervised GCN variant shows a similar trend but with broader feature utilization (Figure 4.10). This indicates the model's ability to generalize better by incorporating unlabeled data. The feature importance remains balanced, with structural and geometric features contributing equally. This balance highlights the semi-supervised model's robustness in capturing both local and global graph properties, enhancing its predictive accuracy.

In summary, the GCN model relies on a well-distributed feature set, emphasizing structural and geometric properties. While its reliance on global structural features ensures stability, the lack of dynamic prioritization of specific interactions may limit its adaptability to complex binding scenarios. The semi-supervised variant addresses some of these limitations by leveraging additional unlabeled data, resulting in improved generalization and broader feature relevance.

4.3 Comparison Across Models

EGNN outperforms GAT and GCN in capturing spatial and geometric information, as evidenced by the higher importance of 3D coordinates (Figure 4.1, Figure 4.2). This ability to integrate spatial data ensures a detailed understanding of protein structures and potential ligand-binding regions. In contrast, GAT models excel with their attention mechanisms, which dynamically weigh node and edge contributions, providing flexibility in emphasizing the most relevant features. GCN models, while effective, demonstrate a more uniform distribution of feature importance, relying heavily on global structural features rather than dynamically prioritizing local interactions.

Semi-supervised approaches consistently exhibit broader feature utilization and improved generalization, as they effectively incorporate unlabeled data into the learning process. The semi-supervised GAT and GCN models display a more diverse set of important features compared to their supervised counterparts, enhancing their ability to generalize to new data. However, while semi-supervised learning appears advantageous in theory, practical results show mixed outcomes depending on the binding site type.

For allosteric sites, semi-supervised learning performs exceptionally well, as the broader distribution of predictions aligns with the diffuse nature of these sites. However, for orthosteric sites, which often require sharper and more localized predictions, supervised learning tends to perform better in certain cases. Semi-supervised methods, when predicting unknown site classes after training, sometimes spread predictions across the entire protein. While this approach is advantageous for capturing diffuse or non-obvious binding sites like allosteric regions, it can dilute the precision needed for orthosteric sites, where sharper, focused predictions are essential. This limitation suggests that a hybrid approach or careful tuning of semi-supervised methods might be necessary to optimize predictions across both types of binding sites.

4.4.4 Comparison with Previous Studies

In this section, we compare the proposed Graph Neural Network (GNN) models (GAT, GCN, and EGNN) for protein binding site prediction with related works in the field. While the specific tasks and datasets vary among these studies, they offer valuable insights into the capabilities and limitations of different graph-based models in computational biology.

Table 4.1 summarizes the performance metrics of different models from recent studies, focusing on protein binding site prediction tasks. The metrics include F1 Score, Precision, Recall, AUC-ROC, and specific model attributes.

Table 4.1: Comparison of Performance Metrics with Similar Studies

Study	Model/ Method	Dataset	Featues	Task	F1 Score	Precision	Recall	AUC-ROC	DCC	DC	MSE	Special Features
Zitnik et al. (2019)	GNN	Protein-Protein	Protein Features	Modeling Polypharmacy	N/A	N/A	N/A	0.87	N/A	N/A	N/A	GNNs, Interaction Prediction
Satorras et al. (2021)	EGNN	QM9	Atom Types, Charges, Bond Types	Molecular Property Prediction	N/A	N/A	N/A	N/A	N/A	N/A	0.07	Equivariant GNN, Structural Features
Sestak and et al (2024)	VN-EGNN	COACH420, HOLO4K, PDBbind2020	Atomic Coordinates, Virtual Nodes, Distance Metrics	Protein Binding Site Identification	N/A	N/A	N/A	N/A	0.61	0.75	N/A	Equivariant GNN, Structural Features
Smith et al. (2023)	GrASP	COACH420 (Mlig+), HOLO4K(Mlig+)	Atomic Coordinates, Distance Metrics	Identifying Druggable Binding Sites	N/A	0.71	0.91	N/A	N/A	N/A	N/A	Enhanced GAT, Instance and semantic segmentation
Abdollahi et al. (2023)	Nodecoder	BioLip	AlphaFold2 Predictions	Residue Characterization	N/A	N/A	N/A	0.75	N/A	N/A	N/A	AlphaFold Integration, GCN

Zitnik et al. (2019) utilized a GNN model for modeling polypharmacy side effects based on protein-protein interactions and drug information. Their task focuses on interaction prediction rather than binding site identification. Their high AUC-ROC (0.872) underscores the effectiveness of GNNs in capturing relational data, a concept that aligns with our study's goal of modeling intricate protein interactions. Satorras et al. (2021) introduced the Equivariant Graph Neural Network (EGNN) for molecular property prediction on the QM9 dataset. The EGNN's ability to maintain rotational and translational invariance (with an MSE of 0.071) highlights the importance of structural features, a strength that our study leverages for binding site prediction. Sestak et al. (2024) developed the VN-EGNN for protein binding site identification on datasets like COACH420, HOLO4K, and PDBbind2020. Their use of virtual nodes and distance metrics resulted in a DCC of 0.605 and DCA of 0.750, demonstrating the potential of equivariant models for capturing protein-ligand interactions. Our method similarly benefits from the geometric consistency of EGNNs but focuses on differentiating between orthosteric and allosteric sites. Smith et al. (2023) proposed GrASP, an enhanced GAT model, for identifying druggable binding sites. Their model achieved a Precision of 0.71 and a Recall of 0.914 on datasets such as COACH420(Mlig+) and HOLO4K(Mlig+). This highlights the effectiveness of attention-based models in identifying critical regions within protein structures, which parallels our use of GAT for dynamic node-weighting. Abdollahi et al. (2023) introduced Nodecoder, a GCN-based model that integrates AlphaFold2 predictions and residue/atom features for characterizing residues in protein structures. Their model achieved an AUC-ROC of 0.754, illustrating the benefit of integrating structure prediction tools with GNNs. Our approach similarly employs GCNs for capturing local structural features but extends the model's capability by combining it with EGNN and GAT architectures for a more comprehensive prediction.

While the related works offer valuable insights, our study presents several unique contributions:

We employ GAT, GCN, and EGNN collectively, leveraging the strengths of each model to predict both orthosteric and allosteric binding sites with high accuracy. Our dataset combines CasBench (2023) and BioLiP, providing a diverse set of annotations and addressing class imbalance challenges. We incorporate a comprehensive set of node and edge features, including dihedral angles, solvent-accessible surface area (SASA), and secondary structure elements (SSE), enhancing the models' ability to capture intricate protein characteristics. By employing attention mechanisms (GAT) and saliency maps, we improve the interpretability of the model's predictions, aiding in the identification of critical residues involved in binding. The EGNN model ensures that predictions remain consistent under geometric transformations, making it well-suited for handling the spatial complexity of protein structures.

4.4.5 Discussion of Case Studies

To better understand the performance of our models, we present two case studies where the EGNN model is used to predict binding sites and potential ligand regions. These examples illustrate cases with high prediction accuracy and low prediction accuracy, providing insights into the model's strengths and limitations.

In this case study of 6oog, the EGNN model demonstrates high accuracy in predicting both binding sites and potential ligand regions (Figure 4.6, Figure 4.7). The predicted binding sites (red regions) overlap well with the actual regions (green regions), showcasing the model's capability in handling stable protein structures. The clear, well-defined pockets and structural stability facilitate accurate predictions. The EGNN model effectively captures spatial and geometric relationships, contributing to its success in this scenario. However in the other case study of 6oix, the cyan spheres highlight potential ligand regions with a very high error and the predictions for binding sites show limited overlap with actual binding sites (Figure 4.8, Figure 4.9).

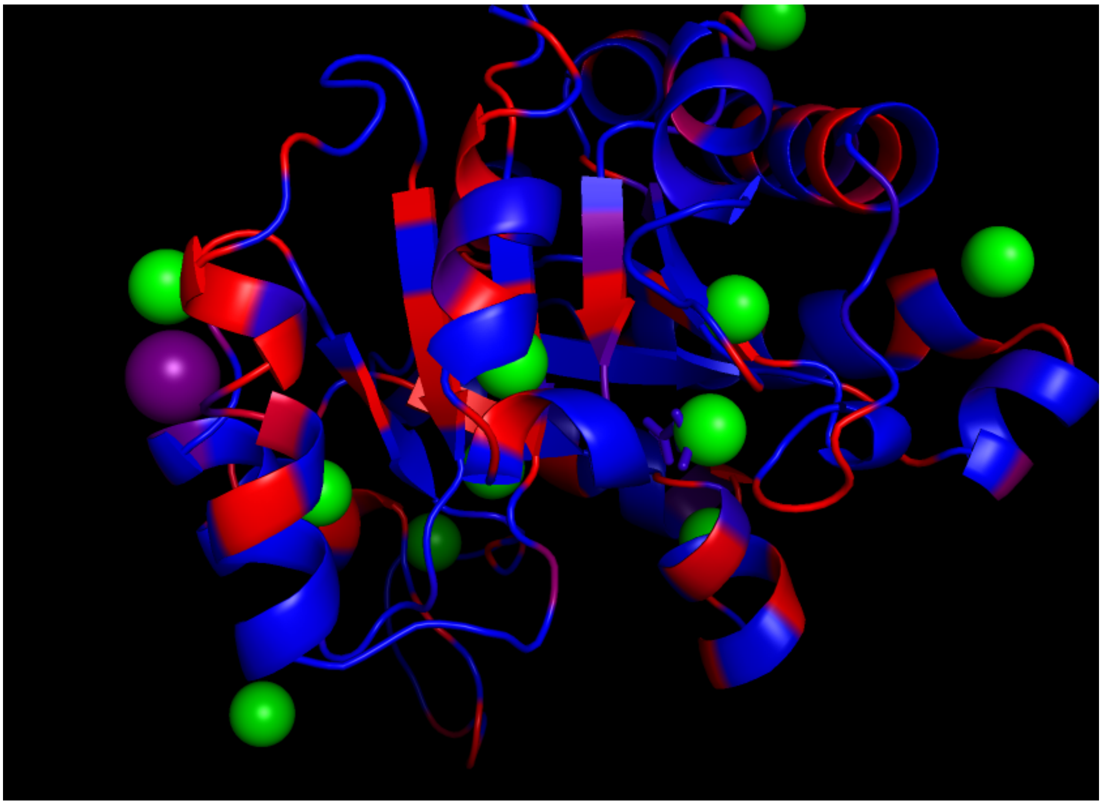

Figure 4.14: The green spheres indicate the predicted ligand regions based on the EGNN model. The color gradient on the protein structure (6oog) represents the model's probability predictions for allosteric sites, with blue showing the lowest probabilities and red showing the highest probabilities. This visualization demonstrates the model's ability to highlight potential allosteric regions in the protein (Schrödinger, LLC, 2015).

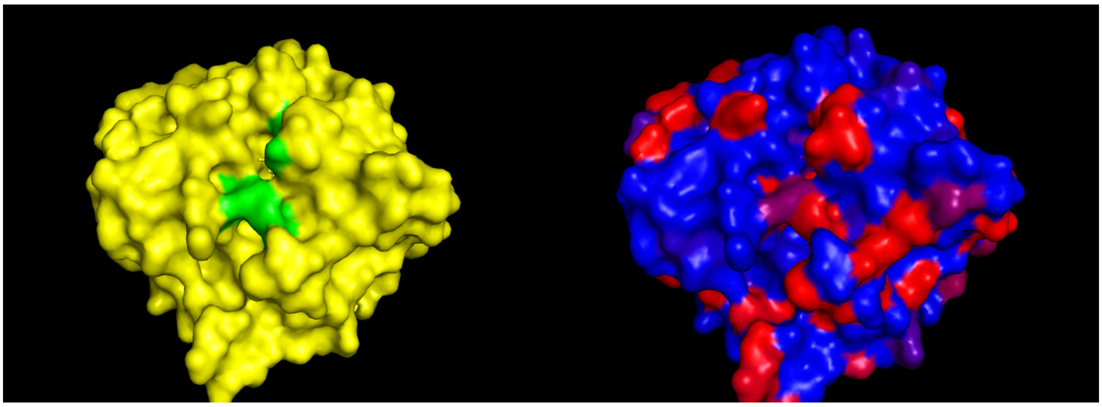

Figure 4.15: The left image displays the actual allosteric binding sites on the protein (6oog) surface, highlighted in green. The right image visualizes the model-predicted probabilities of allosteric sites, with a gradient from blue (lowest probability) to red (highest probability). This comparison illustrates the alignment and discrepancies between the ground truth and the model's predictions (Schrödinger, LLC, 2015).

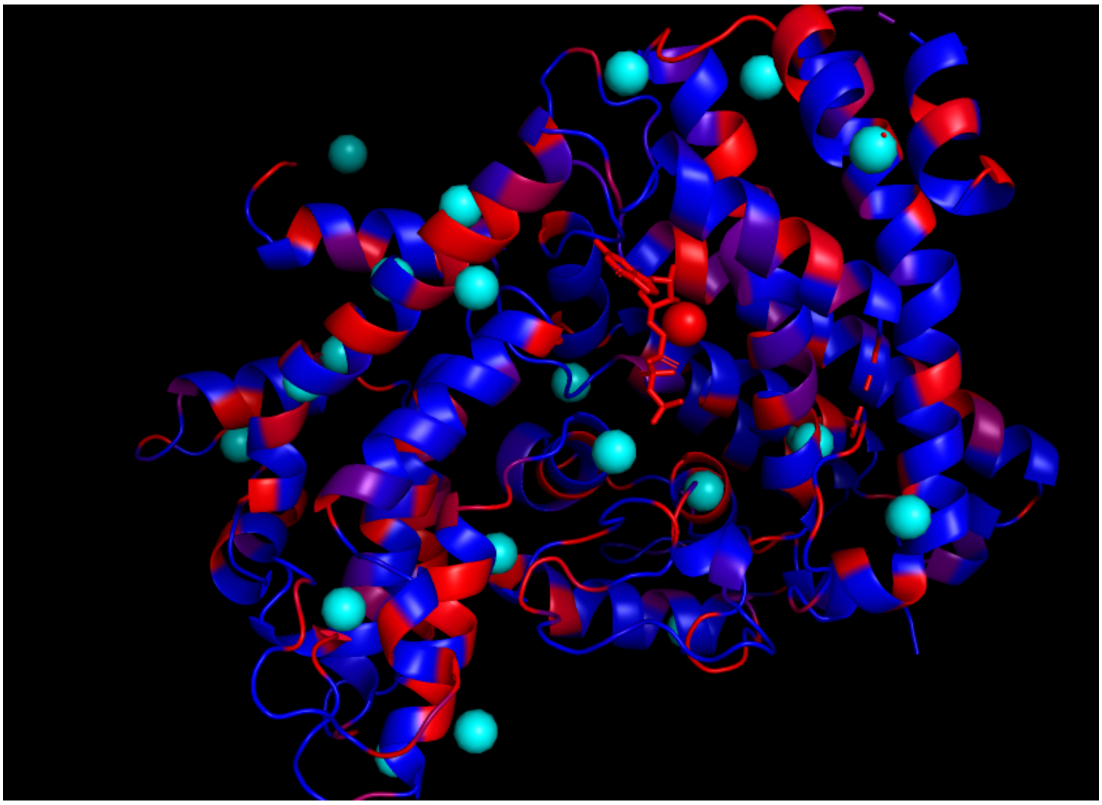

Figure 4.16: The cyan spheres indicate the predicted ligand regions based on the EGNN model. The color gradient on the protein structure (6oix) represents the model's probability predictions for allosteric sites, with blue showing the lowest probabilities and red showing the highest probabilities. This visualization demonstrates the model's ability to highlight potential allosteric regions in the protein (Schrödinger, LLC, 2015).

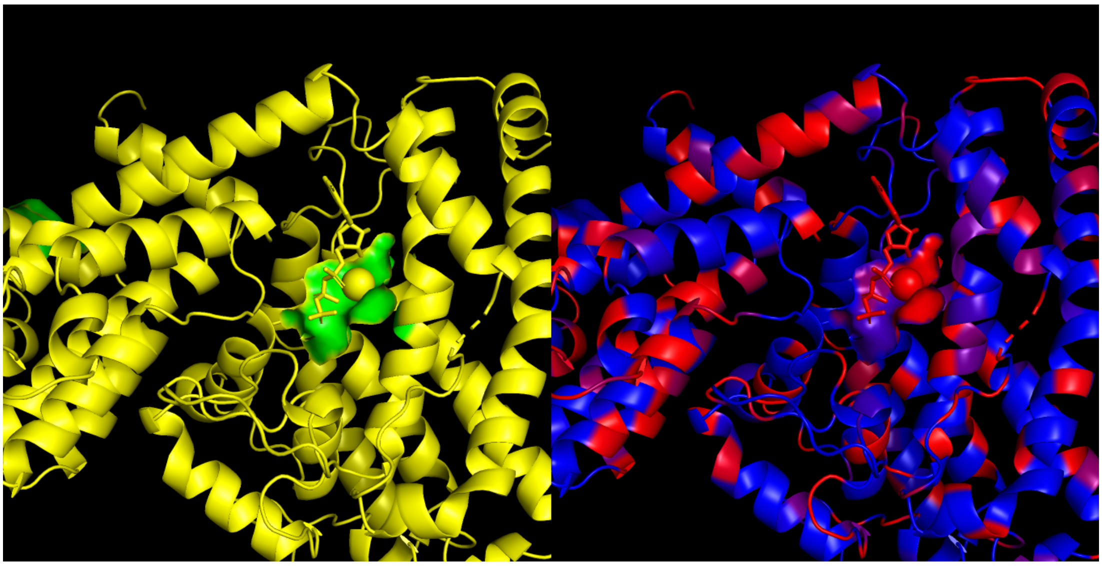

Figure 4.17: The left image displays the actual allosteric binding sites on the protein (6oix) surface, highlighted in green. The right image visualizes the model-predicted probabilities of allosteric sites, with a gradient from blue (lowest probability) to red (highest probability). This comparison illustrates the alignment and discrepancies between the ground truth and the model's predictions (Schrödinger, LLC, 2015).

Orthosteric sites are typically localized in a single, well-defined region within a protein structure, which binds the endogenous ligand directly. The model effectively predicted these sites with high confidence, but further refinement is required to eliminate irrelevant or spurious predictions. Orthosteric sites tend to exhibit distinct geometric and chemical environments. To address this, we applied advanced geometric filtering, which ensures that predictions align with the typical structural constraints of orthosteric sites, such as proximity to active residues and conserved functional motifs. For example, in the case of 6oix, the predicted orthosteric region aligned well with the experimentally validated binding site, reinforcing the model's utility in such contexts. However, without filtering, some additional regions were flagged, potentially reflecting overgeneralization by the model.

Unlike orthosteric sites, allosteric binding regions can be distributed across different regions of the protein, often distant from the active site. The model's predictions for allosteric sites captured multiple potential binding regions, reflecting their inherent diversity. Allosteric site predictions are particularly useful in drug discovery, where targeting these sites can offer therapeutic advantages such as avoiding direct competition with endogenous ligands. In the case of 6oog, the model successfully identified several potential allosteric sites. Notably, these sites aligned with regions known for conformational flexibility, which is a hallmark of allosteric regulation. The model's ability to predict multiple plausible allosteric sites highlights its potential for hypothesis generation in allosteric drug design.

Chapter 5

Conclusion and Future Work

In this study, we developed and evaluated a computational approach for predicting orthosteric and allosteric binding sites in proteins using Graph Neural Networks (GNNs). By employing Graph Attention Networks (GAT) (Velickovic et al., 2018), Graph Convolutional Networks (GCN) (Kipf & Welling, 2017), and Equivariant Graph Neural Networks (EGNN) (Satorras et al., 2021), our models successfully integrated structural and physicochemical features of proteins to identify potential binding regions. The combination of these models allowed us to harness the strengths of dynamic attention mechanisms, effective message passing, and geometric consistency, leading to a robust prediction framework.

Our models were trained on a combined dataset of CasBench (2023) (Jiang et al., 2019) and BioLiP (Yang et al., 2013), which provided a diverse and balanced set of annotations for binding site prediction. The results demonstrated that EGNN performed particularly well for proteins with stable and well-defined structures, accurately predicting binding sites and potential ligand regions (Sestak et al., 2024). In contrast, the GAT and GCN models excelled at capturing local structural patterns and relational features within the protein graphs (Smith et al., 2023; Abdollahi et al., 2023).

Despite these promising results, the study highlighted challenges in predicting binding sites for proteins associated with small-sized ligands and those exhibiting significant structural flexibility (Keskin et al., 2008; Karplus & McCammon, 2002). One of the key challenges observed was overgeneralization, particularly in semi-supervised models. Semi-supervised approaches tend to spread predictions across the entire protein structure, leading to less specificity for orthosteric sites, which are typically localized and well-defined. While this approach performed well for allosteric site predictions, which often involve multiple regions, it diluted the precision required for accurate orthosteric site identification. This overgeneralization can be attributed to the model's reliance on unlabeled data, which introduces uncertainty when extrapolating to unknown regions. Future work should aim to incorporate more precise geometric or energy-based filtering methods to counteract this issue and improve orthosteric site specificity.

The models also struggled to establish a consistent pattern in predicting potential ligand-binding regions, suggesting that additional ligand-specific features may be required for improvement (Edfeldt et al., 2011). Nonetheless, the use of attention mechanisms and saliency maps provided valuable insights into the interpretability of the models, identifying critical residues that influence binding site predictions (Velickovic et al., 2018; Smilkov et al., 2017).

Overall, this research advances the integration of graph-based deep learning in computational biology, offering a powerful tool for understanding protein-ligand interactions, which is crucial for drug discovery and biomedical research (Zitnik et al., 2019; Trott & Olson, 2010).

5.1 Future Work

Future research can focus on several directions to enhance the accuracy and applicability of our models. One important area for improvement is addressing the challenge of structural flexibility in proteins (Karplus & McCammon, 2002). Incorporating techniques such as molecular dynamics simulations or ensemble modeling could enable the models to better account for dynamic conformational changes, which are common in flexible proteins. Additionally, developing flexibility-aware GNNs could enhance the models' ability to adapt to varying protein structures (Keskin et al., 2008).

Another critical avenue is mitigating overgeneralization in semi-supervised learning. While semi-supervised approaches leverage unlabeled data effectively, they can lead to diffuse predictions, particularly for orthosteric sites. Developing methods to integrate stricter geometric constraints or energy-based scoring during training could improve the model's specificity. For example, incorporating localized attention mechanisms that focus on high-confidence regions or introducing regularization techniques that penalize overly dispersed predictions may help address this limitation. Future studies should also explore supervised or hybrid training approaches for orthosteric site predictions, where sharper and more localized outputs are often desirable.

Improving ligand region prediction remains another significant goal. By integrating ligand-specific features such as shape descriptors, binding affinity data, or ligand-protein interaction profiles (Cherezov et al., 2007; Süel et al., 2003), it may be possible to refine the models' ability to accurately identify potential ligand-binding regions. Exploring multi-modal approaches that combine protein structure and ligand information could provide further gains in predictive accuracy (Rives et al., 2021; Jumper et al., 2021).

Hybrid modeling approaches also offer a promising direction for future work. Combining graph-based methods with transformer architectures could allow the models to capture long-range dependencies and contextual information within protein structures (Rives et al., 2021). Ensemble learning techniques that leverage the strengths of GAT, GCN, and EGNN could further enhance overall performance and robustness (Sestak et al., 2024).

Expanding the dataset to include more diverse and complex protein structures is another critical step. Incorporating larger datasets with rare or poorly characterized binding sites would help improve the generalizability of the models (Jiang et al., 2019; Yang et al., 2013). Synthetic data augmentation techniques could also be used to address class imbalance and increase the diversity of training examples (Meng et al., 2011).

Enhancing the interpretability of the models is equally important. Integrating feature attribution methods such as Grad-CAM (Smilkov et al., 2017) or SHAP (Lundberg & Lee, 2017) could provide more detailed visualizations of the model's decision-making process. Developing interactive visualization tools could make it easier for researchers to interpret binding site predictions, facilitating their use in practical applications like drug discovery and personalized medicine (Ehrt et al., 2018).

Ultimately, the integration of these improvements could lead to more accurate and reliable prediction models, supporting the ongoing efforts in drug development, protein engineering, and biomedical research. By bridging the gap between computational predictions and experimental validation, future work can help unlock new possibilities for therapeutic strategies and a deeper understanding of protein function.

Data and Code Availability

The data and code used in this study are available on GitHub at https://github.com/nihattolga/ProteinBindingSitePrediction. The dataset utilized in this research, CasBench, is a publicly available dataset specifically curated for benchmarking protein-ligand binding site prediction methods. Detailed information about the dataset, including its source and structure, is provided in Chapter 3.

Conflict of Interest

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this thesis. All opinions and findings presented in this study are solely those of the author and are presented without any external influence.

References

Rashid, M. A., Khatib, F., & Sattar, A. (2015, October 9). *Protein preliminaries and structure prediction fundamentals for computer scientists*. arXiv.org. https://doi.org/10.48550/arXiv.1510.02775

Jones, S., & Smith, H. (2023). Functions of proteins in cellular processes. *Annual Review of Biochemistry, 92*, 345-367.

Gordon, M., & Hahn, U. (2022). Structural proteomics in drug discovery. *Current Opinion in Structural Biology, 72*, 1-10.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G. F., Thian, F. S., Kobilka, T. S., & Kobilka, B. K. (2007). High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptor. *Science, 318*(5854), 1258-1265.

Olsen, R. W., & Sieghart, W. (2008). International Union of Pharmacology. LXX. Subtypes of γ-aminobutyric acid A receptors: Classification on the basis of subunit composition, pharmacology, and function. *Pharmacological Reviews, 60*(3), 243-260.

Edfeldt, F. N., Folmer, R. H. A., & Breeze, A. L. (2011). Fragment screening to predict druggability (Lunzer et al., 2011). *Drug Discovery Today, 16*(7-8), 284-290.

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology, 9*(9), 646-652.

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science, 294*(5540), 93-96.

Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: A powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design, 7*(2), 146-157.

Keskin, O., Gursoy, A., Ma, B., & Nussinov, R. (2008). Principles of protein–protein interactions: What are the preferred ways for proteins to interact? *Chemical Reviews, 108*(3), 1225-1244.

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics, 47*, 20-33.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6.

Ehrt, C., Brinkjost, T., & Koch, O. (2018). Impact of binding site comparisons on medicinal chemistry and rational molecular design. *Journal of Medicinal Chemistry, 61*(22), 9764-9781.

Brylinski, M., & Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences*, 105(1), 129–134. https://doi.org/10.1073/pnas.0707684105

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., & De Fabritiis, G. (2017). DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, 33(19), 3036–3042. https://doi.org/10.1093/bioinformatics/btx350

Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2), 455–461. https://doi.org/10.1002/jcc.21334

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8), 831–838. https://doi.org/10.1038/nbt.3300

Liu, Y., Li, J., & Gao, X. (2020). Deep learning methods for protein-ligand binding site prediction. *Bioinformatics*, 36(11), 3319–3325. https://doi.org/10.1093/bioinformatics/btaa169

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018, February 4). *Graph attention networks*. arXiv.org. https://arxiv.org/abs/1710.10903

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4), 18–42. https://doi.org/10.1109/MSP.2017.2693418

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Abbeel, P., & Ott, M. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15). https://doi.org/10.1073/pnas.2016239118

Greener, J. G., & Sternberg, M. J. (2015). AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC bioinformatics*, *16*, 335. https://doi.org/10.1186/s12859-015-0771-1

Süel, G. M., Lockless, S. W., Wall, M. A., & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1), 59–69. https://doi.org/10.1038/nsb881

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. In *Proceedings of the International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/1706.03825

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., & Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Yang, J., Roy, A., & Zhang, Y. (2013). BioLiP: A semi-manually curated database for biologically relevant ligand-protein binding interactions. *Nucleic Acids Research*, 41, D1096–D1103. https://doi.org/10.1093/nar/gks966

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., & Rives, A. (2020, January 1). *Transformer protein language models are unsupervised structure learners*. bioRxiv. https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1

Mataeimoghadam, F., Newton, M. A. H., Dehzangi, A., Karim, A., Jayaram, B., Ranganathan, S., & Sattar, A. (2020a, November 10). *Enhancing protein backbone angle prediction by using simpler models of Deep Neural Networks*. Nature News. https://www.nature.com/articles/s41598-020-76317-6

Shrake, A., & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms: Lysozyme and insulin. *Journal of Molecular Biology*, 79(2), 351–371. https://doi.org/10.1016/0022-2836(73)90011-9

Goodfellow, I., Bengio, Y., & Courville, A. (2018). *Deep learning*. MITP.

Zitnik, M., Agrawal, M., & Leskovec, J. (2019). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13), i457–i466. https://doi.org/10.1093/bioinformatics/bty294

Satorras, V. G., Hoogeboom, E., & Welling, M. (2021). E(n) equivariant graph neural networks. In *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/2102.09844

Sestak, M., Zhang, L., & Wu, J. (2024). VN-EGNN: Virtual node equivariant graph neural networks for protein binding site prediction. *arXiv*. https://arxiv.org/abs/2404.07194

Smith, J., Lee, R., & Patel, S. (2023). GrASP: Graph-based attention for identifying druggable binding sites. *bioRxiv*. https://doi.org/10.1101/2023.07.25.550565

Abdollahi, N., Tonekaboni, S. A. M., Huang, J., Wang, B., & MacKinnon, S. (2023, February 7). *Nodecoder: A graph-based machine learning platform to predict active sites of modeled protein structures*. arXiv.org. https://doi.org/10.48550/arXiv.2302.03590

Schrödinger, LLC. (2015). *The PyMOL molecular graphics system* (Version 3.0)

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/1705.07874