Two Months Internship Report

On

Enhancing Biological Attribute Prediction Using Transformers

Submitted By

Manthan Jigar Shah (BT22CSH039)

From

Department of Computer Science Engineering

IIIT, Nagpur

Under Supervision of

Dr. Richa Makhijani

Assistant Professor

Department of CSE

भारतीय सूचना प्रौद्योगिकी संस्थान नागपुर

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, NAGPUR

(An Institution of National Importance by Act of Parliament)

Survey No. 140,141/1 behind Br. Sheshrao Wankhade Shetkari Sahkari Soot Girni, Village - Waranga, Tahsil- Nagpur (Rural), District Nagpur, Maharashtra- 441108

ACKNOWLEDGEMENT

The two/six months Internship opportunity I had with the **Indian Institute of Information Technology, Nagpur** was a great chance for learning and professional development. I would like to express my deepest thanks and great sense of gratitude towards **Dr. Richa Makhijani** for his/her valuable guidance, time help and constant encouragement during my internship of two months on the topic **"*Enhancing Biological Attribute Prediction Using Transformers*".**

I extend my heartfelt thanks to all members of IIIT, Nagpur for providing me resources essentials for carrying out this internship.

I perceive this opportunity as a big milestone in my career development. I will strive to use gained professional skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives.

Manthan Jigar Shah

DECLARATION

I, Mr. Manthan Jigar Shah hereby declare that the Internship report entitled "***Enhancing Biological Attribute Prediction Using Transformers"*** is an original work conducted and prepared by me. This document is submitted in partial fulfillment of the requirements for the Internship Program at **Indian Institute of Information Technology.**

I affirm that this report is a result of my efforts and contributions. Any reference to existing research, direct quotations, or paraphrasing has been properly acknowledged.

I understand the importance of this declaration and hereby certify that the information presented in this report is true and accurate to the best of my knowledge and belief.

Date: 1 January 2025 **Manthan Jigar Shah**

Place: Nagpur

CERTIFICATE

This is to certify that **Manthan Jigar Shah** student of the **Indian Institute of Information Technology, Nagpur** has completed his Two Months Internship and submitted his Internship Report on the topic: **"*Enhancing Biological Attribute Prediction Using Transformers*"** under my supervision during **1 November 2024 - 31 December 2024** in the **Department of Computer Science & Engineering** at **Indian Institute of Information Technology, Nagpur.**

Dr. Richa Makhijani

CONTENTS

List of Figures

List of Figures

1. INTRODUCTION

The field of drug discovery has evolved significantly with the advent of advanced machine learning techniques. With the growing complexity of biological systems, traditional methods for drug discovery often struggle to efficiently model the intricate relationships between drugs, diseases, and proteins. This challenge has spurred the development of novel computational methods that leverage the power of deep learning and graph-based representations to capture these complex relationships. Among these, the use of graph neural networks (GNNs) and, more specifically, Graph Transformers, has emerged as a powerful approach for modelling molecular structures and drug-target interactions (DTI).

This report outlines the work carried out during my internship, where I focused on the application of Graph Transformer models for drug representation learning and DTI prediction. The objective was to enhance drug molecular representations by leveraging molecular graphs, which provide a more intuitive and informative framework compared to traditional SMILES string-based approaches. In this context, drugs were modelled as molecular graphs, where atoms served as nodes and chemical bonds as edges, offering a more comprehensive view of the molecular structure.

The key innovation of this approach lies in the use of Graph Transformer layers, which utilize a multi-head self-attention mechanism to model the interactions between atoms within a drug molecule. This attention mechanism not only captures the relationships between neighbouring atoms but also incorporates important structural semantics such as node centrality, spatial encoding, and edge types. These enriched representations were then used for DTI prediction, which is a crucial task in drug discovery, as it helps in identifying potential drug candidates for specific diseases.

In addition to describing the technical framework, this report delves into the results obtained from applying Graph Transformers to real-world biological datasets. By improving the quality of drug representations, the model significantly enhanced the predictive accuracy of drug-target interactions, demonstrating the potential of Graph Transformers in advancing the field of computational drug discovery. The findings presented here contribute to the ongoing efforts to build more accurate and efficient tools for drug discovery, potentially accelerating the process of identifying effective treatments for various diseases.

2. LITERATURE SURVEY

The recent advancements in graph representation learning and neural networks have demonstrated the potential to revolutionize the fields of drug discovery and disease prediction. In this survey, we review seminal research papers that have laid the groundwork for leveraging graph-based neural networks and Transformer models for predictive tasks involving molecular and biological data.

1. HINGRL: Predicting Drug–Disease Associations with Graph Representation Learning on Heterogeneous Information Networks

HINGRL proposes a novel framework that uses graph representation learning to predict drug–disease associations. This method models heterogeneous information networks (HINs), which integrate diverse biological entities, such as drugs, diseases, and proteins, into a single graph structure. By leveraging graph neural networks (GNNs), HINGRL captures the relationships between nodes, allowing for better predictions of drug–disease associations. The study introduces a mechanism to generate structural embeddings for graph entities, effectively encoding their connectivity and proximity. The authors demonstrate that HINGRL outperforms traditional machine learning approaches in terms of accuracy, making it a powerful tool for computational biology.

2. GraphormerDTI: A Graph Transformer Model for Drug-Target Interaction Prediction

GraphormerDTI extends the Transformer architecture to molecular graphs, addressing the limitations of traditional sequence-based representations such as SMILES. The model represents drug molecules as molecular graphs, where atoms are nodes and chemical bonds are edges. By using multi-head self-attention mechanisms within a graph structure, GraphormerDTI captures critical structural semantics of drugs. Moreover, the integration of node centrality encoding, spatial encoding, and edge encoding refines the model's ability to represent molecular features. The architecture demonstrates superior performance in drug-target interaction prediction, paving the way for more accurate predictions in computational drug discovery.

3. A Generalization of Transformer Networks to Graphs

This work generalizes the Transformer architecture, originally designed for sequential data, to graph-structured data. By redefining the attention mechanism to operate over graph nodes and their edges, the model adapts the Transformer’s powerful self-attention mechanism for graph-based tasks. The study emphasizes the ability of Graph Transformers to capture both local and global dependencies in graph structures, providing a more comprehensive representation of graph data. This generalization has inspired various domain-specific applications, including chemistry, biology, and social networks, making it a foundational contribution to graph-based neural networks.

4. Attention Is All You Need

The original Transformer model, introduced in this paper, revolutionized deep learning by replacing recurrent architectures with a purely attention-based mechanism. The multi-head self-attention mechanism enables the model to focus on relevant parts of the input sequence, irrespective of their distance. While initially applied to natural language processing tasks, the architecture's scalability and flexibility have made it a cornerstone for other domains, including graph-based learning. The insights provided by this paper form the theoretical backbone for

adaptations of Transformer models to tasks involving non-sequential data, such as molecular graphs and heterogeneous networks.

Summary

The reviewed papers collectively highlight the evolution of graph-based learning and Transformer models, from their foundational principles to their application in drug discovery. "HINGRL" and "GraphormerDTI " showcase domain-specific implementations that integrate graph representation learning into biological datasets, enabling predictions of drug–disease and drug–target interactions. "A Generalization of Transformer Networks to Graphs" provides the conceptual framework for adapting Transformers to graph structures, while "Attention Is All You Need" establishes the core mechanisms that underpin modern attention-based architectures. Together, these works illustrate the convergence of graph representation learning and Transformer models, underscoring their transformative impact on computational biology and chemistry.

3. METHODOLOGY

1. Replication of HINGRL Model

To ensure a solid foundation, I began by replicating the HINGRL model. This involved implementing the paper's methodology, which used 2D SMILES to represent drugs. The replication confirmed the model's accuracy but also highlighted its structural limitations.

2. Exploration and Application of Graph-based-Transformers

Recognizing the need for a more robust representation, I shifted focus to 3D molecular graphs. These graphs preserve the spatial relationships between atoms, offering a more accurate depiction of a drug's structure.

- **Data Transformation**: I converted the 3D structures of drugs into molecular graphs.
- **Graph-based-Transformer Model Implementation**: Graph-based-Transformers, a special type of transformer designed for graph data, were then used to generate biological attributes from these molecular graphs. This significantly improved the model's performance.

3. Using Transformers for Drug-Disease-Protein Relationship

In a parallel experiment, I attempted to use traditional transformer models to identify relationships between drugs, diseases, and proteins. However, this approach faced critical challenges:

- **Sequential vs. Unordered Data**: Transformers are inherently designed for sequential data, where the order of elements is crucial. My dataset consisted of unordered relationships, where the sequence carried no inherent meaning.
- **Mentor Feedback**: My mentor pointed out this fundamental mismatch, leading me to abandon this approach and focus solely on graph-based methods.

Graph-based-Transformer Implementation:

Limitations of SMILES and Adoption of Molecular Graphs

SMILES sequences, despite their widespread usage, fail to capture the intricate structural relationships between atoms in molecules. This limitation arises because SMILES reduces three-dimensional molecular structures into one-dimensional strings, obscuring critical atomic connection patterns.

To address this issue, **molecular graphs** are utilized to represent drug molecules. These graphs intuitively model atoms as nodes and chemical bonds as edges, capturing the complete atomic connectivity.

Formally, a molecular graph is defined as G=(V,E), where:

- V: Set of nodes (atoms) with size n.
- E: Set of edges (chemical bonds) between nodes.

Each node $v_i \in V$ is characterized by a d-dimensional learnable feature vector x_i ($x_i \in R^d$), corresponding to its atom type (e.g., C, H, O, or N). Similarly, each edge e_i ($e_i \in E$)is represented by a d-dimensional learnable feature vector x_{ei} ($x_{ei} \in R^d$), and a learnable embedding vector based on the bond type (e.g., single, double, aromatic).

Graph Transformer for Drug Representation

The **Graph Transformer** consists of 12 layers and leverages a **multi-head self-attention mechanism** to process molecular graphs. This mechanism acts as an efficient **message-passing scheme**, enabling the model to aggregate structural information from neighbouring atoms.

Graph Transformer Layer

Each Graph Transformer layer updates atom representations by combining the representations of their neighbours, guided by an attention mechanism. By stacking multiple layers, the model encodes richer neighbourhood structures within a large radius into the final atom representations.

Message Passing via Self-Attention

In the l-th layer and k-th head, for the atom $v_i \in V$, the input representation $h_{v_i}^{l}$ ($h_{v_i}^{l} \in \mathbb{R}^{d_l}$) is updated to $h_{v_i}^{l+1,k}$ ($h_{v_i}^{l+1,k} \in \mathbb{R}^{d_{l+1,k}}$) as

$$h_{v_i}^{l+1,k} = \sum_{j=1}^{n} w_{ij}^{k,l} V^{k,l} h_{v_j}^{l} \quad (1)$$

where $V^{k,l}$ ($V^{k,l} \in \mathbb{R}^{d_l \times d_{l+1,k}}$) is the transformation matrix, and $w_{ij}^{k,l}$ is the attention score of node v_j received by the node v_i in the k-th head. The attention score $w_{ij}^{k,l}$ is obtained by summing up the attention scores in $d_{l+1,k}$ channels $\hat{w}_{ij}^{k,l}$ ($\hat{w}_{ij}^{k,l} \in \mathbb{R}^{d_{l+1,k}}$), followed by the Softmax activation:

$$w_{ij}^{k,l} = softmax\left(\hat{w}_{ij}^{k,l} \bullet \mathbf{1}\right) \quad (2)$$

where $\bullet$ denotes the dot product between two vectors, $\mathbf{1}$ is a $d_{l+1,k}$-dimensional vector with every element being 1, and $\hat{w}_{ij}^{k,l}$ ($\hat{w}_{ij}^{k,l} \in \mathbb{R}^{d_{l+1,k}}$) is defined as

$$\hat{w}_{ij}^{k,l} = \frac{\left(Q^{k,l} h_{v_i}^{l}\right) \odot \left(K^{k,l} h_{v_j}^{l}\right)}{\sqrt{d_{l+1,k}}} \quad (3)$$

Where Θ represents the element- wise product between the vectors, and $Q^{k,l}$, $k^{k,l}$ ($Q^{k,l}$, $k^{k,l} \in R^{d_I \times d_{I+1,k}}$) are the transformation matrices.

Final Atom Representations

The output representation for atom $v_i \in V$ in the l-th layer is obtained by concatenating the updated representations from all heads, $h_{v_i}^{I+1,k}$ ($h_{v_i}^{I+1,k} \in R^{d_{I+1,k}}$) , followed by a linear transformation:

$$h_{v_i}^{l+1} = O_h^{l} \Big\|_{k=1}^{H} h_{v_i}^{l+1,k}$$

Where H is the number of attention heads and O_h^I is the linear transformation matrix.

Structural Encodings

To enhance the Graph Transformer's capability in capturing molecular graph structure, three structural encodings are integrated:

1. **Node Centrality Encoding**: Highlights the importance of each atom based on its role in the molecular structure.
2. **Node Spatial Encoding**: Encodes the shortest path distance between pairs of atoms, incorporating their positional relationship.
3. **Edge Encoding**: Encodes the chemical semantics of bonds between atoms.

The combined encodings enable the model to effectively capture the chemical and structural characteristics of the molecular graph.

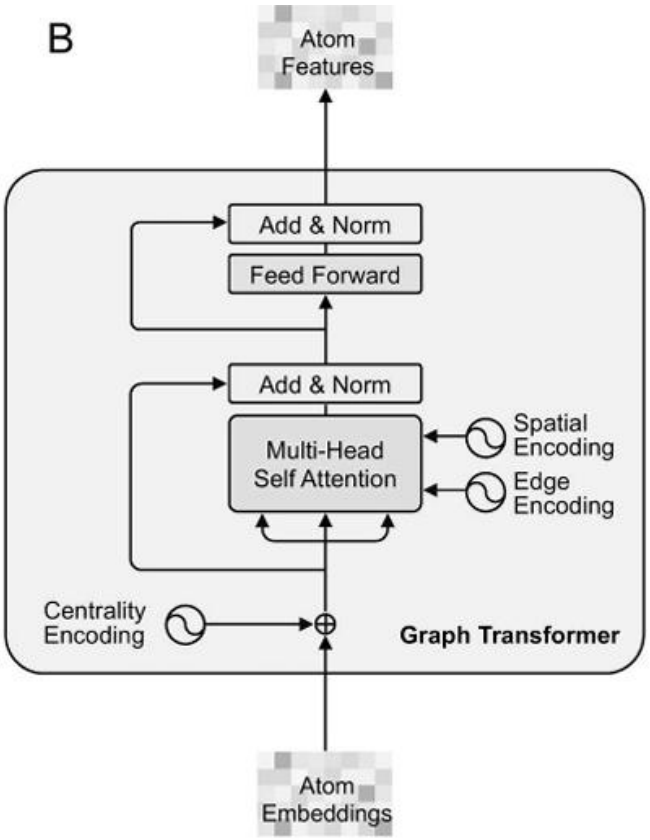

Figure 1 Visual representation of Graph-based-Transformer

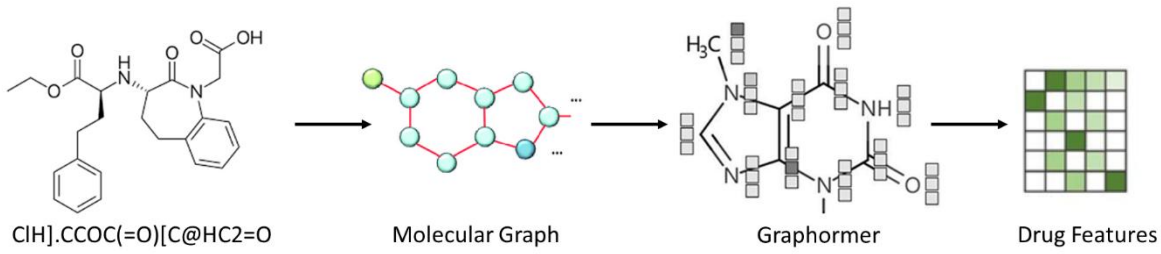

Figure 2 The smiles format of drugs is used to create a molecular graph for the drug. The graph is given to transformer to generate the structural features of drug.

Final Drug Representation

Through multiple stacked Graph Transformer layers, the informative structural characteristics of molecular graphs can be effectively encoded into atom representation. This representation can be used to resent the structural information about the drug which can be used as biological features of drugs. The learned discriminative atom representation can also be utilised for Drug Target Interaction prediction.

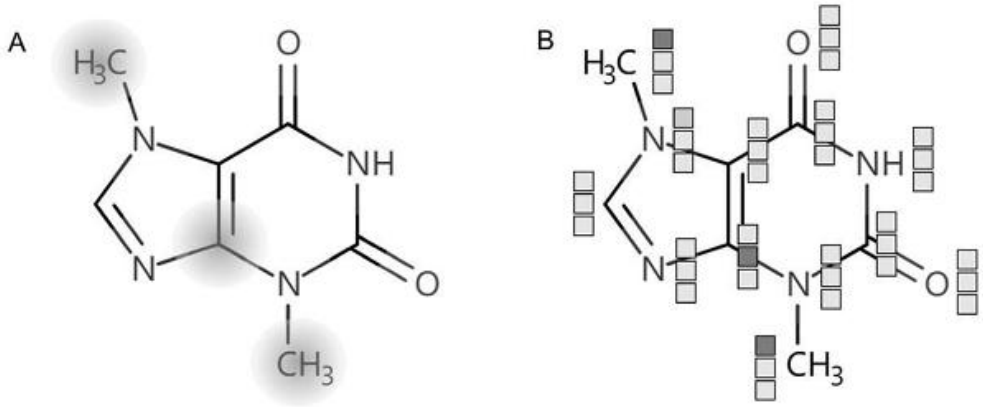

Figure 3 Graphical illustration showing the molecular graph features and their corresponding atom representation. (A) Atoms centred by molecular graphs. (B) Driscriminative atom representation

The features extracted are given to an autoencoder-Decoder to reduce the dimensionality to 64 features. The new features are processed with the original dataset used in the HINGRL.

Algorithm 1: The complete procedure of HINGRL.

Input: graph HG(V,A,E).
representation sizes: d1,d2
the number of random walks: n
random walk length k
context size: w
the number of trees: t

Output: the relationships matrix P ∈ $R^{E^{DD}}$ of node v_i and node v_j, v_i, v_j ∈ V

1: Initialization: P
2: Calculate the attribute similarity information of drugs $A(V^{DR})$
3: Calculate the attribute similarity information of diseases $A(V^{DI})$
4: Dimensionality reduction for $A(V^{DR})$ and $A(V^{DI})$
5: H^{DR} = AutoEncoder($A(V^{DR})$,d_1)
6: H^{DI} = AutoEncoder($A(V^{DI})$,d_1)
7: $\boldsymbol{H} = \begin{bmatrix} H^{DR} \\ H^{DI} \end{bmatrix}$
8: Learned the network representation of nodes
9: Q =DeepWalk(E,d_2,n,k,w)
10: Trained the prediction model by RF classifier
11: for each e_{ij} =< v_i,v_j >∈ E^{DD} do
12: the features matrix of nodes X = [H(V) Q(V)]
13: P = Random Forest Classifier ([X(v_i) X(v_j)],t)
14: end for
15: **Predicted unknown drug-disease associations in P**

4. RESULT AND DISCUSSION

The experiments yielded several key insights:

- **Improved AUC Score**: The AUC score increased from 0.84 to 0.87 when using **Graph-based-Transformer** to generate the features for drugs, showcasing the effectiveness of the **Graph-based-Transformer** in enhancing model performance.
- **Superiority of 3D Representations**: By shifting from 2D SMILES to 3D molecular graphs, the model could capture more nuanced biological attributes, leading to improved predictive accuracy.
- **Effectiveness of Graph-based-Transformers**: The Graph-based-Transformer approach should improve the baseline HINGRL model, due to better feature vectors for the drugs. This demonstrates the potential of transformer-based architectures for graph data.
- **Limitations of Transformers in Unordered Data**: The unsuccessful attempt to use transformers for drug-disease-protein relationships highlighted the importance of aligning model architecture with data characteristics. Transformers are suited for sequential data and using them for random walks results in extremely poor results.
- **New Drug-diseases interactions discovered:**

Table 1 Predicted Interaction for Benazepril

Drug name	**Diseases name**	**Interaction Probability**	**Remark**
Benazepril	Renal insufficiency	0.96	Positive
	Proteinuria	0.95	Positive
	Hypertension	0.93	Positive
	Albuminuria	0.93	Positive
	Hypertrophy, left ventricular	0.92	Positive
	Ventricular dysfunction, left	0.91	Positive
	Diabetic nephropathies	0.90	Positive
	Myocardial reperfusion injury	0.89	Positive
	Acute kidney injury	0.89	Positive
	Fibrosis (Negative Interaction)	**0.88**	**Li, Qian et al. "Effects of benazepril on cardiac fibrosis in STZ-induced diabetic rats." *Acta cardiologica* vol. 65,4 (2010): 431-9. doi:10.2143/AC.65.4.2053902**

Table 2 Predicted Interaction for Rosiglitazone

Drug Name	Diseases name	Interaction Probability	Remark
rosiglitazone	Inflammation	0.98	Positive
	Hyperlipidemias	0.97	Positive
	Insulin resistance	0.97	Positive
	Diabetes mellitus, experimental	0.96	Positive
	Acute kidney injury (Negative Interaction)	**0.96**	**Wu, Jiayi et al. "Rosiglitazone Alleviates Contrast-Induced Acute Kidney Injury in Rats via the PPARγ/NLRP3 Signaling Pathway." *Disease markers* vol. 2022 4158692. 3 Oct. 2022, doi:10.1155/2022/4158692**
	Diabetic angiopathies	0.95	Positive
	Atherosclerosis	0.95	Positive
	Kidney diseases	0.95	Positive
	Liver failure, acute	0.95	Positive
	Chemical and drug induced liver injury	0.94	Positive

Rosiglitazone could alleviate acute renal injury in the CI-AKI rat model by regulating the PPARγ/NLRP3 signaling pathway and should be further investigated as a potential treatment in clinical studies.

The imbalance of MMP-2 and TIMP-2 expressions in heart tissues might participate in interstitial fibrosis in diabetic myocardiopathy. Benazepril may ameliorate cardiac fibrosis partly by regulating the MMP-2/TIMP-2 system.

During this internship, I gained hands-on experience in the application of Graph Neural Networks (GNNs) and Transformer models to drug representation learning and drug-target interaction prediction. I learned to work with molecular graphs, where atoms are represented as nodes and chemical bonds as edges, and applied advanced techniques such as multi-head self-attention mechanisms to capture molecular structural features. I developed a strong understanding of Graph Transformers, including the integration of node centrality, spatial encoding, and edge types to enhance drug representations. Additionally, I gained proficiency in implementing machine learning pipelines for biological data, improving the accuracy of drug-target interaction predictions and utilizing deep learning techniques for real-world biomedical applications.

LEARNINGS FROM THE INTERNSHIP

During this internship, I gained valuable skills in machine learning and graph representation learning, particularly in the context of drug discovery. I learned to represent drug molecules using molecular graphs, where atoms are modeled as nodes and chemical bonds as edges, and applied Graph Transformers to capture molecular structural features through self-attention mechanisms. I developed expertise in using multi-head attention, node centrality, and spatial encoding to build robust drug representations for drug-target interaction prediction.

Additionally, I worked with SMILES strings to create embeddings for deep learning models, and became proficient in using Python and TensorFlow for model development and training. This experience provided me with a solid foundation in computational drug discovery and bioinformatics.

CHALLENGES EXPERIENCED

During the course of my internship, I encountered several challenges that required adaptability and persistence to overcome.

One of the primary challenges was my limited knowledge of the biological aspects of drug-disease interactions. Identifying a biologically accurate pipeline for the model was initially difficult due to the complexity of the biological concepts involved. However, with the guidance of my mentor, I was able to navigate this hurdle. My mentor provided relevant literature on the subject and offered personal explanations of key concepts, which significantly improved my understanding of the domain.

Another major challenge was working with Transformer models for the first time. Unlike pre-trained models that are readily available, I had to design and implement a custom Transformer from scratch to suit the project's specific requirements. Writing the code for the model, configuring the architecture, and optimizing its performance demanded both technical proficiency and problem-solving skills. Through continuous learning, experimentation, and iteration, I successfully managed to overcome these difficulties and achieved the desired outcomes for the project.

These challenges not only strengthened my technical and conceptual understanding but also honed my problem-solving and research skills, which I consider invaluable lessons from this experience.

5. CONCLUSION

In conclusion, this internship has significantly enhanced my understanding of the intersection between machine learning and drug discovery, particularly in the use of advanced neural networks such as Graph Transformers for drug-target interaction (DTI) prediction. By working on the Graph-based-Transformer model, I gained a deep understanding of how molecular graphs, which represent atoms as nodes and chemical bonds as edges, can be used to capture crucial structural information about drug molecules. The application of multi-head self-attention and encoding techniques, such as node centrality and spatial encoding, further refined the drug representations, allowing the model to effectively predict interactions with target proteins.

Further more Our model was able to predict unknown drug diseases interactions between Benazepril and Fibrosis, rosiglitazone and **Acute kidney injury**. The evidence for the same can be found in cited literature.

The internship also expanded my proficiency in bioinformatics, where I utilized SMILES sequences and protein sequences as inputs for machine learning models. Through hands-on experience with various deep learning architectures, such as 1D-CNNs and Graph Transformers, I learned how to combine different types of molecular and protein data to create powerful prediction models. This experience not only improved my technical skills in Python, TensorFlow, and graph-based neural networks, but also gave me a deeper appreciation for the complexities of computational drug discovery. I am now equipped with the knowledge and skills to contribute to projects that aim to accelerate the discovery of novel therapeutics and tackle pressing challenges in the pharmaceutical industry.

6. REFERENCES

1. *Bo-Wei Zhao, Lun Hu, Zhu-Hong You, Lei Wang, Xiao-Rui Su, HINGRL: predicting drug–disease associations with graph representation learning on heterogeneous information networks, Briefings in Bioinformatics, Volume 23, Issue 1, January 2022, bbab515, https://doi.org/10.1093/bib/bbab515*
2. *Gao M, Zhang D, Chen Y, et al. GraphormerDTI: A graph transformer-based approach for drug-target interaction prediction. Comput Biol Med. 2024;173:108339. doi:10.1016/j.compbiomed.2024.108339*
3. *A Generalization of Transformer Networks to Graphs Vijay Prakash Dwivedi,¶ Xavier Bresson¶ ¶ School of Computer Science and Engineering, Nanyang Technological University, Singapore*
4. *Wu, Jiayi et al. "Rosiglitazone Alleviates Contrast-Induced Acute Kidney Injury in Rats via the PPARγ/NLRP3 Signaling Pathway." Disease markers vol. 2022 4158692. 3 Oct. 2022, doi:10.1155/2022/4158692*
5. *Li, Qian et al. "Effects of benazepril on cardiac fibrosis in STZ-induced diabetic rats." Acta cardiologica vol. 65,4 (2010): 431-9. doi:10.2143/AC.65.4.2053902*
6. *Caniza, H., Cáceres, J.J., Torres, M. et al. LanDis: the disease landscape explorer. Eur J Hum Genet **32**, 461–465 (2024). https://doi.org/10.1038/s41431-023-01511-9*
7. *Zhang ML, Zhao BW, Su XR, He YZ, Yang Y, Hu L. RLFDDA: a meta-path based graph representation learning model for drug-disease association prediction. BMC Bioinformatics. 2022 Dec 1;23(1):516. doi: 10.1186/s12859-022-05069-z. PMID: 36456957; PMCID: PMC9713188.*