AI Driven Computational Drug Discovery using Graph Neural Networks for Dengue virus DENV-2

Abraham Demeke Alemu

A Thesis Submitted to

The Department of Computer Science and Engineering

College of Electrical Engineering and Computing

Presented in Partial Fulfilment of the Requirement for the Degree of

Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

May, 2025

Adama, Ethiopia

AI Driven Computational Drug Discovery using Graph Neural Networks for Dengue virus DENV-2

Abraham Demeke Alemu

Advisor: Dr. Sintayehu Hirpassa (Ph.D)

A Thesis Submitted to

The Department of Computer Science and Engineering
College of Electrical Engineering and Computing

Presented in Partial Fulfilment of the Requirement for the Degree of
Master's in Computer Science and Engineering

Office of Graduate Studies
Adama Science and Technology University

May, 2025

Adama, Ethiopia

DECLARATION

I hereby declare that this Master's thesis entitled "**AI Driven Computational Drug Discovery using Graph Neural Networks for Dengue virus DENV-2**" is my original work. It has not been submitted for the award of any academic degree, diploma, or certificate at any other university. All sources of materials used in this thesis have been duly acknowledged by proper citation.

Abraham Demeke Alemu	_____	_____
Name of Student	Signature	Date

RECOMMENDATION OF ADVISORS

I, the advisor of this thesis, hereby certify that I have read the revised version of the thesis entitled "**AI Driven Computational Drug Discovery using Graph Neural Networks for Dengue virus DENV-2**" prepared under my guidance by **Abraham Demeke Alemu** submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science and Engineering. Therefore, I recommend the submission of the revised version of the thesis to the department following the applicable procedures.

Dr. Sintayehu Hirpassa (Ph.D)	_____	_____
Major Advisor	Signature	Date

APPROVAL PAGE

I, the advisor of the thesis entitled "**AI Driven Computational Drug Discovery using Graph Neural Networks for Dengue virus DENV-2**" and developed by **Abraham Demeke Alemu**, hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Dr. Sintayehu Hirpassa (Ph.D) _____ _____

Major Advisor Signature Date

We, the undersigned, members of the board of examiners of the thesis by **Abraham Demeke Alemu** have read and evaluated the thesis entitled "**AI Driven Computational Drug Discovery using Graph Neural Networks for Dengue virus DENV-2**" and examined the candidate during open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

_____ _____ _____

Chair Person Signature Date

_____ _____ _____

Internal Examiner Signature Date

_____ _____ _____

External Examiner Signature Date

Finally, approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and College Graduate committee (CGC).

_____ _____ _____

Department Head Signature Date

_____ _____ _____

College Dean Signature Date

_____ _____ _____

Office of Postgraduate Studies, Dean Signature Date

ACKNOWLEDGEMENT

First and foremost, I would like to thank God for blessing me with the health and strength to undertake and complete this thesis. I am grateful for all the opportunities and challenges that have come my way during this process, as they have helped me grow both academically and personally.

I am sincerely grateful to my advisor, Dr. Sintayehu Hirpassa (Ph.D), for his invaluable guidance, understanding, and unwavering support. His positive encouragement was instrumental in helping me complete this thesis. It has been a great honor to have him as my advisor. I also extend my deepest gratitude to Dr. Teklu Urgessa (Ph.D.) for his insightful guidance, constructive comments, and kind support.

I would like to express my heartfelt thanks to my parents, brother and sister for their unconditional support and prayers. I am also deeply grateful to my classmates and colleagues who motivated and supported me throughout this journey.

Lastly, I extend my heartfelt gratitude to all the staff members of the ASTU Computer Science and Engineering department for their invaluable advice, motivation, and guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AI: Artificial Intelligence
ANN: Artificial Neural Network
ADMET: Absorption, Distribution, Metabolism, Excretion, and Toxicity
BCE: Binary Cross Entropy
B&W: Black and White
CDD: Computational Drug Discovery
DENV-2: Dengue Virus Serotype 2
DL: Deep Learning
DF: Dengue Fever
DFT: Density Functional Theory
DHF: Dengue Hemorrhagic Fever
DNN: Deep Neural Network
DSS: Dengue Shock Syndrome
GAN: Generative Adversarial Network
GAT: Graph Attention Networks
GCN: Graph Convolutional Networks
GPU: Graphics Processing Unit
IC50: Half Maximal Inhibitory Concentration
kNN: k-Nearest Neighbors
MD: Molecular Dynamics
ML: Machine Learning
MSE: Mean Squared Error
MAE: Mean Absolute Error
NS2B: Non-Structural Protein 2B
NCBI: National Center for Biotechnology Information
NS3: Non-Structural Protein 3
PCC: Pearson's correlation coefficients
PDB: Protein Data Bank
QSAR: Quantitative Structure-Activity Relationship
RF: Random Forest
SBVS: Structure-Based Virtual Screening
SMILES: Simplified Molecular Input Line Entry System
SVM: Support Vector Machine

VS: Virtual Screening
VGAE: Variational Graph Autoencoder
ZINC: ZINC Database for Virtual Screening
ChEMBL: Chemical Database of Bioactive Molecules
2D: Two-Dimensional
3D: Three-Dimensional

ABSTRACT

Dengue Virus Serotype 2 (DENV-2) poses a critical global health threat due to its severe complications such as dengue hemorrhagic fever and dengue shock syndrome. The absence of effective antiviral therapies and limitations in mosquito control strategies used currently is another concern. This study introduces an Artificial Intelligence (AI)-driven computational drug discovery pipeline by using Graph Neural Networks (GNNs) to identify and design potent antiviral compounds targeting this disease. The main aim of this study is to accelerate drug discovery process by identifying drug candidates. The methodology involved collecting bioactivity datasets from ChEMBL database, drug molecules from DrugBank databases and a target protein on DENV-2 from Protein Data Bank (PDB) database. Then preprocessing the bioactivity data with Simplified Molecular Input Entry Line System (SMILES)-to-graph conversion, Absorption Distribution Metabolism Excretion (ADME) filtering, and Lipinski's Rule of Five (Ro5) compliance and developing a Hybrid Graph Convolution Network-Graph Attention Network (GCN-GAT) architecture to predict negative logarithm (base 10) of the Half-Maximal Inhibitory Concentration (pIC50) values for Non-Structural Protein 3 (NS3) inhibitors. The GNN model trained on augmented datasets achieved a state-of-the-art R^2 score of 0.9938, demonstrating exceptional predictive accuracy for bioactivity. Additionally, a Variational Graph Autoencoder (VGAE) enabled de novo molecule generation. Through Virtual screening of Food and Drug Administration (FDA)-approved drugs from DrugBank database, we identified potential drugs for repurposing. Then VGAE model is employed to generate novel inhibitors for drug designing purpose. Also, we used molecular docking to validate binding affinities (≤ -5 kcal/mol) of top candidates. This work establishes a robust framework for AI-driven computational drug discovery for combating DENV-2 and other challenging diseases. Future directions include expanding datasets with diverse chemical libraries, refining the generative model to identify better drug candidates to bridge the gap between computational predictions and clinical translation.

Keywords: *Dengue Virus, DENV-2, AI-Driven Drug Discovery, GNNs, VGAE, Molecular Docking, Drug Repurposing, Drug Designing, Computational Drug Discovery, Deep Learning, Healthcare, Viral Diseases.*

CHAPTER ONE: INTRODUCTION

1.1BACKGROUND OF STUDY

Dengue virus is a mosquito-borne infectious disease that spreads quickly. And the cases are primarily reported from areas of tropical and subtropical regions of the world. It has a heavy burden in terms of both mortality and morbidity. The World Health Organization has reported a steep increase in the global burden of dengue over the last two decades. Around half of the world's population is at risk of dengue infection, with an estimated 100 to 400 million infections yearly. It is transmitted by the bite of an infected female Aedes mosquito, mainly by Aedes aegypti, and rarely by Aedes albopictus. The virus causes diseases such as dengue fever (DF), dengue hemorrhagic fever (DHF), and dengue shock syndrome (DSS), which could lead to serious health complications that are fatal (Gautam et al., 2023).

Dengue virus (DENV) is a single-stranded positive RNA virus that belongs to the genus Flavivirus and family Flaviviridae. It has four antigenically distinct but closely related serotypes: DENV-1, DENV-2, DENV-3, and DENV-4. These serotypes share about 65-70% sequence homology but differ in their antigenic properties, which means that infection by one serotype confers lifelong immunity against it but only partial and temporary cross-immunity against the others. Among these, DENV-2 is often associated with severe cases of dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), posing a significant challenge for disease management (Halder et al., 2024).

Mosquito control programs have proved to be not totally effective. And no antiviral drugs have been developed so far. The search for therapeutic strategies targeting viral enzymes essential for virus replication has expanded in recent years to achieve inhibition of the virus entry step (Gallo et al., 2024).

A number of drugs and therapeutic methods were tried to combat dengue. Traditional interventions mostly relieve the symptoms, and for a long period of time, no specific antiviral treatment has been approved into clinical practice. The classes of drugs studied within traditional

systems include repurposing of antivirals or new molecules that act on viral replication, entry, or binding of proteins. For example, ribavirin and chloroquine were studied and did not yield clinical improvement because of toxicity or minimum efficacy (Gautam et al., 2024).

Besides, some computational analyses have identified inhibitors such as Non-Structural Protein 2B- Non-Structural Protein 3 (NS2B-NS3) protease inhibitors that target important viral proteins necessary for replication. Molecules tested include balapiravir and celgosivir; however, both molecules showed initial promise but were subsequently discontinued due to safety concerns and suboptimal performance in clinical trials. Moreover, vaccines such as Dengvaxia have had their own limitations, relating to serotype-dependent efficacies and increased risks of severe disease among seronegative individuals (Gautam et al., 2024 & Roney, 2024).

Drug discovery and development is very crucial in the health sector because it can outline and formulate new therapeutics for various diseases. Drug discovery has traditionally been a tedious, expensive process, and often drugs do not act in all patients. Over the last few years, Artificial Intelligence has really emerged as a tremendous resource in the field of drug discovery and development, changing the paradigm. AI technologies, such as machine learning and deep learning, have the potential to accelerate the drug discovery process greatly, be more cost-effective, and improve treatment efficacy. These techniques analyze large datasets, find patterns, make predictions about potential drug targets, design new molecules and predict patient responses to treatment (Lee., 2023).

Drug repurposing is considered an important safety strategy for the discovery of new bioactive molecules, mainly due to reducing the complexity and expenses of experimental pharmaceuticals at later stages of clinical trials. Dengue fever outbreak, especially caused by dengue serotype 2, has turned into a world problem, while its effective treatment is being faced as a great challenge (Halder et al., 2022).

This study aims to utilize GNNs to overcome these challenges and enhance the performance of computational drug discovery to find effective and novel inhibitors for DENV-2.

1.2 MOTIVATION OF THE STUDY

The increasing incidence of dengue fever worldwide, especially in regions with inadequate healthcare facilities like Africa, underscores an urgent need for effective antiviral drugs. Of the four serotypes, DENV-2 poses the most concern because of the severe health implications of infections, which include the potentially fatal dengue hemorrhagic fever and dengue shock syndrome. And the lack of an effective antiviral treatment, coupled with the limitations of mosquito control efforts, highlights the critical gap in this area which needs to be addressed.

AI has a high potential in drug discovery as it fast-tracks promising drug candidate identification. In this study, new antiviral development methods were designed to explore how agents might target unique DENV-2 characteristics. Another objective of this research is contributing to the fight against dengue using AI with advanced techniques in improving and speeding up drug discovery. This will eventually provide a way into practical and effective treatments, which could reduce the burden of the disease and, therefore, increase the resiliency of health-care systems in dengue-endemic regions.

This study aims to bridge the gap between technology and pressing healthcare needs. The application of AI in drug development addresses the economic and logistical challenges faced by traditional research approaches besides identification of the potential inhibitors. By giving priority to a scalable, efficient, and precise methodology, this study is intended to reduce the impact of the dengue virus.

1.3 STATEMENT OF THE PROBLEM

The importance of this research arises from the limitations in current computational drug discovery methodologies used for dengue virus. Existing studies, such as (Gautam et al., 2024), rely heavily on conventional machine learning (ML) models like Support Vector Machine (SVM), Artificial Neural Network (ANN), K- Nearest Neighbor (kNN), and Random Forest (RF) for predicting potential inhibitors. Even if these approaches have demonstrated potential, they predominantly use generic regression techniques that do not exploit advanced AI methods such as Graph Neural Networks (GNNs) or transformers, which are specifically tailored for

molecular data. This constraint limits the capacity of these models to fully capture the needed molecular interactions critical for drug discovery.

The article (Costa et al., 2022) uses molecular docking, molecular dynamics (MD), and free energy calculations to predict binding affinities. And these traditional computational techniques also lack the integration of deep learning models, which can significantly accelerate and broaden the scope of interaction predictions.

This study is intended to enhance the prediction performance in computational drug discovery and identification of novel inhibitors against DENV-2 using deep learning techniques, particularly GNNs, designed to directly learn from graph representations of chemical compounds. By addressing the current methodological limitations stated, this approach tries to find a way for more scalable, efficient, and precise framework

1.4 RESEARCH QUESTIONS

In this study, an attempt will be made to answer the following questions:

1. What are the key molecular targets for DENV-2 that can be used for drug discovery efforts?
2. How can AI-driven computational techniques be employed to identify and design effective antiviral compounds targeting DENV-2?
3. Which existing bioactive molecules or drugs have the potential to be repurposed for treating infections caused by DENV-2?
4. What predictive models and AI algorithms are most effective in identifying drug candidates with high affinity and specificity for DENV-2 targets?

1.5 OBJECTIVE

1.5.1 General Objective

The main objective of this study is to repurpose and design novel potential antiviral compounds or drugs for DENV-2 using Graph Neural Networks.

1.5.2 Specific Objectives

- ✓ To gather and prepare datasets of bioactive molecules and structural data of DENV-2 proteins for computational drug discovery workflows.
- ✓ To utilize GNN models to predict protein interactions and viral targets critical for the replication and transmission of DENV-2.
- ✓ To explore the potential of drug repurposing for existing bioactive molecules, assessing their efficacy and safety in inhibiting DENV-2 replication.
- ✓ To design novel potential antiviral compounds for DENV-2 using VGAE model.
- ✓ To explore the drug candidates' affinity and specificity for DENV-2 target protein using molecular docking technique.

1.6 SCOPE OF THE STUDY

The scope of this study is primarily concentrated on the following areas:

- ✓ Identifying key molecular targets within DENV-2 molecular structure by analyzing structural and functional proteins essential for viral replication and infection. Data is sourced from databases such as ChEMBL, Protein Data Bank (PDB), UniProt, and National Center for Biotechnology Information (NCBI).
- ✓ Applying AI-driven computational techniques to screen, repurpose, and design novel antiviral compounds targeting the virus.
- ✓ Applying molecular docking on the identified top drug candidate molecules with DENV-2 target protein to test their affinity and specificity.

Limitations:

- ✓ Results are limited by the in-silico approach, as laboratory validation, clinical testing, and epidemiological assessments are beyond the immediate scope of this study.
- ✓ Findings may not directly apply to other dengue virus serotypes, as the study focuses exclusively on DENV-2.

1.7 SIGNIFICANCE OF THE STUDY

The main aim of the proposed research is to provide new insights into antiviral drug development for DENV-2 by addressing existing challenges related to the design of effective therapies and contribute to state-of-the-art methodologies that show a higher precision compared to those computational methods previously used.

This study is important since it could lead to more effective computational drug discovery against DENV-2, a virus considered as one of the global health concerns in view of the non-availability of effective antiviral therapeutics. Using advanced AI-based techniques like GNNs, this study attempts to enhance the precision in predictions of molecular interaction and, therefore, reduce the difficulty of finding new inhibitors. Furthermore, computational evaluation of the pharmaceutical candidates covering their stability, specificity, binding affinities, and pharmacokinetic properties sets a strong foundation for the discovery of potential antiviral compounds. It also aims at developing scalable and efficient approaches that could be adapted to address broader health issues around the world in the domain of global health.

1.8 EXPECTED OUTCOMES OF THE STUDY

The successful completion of this research aims at discovering novel antiviral agents targeted at DENV-2 using an application of GNNs. Results are expected to contain drug repurposing candidates as well as novel drug formulations. Therefore, the discoveries facilitate the development of efficient therapies against DENV-2, further instigating the vast applications of artificial intelligence against viral diseases.

CHAPTER TWO: LITERATURE REVIEW

2.1 CONCEPTUAL BACKGROUND

Despite advances in the understanding of biological systems and the development of cutting-edge technologies, the process is still long, costly with a high attrition rate (Singh et al., 2023). Traditional de novo drug discovery, which is a lengthy and costly process often hampered by high failure rates and It often requires two to three billion dollars and 10–17 years per new drug (Pinzi et al., 2024).

Computational techniques provide a unique benefit beyond conventional methods by enabling lead compound identification to occur more quickly and accurately than ever before (Naithani & Guleria, 2024). Also drug repurposing can reduce risks and bring drugs to the market in 3–12 years, with an average of $300 million investment (Pinzi et al., 2024). And new approaches, such as artificial intelligence and novel in vitro technologies bring new drugs to patients faster (Singh et al., 2023).

Computer-aided drug discovery has been around for decades, although the past few years have seen a tectonic shift towards embracing computational technologies in both academia and pharma. This shift is largely defined by the flood of data on ligand properties and binding to therapeutic targets and their 3D structures, abundant computing capacities and the advent of on-demand virtual libraries of drug-like small molecules in their billions. Taking full advantage of these resources requires fast computational methods for effective ligand screening (Bender et al., 2021). Moreover, the Deep Docking (DD) platform enables up to 100-fold acceleration of structure-based virtual screening by docking only a subset of a chemical library, iteratively synchronized with a ligand-based prediction of the remaining docking scores. This method results in hundreds- to thousands-fold virtual hit enrichment (without significant loss of potential drug candidates) and hence enables the screening of billion molecule–sized chemical libraries without using extraordinary computational resources (Gentile et al., 2020).

With the advantages of low cost and fast speed, the ML approaches are revolutionizing and strengthening multiple stages of drug discovery, such as target identification, de novo drug design and drug repurposing (Sarkar et al., 2023). In fact, the process of discovering effective new drugs is time-consuming and predominantly the most challenging part of drug development. With the advantages in learning from data, discerning patterns, and making intelligent decisions, ML-based approaches have emerged as versatile tools that can be applied in multiple stages of drug discovery, including drug design, drug screening, drug repurposing and chemical synthesis (Acharjee et al., 2023).

The rapid growth in computing power, amount of data, and advanced algorithms has led to breakthroughs in AI for drug discovery, especially in the application of deep generative models. The models have emerged as high potential tools to transform the design, optimization, and synthesis of small molecules, and macromolecules (Huang et al., 2024). Furthermore, considerable efforts are dedicated to developing models, tools, software and databases based on the core architecture of ML algorithms, to counter the inefficiencies and uncertainties inherent in traditional drug development methods (Acharjee et al., 2023). Applications of these models have already delivered new partially optimized candidate leads, in some cases in less time typically required by conventional sequential approaches (Huang et al., 2024).

Graph neural networks (GNNs) are one of the fastest growing classes of machine learning models. They are of particular relevance for chemistry and materials science, as they directly work on a graph or structural representation of molecules and materials and therefore have full access to all relevant information required to characterize materials (Reiser et al., 2022).

Over recent years, with the development of artificial intelligence (AI) technologies, data-driven methods have rapidly gained in popularity in this field. Among them, graph neural networks (GNNs), a type of neural network directly operating on the graph structure data, have received extensive attention (Bender et al., 2021).

Among different AI models, GNNs have emerged as a powerful class of artificial neural networks designed to process and learn from data structured as graphs. Graphs consist of nodes (vertices) connected by edges (links or relationships), and they are widely used to represent complex relationships and interactions between different entities. This review provided a

comprehensive overview of various GNN models developed for predicting effective drug combinations, examining the limitations and strengths of different models, and comparing their predictive performance (Dara et al., 2022).

GNNs core building blocks are the message passing layers, which are responsible for combining the node and edge information into the node-embeddings. This is done by iteratively aggregating the information of each node with its neighbors' one, thus obtaining a new embedding that is used to update the representation of each node. In particular, the use of GCNs, GATs and GINs in this work is motivated by their success in the prediction of different molecular properties, such as toxicity, hydrophobicity, blood-brain barrier permeability, and solvation free energy (Sarkar et al., 2023).

The rise in the use of GNNs in drug discovery is due to their ability to handle and interpret complex data, such as molecular graphs and biological networks (Artificial Intelligence Review, 2024). GNN-based models have demonstrated high performance and have yielded promising results in various aspects of drug discovery, including virtual screening, molecular property prediction, protein–ligand binding prediction, and drug repurposing (Dara et al., 2022).

The application of computational drug discovery (CDD) methodologies has revolutionized how identification and optimization of therapeutic agents are carried out, in particular against challenging targets such as DENV. This approach, at its very core, uses algorithms that analyze and predict molecular interactions, significantly reducing the time and cost associated with traditional drug development strategies. Machine learning techniques, such as Random Forest, Support Vector Machines, and deep neural networks, have been highly effective in handling complex biological data (Dara et al., 2021).

As stated in (Gallo et al., 2023) CDD's application in antiviral drug discovery has focused on the DENV envelope (E) protein, a critical target for viral entry into host cells. By employing molecular docking and dynamic simulations, researchers have been able to identify binding hotspots within the hydrophobic pocket of the E protein. These approaches facilitated the discovery of compound 16a, a pyrimidine derivative with robust antiviral activity and favorable pharmacokinetic properties. Such findings underscore the role of structure-guided drug design in addressing unmet medical needs for vector-borne diseases.

Despite such advances, a lot of challenges still stands in the way of applying computational tools to drug discovery. Low solubility, high cytotoxicity, and suboptimal pharmacokinetics remain among the barriers that in silico hits have to overcome to become viable clinical candidates (Acharjee et al., 2023).

The versatility of CDD extends beyond single-target inhibitors to multi-target approaches, addressing the serotype diversity of pathogens like DENV. Tools such as Quantitative Structure-Activity Relationship (QSAR) modeling and virtual screening have identified potential broad-spectrum inhibitors that mitigate serotype-related challenges. This multidimensional application of CDD highlights its potential to streamline the discovery of cross-reactive therapeutic agents for complex diseases (Neuner et al., 2020).

Besides, aided by modern computational power, organized studies of molecular modifications targeting the improvement of pharmaceutical properties have been facilitated with tools such as SwissBioisostere. Introduced changes in structure at key positions specifically, at the 6-position of pyrimidine derivatives revealed significant improvements in binding affinity and reduced off-target interactions. Such progress testifies to the precision of CDD methodologies in tailoring compounds for better performance, as substantiated by the study of (Gallo et al., 2023).

In summary, computational drug discovery is a revolutionary technique to combat various viral and other type of diseases. Through the integration of predictive algorithms with experimental protocols, CDD offers a fast track to the design and improvement of therapeutic candidates. And recently much emphasis has been placed on using deep learning algorithms like GNNs for CDD.

2.2 RELATED WORKS

AI-driven methodologies have revolutionized the field of drug discovery by efficiently identifying targets and optimizing candidate molecules. Deep learning (DL) models, including Convolutional Neural Networks (CNNs), have achieved more than an 85% accuracy in the prediction of ligand-receptor interactions, significantly enhancing the docking simulations of virus inhibitors. Generative AI models, such as Generative Adversarial Networks, boast a 90%

success rate in generating novel molecular candidates with desirable Absorption Distribution Metabolism Excretion (ADME) properties, hence speeding up the discovery pipeline. Also, omics data have been mined with knowledge graphs, which have more than 80% success rates in identifying potential drug targets upon experimental validation (Acharjee et al., 2023).

The effectiveness of CDD methods based on artificial intelligence continues through the clinical stages, with an 80–90% striking success rate in Phase I trials by AI-identified compounds targeting the dengue virus, compared to traditional methods. Furthermore, these compounds showed approximately 40% success rates in Phase II trials, which is in line with historical industry benchmarks and demonstrates the potential of AI in optimizing candidate selection (Acharjee et al., 2023).

(Costa et al., 2022) employed a computational approach to explore allosteric inhibitors for the NS2B/NS3 protease of DENV, a critical target for viral replication. The study identified compounds such as ZINC000001680989 and ZINC000001679427, which demonstrated high stability in molecular dynamics simulations and formed key hydrogen bond interactions with residues Asn152, Leu149, and Ala164. These compounds showed binding energies superior to traditional inhibitors, suggesting their potential as effective therapeutic agents.

(Gautam et al., 2024) developed a predictive algorithm by using QSAR coupled with machine learning methodologies such as SVM, ANN, kNN, and RF. The generated models showed PCC values of 0.71 for the training/test dataset and 0.81 for the independent validation dataset. These results demonstrate the success of computational methods in predicting potent dengue inhibitors, which ultimately led to the finding of promising candidates such as goserelin and nafarelin for drug repurposing.

(Huang et al, 2024) highlighted the transformative potential of AI in the biopharmaceutical sector, including its application to DENV. AI-driven virtual screening and QSAR modeling have significantly reduced the time and cost associated with experimental drug screening. For instance, generative adversarial networks (GANs) combined with reinforcement learning improved lead optimization processes by 30% compared to traditional methods. Moreover, AI-assisted ADMET predictions have achieved accuracies as high as 93% in hepatotoxicity assessments, emphasizing the role of computational intelligence in enhancing drug safety profiles.

Importance of Graph Neural Networks (GNNs) for CDD:

Graph Neural Networks (GNNs) have demonstrated exceptional accuracy in various tasks within Computational Drug Discovery (CDD). For drug-target interaction (DTI) prediction, the DGraphDTA model achieved state-of-the-art results on benchmark datasets. On the Davis dataset, it reported a concordance index (CI) of 0.904 and a mean squared error (MSE) of 0.202, while on the KIBA dataset, it achieved a CI of 0.904 and an MSE of 0.126. Additionally, the model exhibited a Pearson correlation coefficient of up to 0.903, highlighting its strong reliability in predicting binding affinities between drugs and target proteins (Jiang et al., 2020).

As showed in the article (Jiang et al., 2020) the GNN models effectively represent molecular structures as graphs, where atoms and bonds are nodes and edges, respectively and capturing complex relationships in a highly interpretable manner. These models excel in predicting drug-target interactions, molecular dynamics, and bioactivity, making them invaluable for virtual screening and structure-based drug design. The ability of GNNs to learn from and integrate diverse chemical and biological datasets make them a critical tool in advancing CDD.

Table 1 Summary for Literature Review

No.	Authors	Dataset	Method	Modality	Classification Type	Gap or limitation
1	(Costa et al., 2022)	Molecular dynamics data for NS2B/NS3 protease inhibitors	Computational exploration using molecular docking and dynamics simulations	Molecular docking	Binary	Lack of utilizing Graph Neural Networks (GNNs) for molecular modeling.
2	(Gautam et al.,2024)	QSAR datasets for predictive modeling	QSAR, SVM, ANN, kNN, and Random Forests for predicting dengue inhibitors	Drug repurposing	Binary & Multiclass	Lack of utilizing Graph Neural Networks (GNNs) for molecular modeling.
3	(Huang et al., 2024)	ADMET and drug property prediction datasets	GANs combined with reinforcement learning for lead optimization, ADMET predictions	Virtual screening, ADMET	Binary & Multiclass	Lack of utilizing Graph Neural Networks (GNNs) for molecular modeling.

CHAPTER THREE: METHODOLOGY

This methodology will outline the approach in using deep learning techniques for drug discovery, particularly on targeting DENV-2. The research will proceed into a few stages like data collection, preprocessing of data, deep learning model development, and performance evaluation. The aim is to come out with possible drug candidates against DENV-2 using Graph Neural Networks (GNNs).

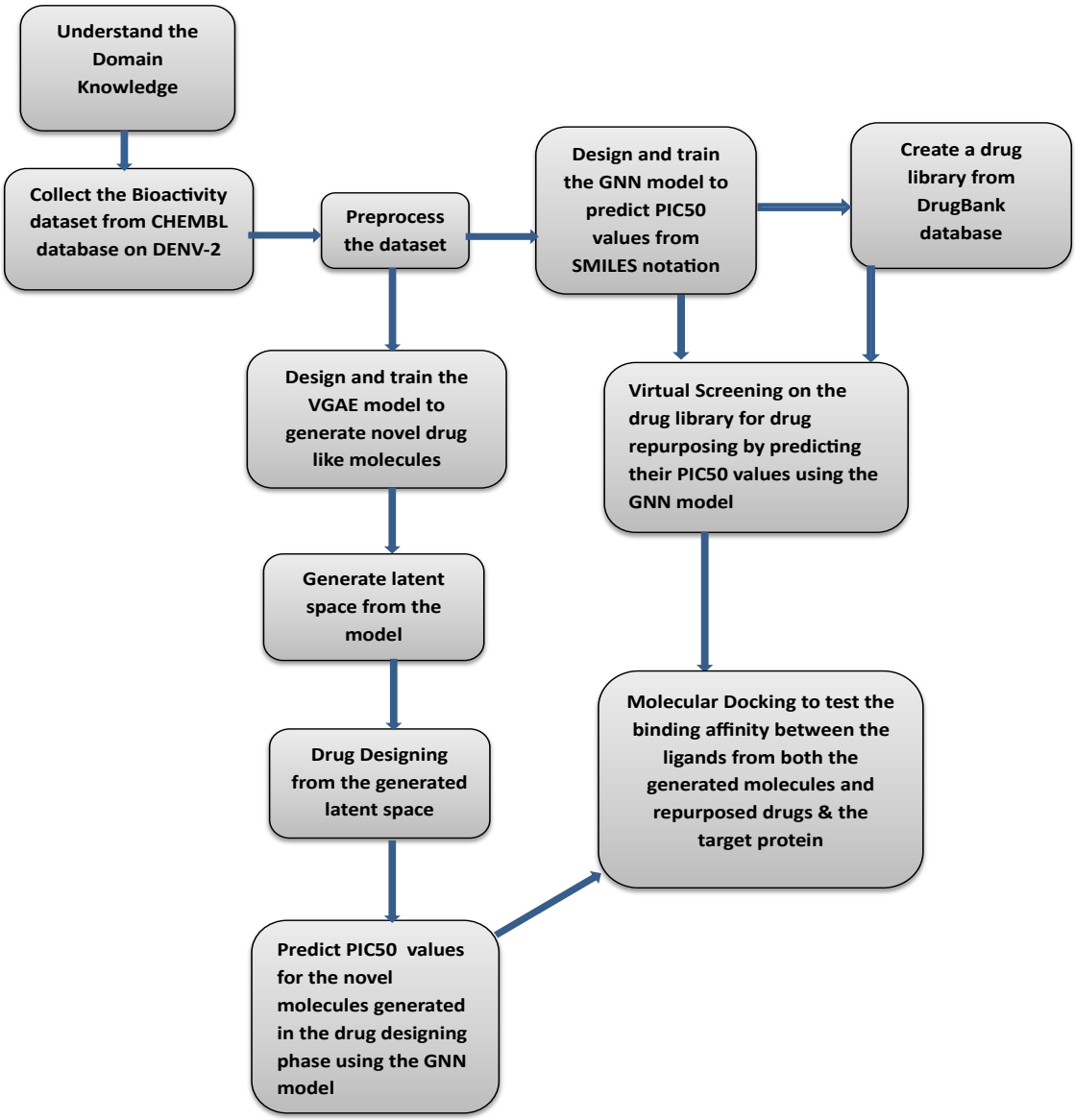


Figure 1: Overall Methodology

3.1DATASETS COLLECTION

The first step of the pipeline is to develop a high-quality dataset that will allow for efficient training of deep learning models. This step involves the extraction of comprehensive information about the target protein, Dengue virus protein, from the ChEMBL database using its UniProt ID.

ChEMBL is a manually curated chemical database of bioactive molecules with drug inducing properties. It is maintained by the European Bioinformatics Institute (EBI), of the European Molecular Biology Laboratory (EMBL), based at the Wellcome Trust Genome Campus, Hinxton, UK. The ChEMBL database contains compound bioactivity data against drug targets. Data can be filtered and analyzed to develop compound screening libraries for lead identification during drug discovery (Zdrazil et al., 2024).

(The UniProt Consortium., 2025) UniProt is a freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from the research literature. It is maintained by the UniProt consortium, which consists of several European bioinformatics organisations and a foundation from Washington, DC, USA.

UniProt was utilized to validate and map the DENV-2 NS3 target protein to its corresponding bioactivity data in ChEMBL through the UniProt Accession ID P29990. This ensures biological relevance and specificity in dataset curation.

The extracted data gives an overview of the characteristics of the protein, including the organism of origin, the target type, and its unique ChEMBL ID. These data represent the biological basis for the workflow of drug discovery in order to ensure a precise understanding of the target protein and its role in the disease.

Bioactivity data is simultaneously collected with its focus on the IC50 values for different compounds interacting with the target protein. The IC50 value gives the concentration of a compound that could inhibit 50% of the protein activity; this forms a critical measure of compound efficiency. To make these more interpretable and also be machine learning-friendly, these IC50 values are converted into pIC50 using negative logarithm function. This logarithmic transformation normalizes the IC50 values to makes them more suitable to modeling by reducing skewness and improving numerical stability.

Lastly, molecular data in SMILES string representation is converted into molecular graphs using RDKit. This transformation will represent the structural and chemical properties of each compound, with atoms as nodes and bonds as edges. This graph-based representation is basic to GNNs for effective learning of relationships within molecular structures and predicting bioactivity.

3.2 PROPOSED MODEL

The proposed methodology describes a comprehensive computational drug discovery pipeline using state-of-the-art graph neural networks for the identification and optimization of potential drug candidates targeting the Dengue virus protein. The pipeline is divided into different well-defined stages, each dedicated to a critical aspect of the drug discovery process: dataset preparation, model training, virtual screening, and drug design. In summary, it focuses on how to choose drug candidates with better effects but lower toxicity using advanced frameworks of machine learning, such as Graph Attention Networks (GAT) in compound-protein interaction predictions, Graph Convolutional Networks (GCN) in virtual screening, and Variational Graph Autoencoders (VGAE) in drug design.

3.2.1 Feature Engineering

The feature engineering step polishes the molecular graphs from the previous step by adding more structural and physicochemical features. This makes the data complete and provides the GNN with all the information that may be necessary for better predictions. Each molecular graph

is augmented with node features, such as atomic numbers representing the types of atoms present in the molecule, and bond connectivity encoding the relationships between atoms. Beyond this basic graph structure, global molecular properties such as molecular weight, hydrogen bond donor and acceptor count, and lipophilicity are integrated. These properties provide additional context that allows the model to learn about chemical behaviors important for predicting molecular bioactivity.

Once the graphs are enriched with these features, they are encoded into the PyTorch Geometric Data format. This standardized format makes the data compatible with advanced GNN architectures, enabling efficient processing and training. By incorporating both local (atomic-level) and global (molecular-level) features, the feature engineering stage ensures the GNN can extract subtle patterns and relationships within the data. This detailed representation improves the model's ability to generalize and make accurate predictions about the bioactivity of novel compounds.

3.2.2 GNN Model Development

The model development stage mainly focuses on the architecture for GNN, which would be used to predict the bioactivity of compounds with respect to negative logarithm of Half-Maximal Inhibitory Concentration (pIC50) values against the target protein Dengue virus protein. In compound-protein interaction prediction, Graph Attention Networks (GAT) are applied. GATs are chosen because of their ability to assign different attention scores to the neighboring atoms and bonds, allowing the model to focus on the most relevant parts of the graph. This mechanism is critical in predicting how a compound will bind to the target protein.

Graph Convolutional Networks (GCNs) are used for virtual screening because they can process big data very efficiently. The GCN operates by aggregating information of neighboring nodes, hence the suitability for handling big compound libraries and classifying compounds into active or inactive with efficiency, according to their predicted bioactivity. GCNs are appropriate for the fast screening of big libraries and identification of those compounds that would probably be efficient while interacting with the target protein.

3.2.3 VGAE Model Development

Variational Graph Autoencoders (VGAEs) are in use for the drug design stage following virtual screening. VGAE model is useful for generating novel molecules. They learn a probabilistic representation of molecular graphs, allowing them to sample the latent space of molecular features and generate new molecules possessing properties similar to those known to be active. VGAEs are important for the exploration of chemical space and the design of novel drug candidates with better bioactivity or pharmacokinetic properties.

3.2.4 Model Training

The next step is to train this architecture with a supervised learning method, in which the model will have learned the pIC50 estimation for the compounds with given molecular graphs. It tries to minimize the loss function to bring the predicted outputs close to the actual values. The dataset is prepared in batches using the PyTorch Geometric DataLoader to work with large datasets.

The Adam optimizer updates model parameters for effective and fast convergence. The model is trained for many epochs, during which time it learns to recognize patterns in the molecular graphs that relate to bioactivity. It trains in a computationally efficient manner while maximizing the generalization capability of the model for new, unseen compounds.

3.2.5 Model Evaluation

The performance of the model is then verified using various evaluation metrics. Some of the most important performance metrics, such as Mean Squared Error (MSE), Huber Loss, Mean Absolute Error (MAE), and R^2 Score, are used to check how well the model is performing in predicting values. MSE provide insights into the magnitude of prediction errors while MAE measures the average absolute error, helping to spot how consistent the predictions are. Huber loss complements these metrics by combining their strengths: it behaves like MSE for small residuals (errors close to zero), ensuring sensitivity to subtle deviations, and transitions to MAE-like behavior for larger errors beyond a predefined threshold. R^2 Score is very useful because it

gives the proportion of variance in the pIC50 values explained by the model. The higher the R^2 score, the closer it gets to 1.0, indicating that the model can explain a greater percentage of the variance and hence possesses strong predictive power.

3.2.6 Create a drug library

Building a drug library is needed for the virtual screening pipeline that enables repurposing of existing small-molecule drugs against the Dengue virus NS3 protein. This library is derived from the DrugBank database.

DrugBank, as detailed in (Wishart et al., 2018) is a comprehensively annotated resource that integrates chemical, pharmacological, and therapeutic data for drug discovery. It provides access to FDA-approved drugs, experimental compounds, and bioactive molecules, offering curated annotations on drug-target interactions, ADMET properties (absorption, distribution, metabolism, excretion, toxicity), and pharmacokinetic profiles. The database supports computational workflows such as virtual screening and drug repurposing by enabling standardized molecular representations (e.g., SMILES notation) and prioritizing drug-likeness criteria (e.g., Lipinski's Rule of Five compliance).

The process begins with querying DrugBank for small-molecule drugs (molecular weight < 500 Da) to ensure drug-likeness and bioavailability. Compounds are filtered to retain those with valid SMILES (Simplified Molecular Input Line Entry System) notation, which encodes their molecular structures in a machine-readable format. Duplicate entries are removed to ensure chemical uniqueness and simplify computational processing. This curated dataset forms the foundation of the drug library, comprising diverse scaffolds with potential cross-reactivity against the Dengue virus target.

The resulting library serves as a virtual screening dataset for the trained GNN model which predicts pIC50 values using their Simplified Molecular Input Line Entry System (SMILES) notation. By applying the model to this library, the pipeline identifies existing drugs with high inhibitory potential against NS3, enabling drug repurposing. This approach significantly reduces

development time and costs, as prioritized candidates can advance directly to experimental validation or clinical trials.

3.2.7 Virtual Screening

During the virtual screening stage, the trained model is used to predict the pIC50 values of compounds from a drug molecule library. New compounds are represented as SMILES strings, which are converted into molecular graphs. These graphs are processed with the trained GNN model to predict the bioactivity of each compound. Hence, only the compounds with the highest predicted pIC50 values are shortlisted for further analysis, as they can show the strongest inhibitory potential against the Dengue virus NS3 protein.

3.2.8 Drug Design

During this stage, the generation of novel molecular structures is performed with the use of Variational Graph Autoencoders (VGAE). These VGAEs encode molecular graphs into a latent space and sample from this space to generate new molecular structures similar to the active compounds used in training. This allows the model to explore chemical space and come up with new compounds that could have better binding affinities or improved pharmacokinetic properties. VGAEs are helpful in the design of compounds with specific properties such as enhanced bioactivity, selectivity, or lowered toxicity. By applying such a generative model, it is possible to suggest new candidates of drugs that may prove more active or with a superior therapeutic profile.

3.2.9 Molecular Docking

Molecular docking is another important step in the drug discovery pipeline. It is used to predict the preferred orientation and binding affinity of a ligand (drug candidates) when interacting with its target protein. In this case it evaluates the molecular interactions between the Dengue virus NS3 protein and the shortlisted compounds from virtual screening or those generated via drug design. By simulating the binding process, molecular docking provides insights into the

structural compatibility and energetics of compound-protein complexes, enabling the prioritization of candidates with optimal binding characteristics for experimental validation.

The docking process employs algorithms that explore the conformational space of ligands within the target protein's active site, generating multiple binding poses ranked by scoring functions. These scoring functions estimate binding affinity using metrics such as docking scores or binding energy, which correlate with the likelihood of stable interactions. This step is particularly valuable for identifying key residues in the NS3 protein involved in compound binding which are critical for inhibitory activity. And for this purpose, we use online docking tool called AutoDock vena from SwissDock website.

SwissDock is a web-based platform for protein-ligand docking simulations, designed to predict the preferred binding orientation and affinity of small molecules to target proteins. And specifically, AutoDock Vina is the docking engine used. It uses a widely known algorithm for calculating binding poses and energies (Bugnon et al., 2024) & (Eberhardt et al., 2021).

The output of molecular docking includes quantitative metrics such as binding energy (kcal/mol) and qualitative visualizations of ligand-protein complexes. Lower binding energy in negative values (<= -5 kcal/mol) indicate stronger interactions or binding affinity.

By bridging computational predictions with biophysical validation, molecular docking ensures that proposed drug candidates are both theoretically potent and structurally viable. This significantly enhances confidence in the pipeline's outputs, directing resources toward compounds most likely to succeed in experimental assays and subsequent stages of drug development.

Generally, this in silico pipeline presents a robust framework for identifying and optimizing antiviral drug candidates targeting the Dengue virus NS3 protein by integrating compound-protein interaction prediction and virtual screening using GNNs, and for drug designing using VGAE. The performance of the model will be evaluated using several key metrics with the view that the predictions are accurate and reliable. This approach accelerates the early stage of drug

discovery by reducing reliance on experimental methods, hence allowing for the efficient identification and optimization of antiviral drug candidates.

3.3 Evaluation Methods

The proposed computational drug discovery framework is evaluated through a series of rigorous tests for the reliability, accuracy, and functional applicability of its predictions. This evaluation approach incorporates several key metrics, each targeted at different performance dimensions of the model.

3.3.1 Training Accuracy

Training accuracy assesses how effective the GNN models are in predicting outcomes on the training dataset. It is a measure of the ability of the models to learn patterns and relationships in data, such as molecular interactions, ADMET profiles, and binding affinities. High training accuracy, therefore, means that the models have been able to learn these underlying features in the training dataset.

3.3.2 Validation Accuracy

Validation accuracy describes the performance of a model on a separate and distinct validation dataset excluded in training. This metric examines how generalizable the models are and confirms their effectivity on data that is unused so far. High validation accuracy can give an indication about how good the models must have learned patterns from past compound sets to apply that for testing new compounds toward essential pharmaceutical predictions.

3.3.3 Training Loss

The training loss gives insight into the learning dynamics of the GNN models and quantifies the error produced during training. It measures how well the model minimizes the difference between predicted and actual values, usually through loss functions like Mean Squared Error

(MSE). This metric helps identify possible over-fitting or under-fitting, guiding changes in model parameters or architecture through monitoring training loss.

3.3.4 Validation Loss

The validation loss is the quantification of the error rate of the models when evaluated on the validation dataset. The metric gives a real sense of how good the models are in effectiveness with unseen data and shows their strength and reliability. Lower validation loss could indicate that the models are better at generalizing away from the training dataset.

3.4 Development Tools

The development and implementation of the GNN-based framework for the computational drug discovery involve a use of different tools for the specific stages of the model development process as stated below.

3.4.1 Hardware Tools

To design the graph neural network models, GPUs from **Google Colab** is utilized. These GPUs offer powerful computational resources that facilitate the construction and training of deep learning models. Use of cloud-based GPUs from Google Colab removes the bottlenecks of handling large-scale molecular and protein graph data, enabling the faster execution of complex tasks in training, data augmentation, and molecular docking simulations. And this eliminates the limitations of using local hardware by providing flexibility and scalability throughout the model development process.

3.4.2 Software Tools

Python is the main programming language, since it has a large ecosystem of libraries and frameworks for machine learning and data science. Advanced deep learning frameworks like **TensorFlow** and **PyTorch** will be used to design, train, and evaluate the models. TensorFlow

brings the best scalability with a flexible ecosystem of building complex GNN architectures, while PyTorch has dynamic computation graph capabilities that are essential for experimenting with novel designs. Additional tools such as **RDKit** support the generation and manipulation of molecular descriptors.

CHAPTER FOUR: IMPLEMENTATION

4.1 DATA-COLLECTION AND PREPROCESSING

First, we Retrieved the bioactivity dataset From ChEMBL Database on DENV-2 target protein called NS3. Then we fetched its information using the Uniprot Accession ID P29990, corresponding to the target protein CHEMBL5980. This protein is derived from Dengue virus type 2 (strain Thailand/16681/1984) (DENV-2). The collected dataset contains 236 rows of data. Then we fetch some important variable of the data.

IC50 is a standard unit used in biochemical assays for inhibition. It stands for Inhibitory concentration at 50 or Half maximal inhibitory concentration. The lower value indicates higher potency of a drug. **pIC50**, that is = -log10(IC50). It makes the opposite of IC50, meaning higher pIC50 means higher potency of a drug. (Kazakova, R. R., & Masson, P., 2022).

A **low potency** compound typically has an IC50 value greater than 10 µM, corresponding to a pIC50 value less than 5. These compounds exhibit weak inhibition and are generally considered less effective. **Moderate potency** compounds have IC50 values ranging between 1 µM and 10 µM, with corresponding pIC50 values between 5 and 6, indicating reasonable inhibition and moderate drug effectiveness. In contrast, **high potency** compounds are defined by IC50 values below 1 µM (often in the nanomolar range), corresponding to pIC50 values greater than 6. These compounds demonstrate strong inhibition and are considered highly potent drugs. (Pérez, & et al., 2013).

Data Processing and Filtration

There is no missing value in the dataset. But there are 2 non-nM entries and 43 duplicate rows. After removing them from the dataset shape is 192 rows by 11 columns.

Fetch compound data from ChEMBL

Now we are going to fetch compound ChEMBL IDs and structures for the compounds linked to our filtered bioactivity data.

Extract the SMILES (Simplified Molecular Input Lines Entry System)

The next figure shows the extracted SMILES structure for each compounds in the dataset.

	molecule_chembl_id	smiles
0	CHEMBL164	O=c1c(O)c(-c2cc(O)c(O)c(O)c2)oc2cc(O)cc(O)c12
1	CHEMBL320007	O=c1c2ccccc2sn1-c1ccc([N+](=O)[O-])cc1
2	CHEMBL1324	Cc1ccc(C(=O)c2cc(O)c(O)c([N+](=O)[O-])c2)cc1
3	CHEMBL379110	COc1cc(O)c(C(=O)[C@@H]2[C@H](CC=C(C)C)C(C)=CC[...
4	CHEMBL506247	O=C(OC[C@H]1O[C@@H](OC(=O)c2cc(O)c(O)c(OC(=O)c...

Figure 2: Fetched SMILES data

Merged datasets with desired columns (molecule ID, IC50, Units, SMILES and calculated PIC50 value from IC50)

	molecule_chembl_id	IC50	units	smiles	pIC50
0	CHEMBL4452943	1700.0	nM	COc1ccc(C[C@H]2C(=O)N[C@@H](CCCCN)C(=O)N(C)[C@...	5.769551
1	CHEMBL4440832	4200.0	nM	O=C(Nc1nc2cc(O)c(O)cc2s1)c1ccccc1Oc1ccc2cc(O)c...	5.376751
2	CHEMBL3344303	14400.0	nM	N=C(N)Nc1ccc(OC(=O)c2ccc(NC(=N)N)cc2)cc1	4.841638
3	CHEMBL506092	12200.0	nM	N=C(N)NCCC[C@@H](C=O)NC(=O)[C@H](CCCCN)NC(=O)[...	4.913640
4	CHEMBL4437334	1800.0	nM	O=C1C=C/C(=N/Nc2ccc([N+](=O)[O-])cc2Cl)C=C1C(=...	5.744727
...	...	...	...	...	...
187	CHEMBL5421390	200.0	nM	CN([C@@H](Cc1ccc(C(=N)N)cc1)C(=O)N[C@@H](CCCCN...	6.698970
188	CHEMBL5421971	120.0	nM	CN([C@@H](Cc1ccc(C(=N)N)cc1)C(=O)N[C@@H](CCCCN...	6.920819
189	CHEMBL5413511	89.0	nM	CN([C@@H](Cc1ccc(C(=N)N)cc1)C(=O)N[C@@H](CCCCN...	7.050610
190	CHEMBL164	8460.0	nM	O=c1c(O)c(-c2cc(O)c(O)c(O)c2)oc2cc(O)cc(O)c12	5.072630
191	CHEMBL3819036	7550.0	nM	O=C(CCCCCCCCc1ccc(O)cc1)c1c(O)cc(O)cc1O	5.122053

192 rows × 5 columns

Figure 3: Merged dataset with SMILES and PIC50 values

Distribution of the dataset Based on pIC50

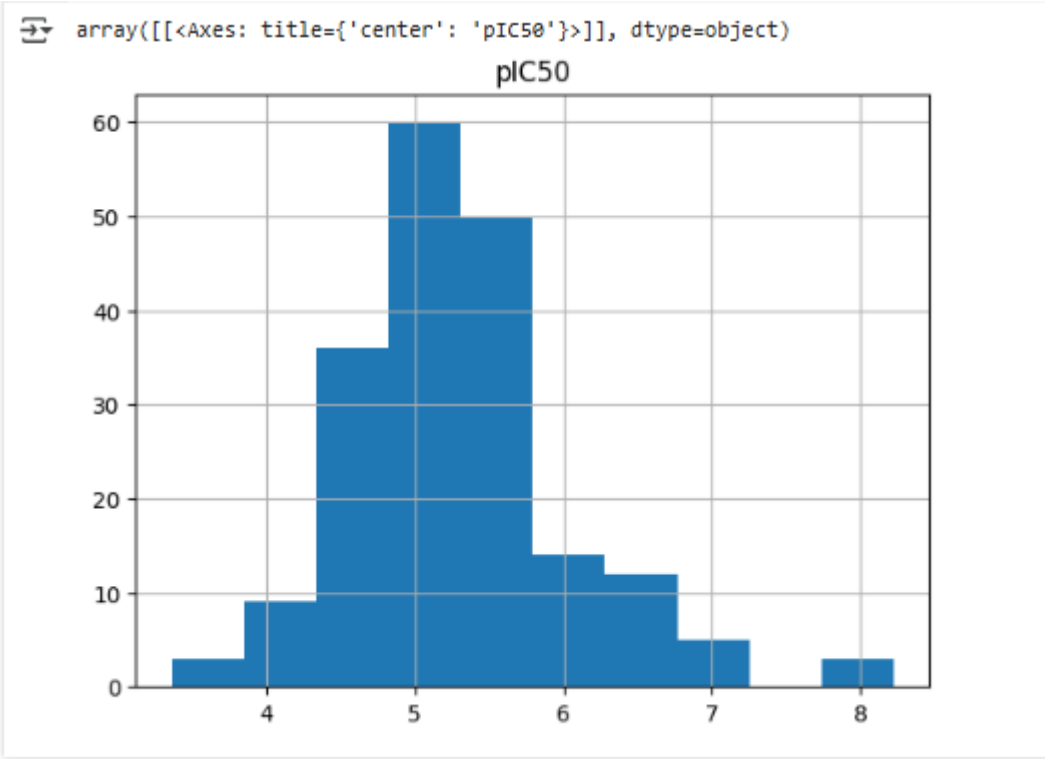
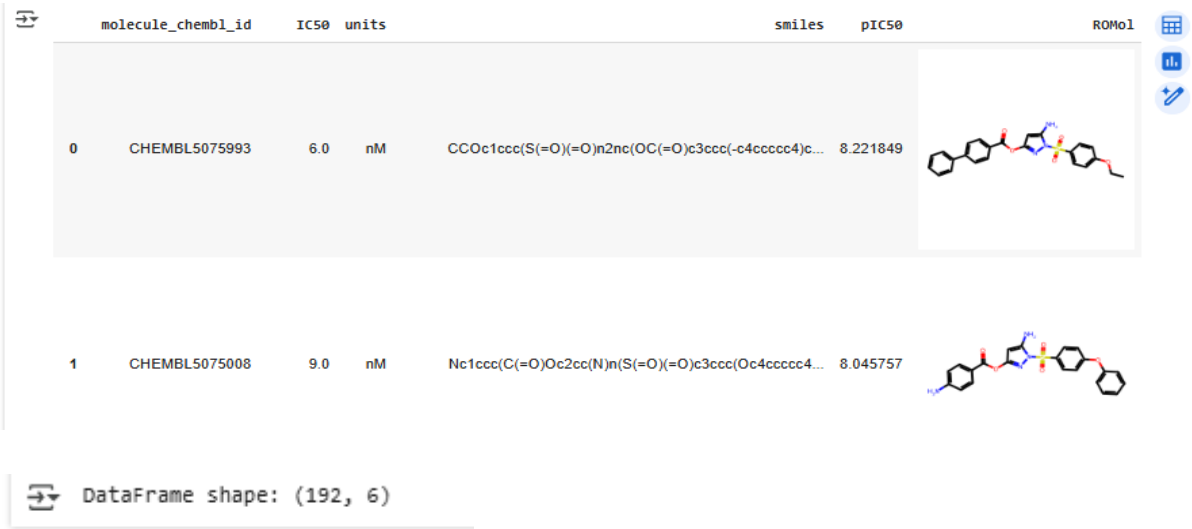


Figure 4: Distribution of the dataset based on PIC50

Now we add a column for RDKit molecule objects to our DataFrame and look at the structures of the molecules with their pIC50 values. Then we save the dataset for next sessions.



	molecule_chembl_id	IC50	units	smiles	pIC50	ROMol
0	CHEMBL5075993	6.0	nM	CCOc1ccc(S(=O)(=O)n2nc(OC(=O)c3ccc(-c4ccccc4)c...	8.221849	
1	CHEMBL5075008	9.0	nM	Nc1ccc(C(=O)Oc2cc(N)n(S(=O)(=O)c3ccc(Oc4ccccc4...	8.045757	

DataFrame shape: (192, 6)

Figure 5: Merged dataset with SMILES, PIC50 & RDkit molecule objects and its shape

4.2 COMPOUND FILTRATION BASED ON ADME AND LEAD-LIKENESS

(Lin & et al., 2003) in a virtual screening (molecular docking), our aim is to predict whether a compound might bind to a specific target. However, if we want to identify a new drug, it is also important that this compound reaches the target and is eventually removed from the body in a favorable way. Therefore, we should also consider whether a compound is actually taken up into the body and whether it is able to cross certain barriers in order to reach its target. Is it metabolically stable and how will it be excreted once it is not acting at the target anymore? These processes are investigated in the field of pharmacokinetics. In contrast to pharmacodynamics ("What does the drug do to our body?"), pharmacokinetics deals with the question "What happens to the drug in our body?".

I. ADME – (Absorption, Distribution, Metabolism, and Excretion**):**

Pharmacokinetics are mainly divided into four steps: Absorption, Distribution, Metabolism, and Excretion. These are summarized as ADME. Often, ADME also includes Toxicology and is thus referred to as ADMET or ADMETox. Below, the ADME steps are discussed in more detail

Absorption: The amount and the time of drug-uptake into the body depends on multiple factors which can vary between individuals and their conditions as well as on the properties of the substance. Factors such as (poor) compound solubility, gastric emptying time, intestinal transit time, chemical (in-)stability in the stomach, and (in-)ability to permeate the intestinal wall can all influence the extent to which a drug is absorbed after e.g. oral administration, inhalation, or contact to skin.

Distribution: The distribution of an absorbed substance, i.e. within the body, between blood and different tissues, and crossing the blood-brain barrier are affected by regional blood flow rates, molecular size and polarity of the compound, and binding to serum proteins and transporter enzymes. Critical effects in toxicology can be the accumulation of highly polar substances in fatty tissue or crossing of the blood-brain barrier.

Metabolism: After entering the body, the compound will be metabolized. This means that only part of this compound will actually reach its target. Mainly liver and kidney enzymes are responsible for the break-down of xenobiotics (substances that are extrinsic to the body).

Excretion: Compounds and their metabolites need to be removed from the body via excretion, usually through the kidneys (urine) or in the feces. Incomplete excretion can result in accumulation of foreign substances or adverse interference with normal metabolism.

II. Lead-likeness and Lipinski's rule of five (Ro5)

Molecular weight (MWT) <= 500 Da

Number of hydrogen bond acceptors (HBAs) <= 10

Number of hydrogen bond donors (HBD) <= 5

Calculated LogP (octanol-water coefficient) <= 5

Additionally -LogP is also called partition coefficient or octanol-water coefficient. It measures the distribution of a compound, usually between a hydrophobic (e.g. 1-octanol) and a hydrophilic (e.g. water) phase.

Hydrophobic molecules might have a reduced solubility in water, while more hydrophilic molecules (e.g. high number of hydrogen bond acceptors and donors) or large molecules (high molecular weight) might have more difficulties in passing phospholipid membranes. (Lipinski, C. A., 2004).

Calculate the Molecular Properties needed for Ro5 from RDKit molecular Objects

	molecular_weight	n_hba	n_hbd	logp
0	463.120192	8	1	3.98720
1	450.099791	9	2	3.29600
2	419.093977	7	1	3.58850
3	435.984088	8	2	2.26620
4	795.237258	8	6	4.04487
...	...	...	...	...
187	607.185569	8	2	6.87200
188	428.083078	4	1	4.44120
189	389.256609	4	0	5.14910
190	404.267508	4	0	4.60580
191	330.012415	8	2	1.68960

192 rows × 4 columns

Figure 6: Calculated molecular properties needed for Ro5 from RDkit molecular objects

Then using those properties, we checked for each compound in the dataset if they fulfill Ro5 properties. And from the original 192 compounds only 95 compounds fulfilled the rule.

Save the filtered datasets with fulfilled Ro5 properties

	molecule_chembl_id	IC50	units	smiles	pIC50	ROMol	molecular_weight	n_hba	n_hbd	logp	molecular_weight	n_hba	n_hbd	logp	ro5_fulfilled
0	CHEMBL5075993	6.0	nM	CCOc1ccc(S(=O)(=O)n2nc(OC(=O)c3ccc(-c4ccccc4)c...	8.221849		463.120192	8	1	3.9872	463.120192	8	1	3.9872	True
1	CHEMBL5075008	9.0	nM	Nc1ccc(C(=O)Oc2cc(N)n(S(=O)(=O)c3ccc(Oc4ccccc4...	8.045757		450.099791	9	2	3.2960	450.099791	9	2	3.2960	True
2	CHEMBL5092085	12.0	nM	Nc1cc(OC(=O)c2ccc(-c3ccccc3)cc2)nn1S(=O)(=O)c1...	7.920819		419.093977	7	1	3.5885	419.093977	7	1	3.5885	True
3	CHEMBL5081752	74.0	nM	Nc1ccc(C(=O)Oc2cc(N)n(S(=O)(=O)c3ccc(Br)cc3)n2...	7.130768		435.984088	8	2	2.2662	435.984088	8	2	2.2662	True

Figure 7: Filtered datasets with fulfilled Ro5 properties

4.3 REMOVAL OF UNWANTED COMPOUNDS FROM THE RO5 PROPERTIES FULFILLED DATASET

There are some substructures we prefer not to include into our screening library. Before going to the implementation, let's see the characteristics of such unwanted substructures.

Unwanted substructures

Substructures can be unfavorable, e.g., because they are toxic or reactive, due to unfavorable pharmacokinetic properties, or because they likely interfere with certain assays. Filtering unwanted substructures can support assembling more efficient screening libraries, which can save time and resources.

(Brenk & *et al., 2008)* have assembled a list of unfavorable substructures to filter their libraries used to screen for compounds to treat neglected diseases. Examples of such unwanted features

are nitro groups (mutagenic), sulfates and phosphates (likely resulting in unfavorable pharmacokinetic properties), 2-halopyridines and thiols (reactive).

Pan Assay Interference Compounds (PAINS)

PAINS are compounds that often occur as hits in HTS even though they actually are false positives. PAINS show activity at numerous targets rather than one specific target. Such behavior results from unspecific binding or interaction with assay components. (Baell, J. B., & Holloway, G. A., 2010) focused on substructures interfering in assay signaling. They described substructures which can help to identify such PAINS and provided a list which can be used for substructure filtering.

In RDKit, PAINS (Pan Assay Interference Compounds) refers to a list of substructures that are known to cause false positives in drug discovery assays. RDKit provides tools to filter out molecules containing these unwanted substructures within the "FilterCatalog" module.

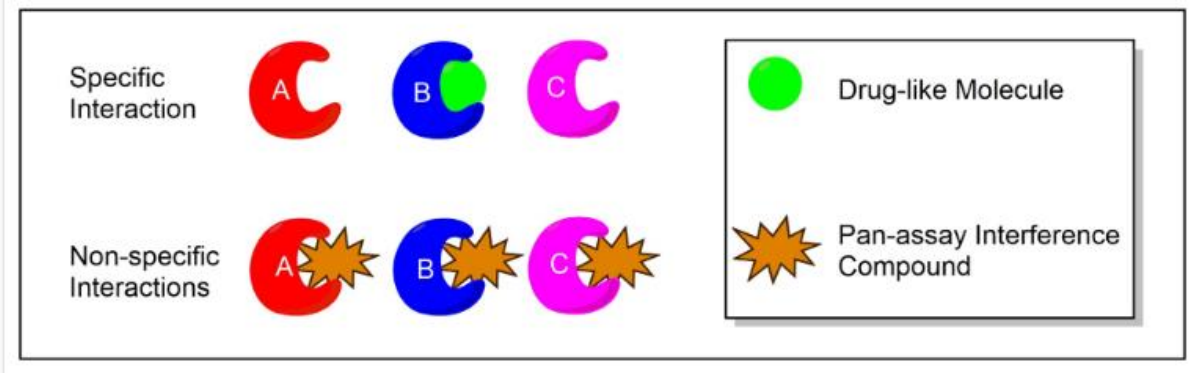


Figure 8: Specific and unspecific binding in the context of PAINS

Filter for PAINS from RDkit library

Using the RDkit library we further filtered for PAINS substructures from the remaining dataset.

Filter for unwanted substructures and save the last and clean NS3 dataset

Here, we use the list provided in the supporting information from (Brenk & *et al., 2008).* Finally, our last clean dataset contains 37 compounds.

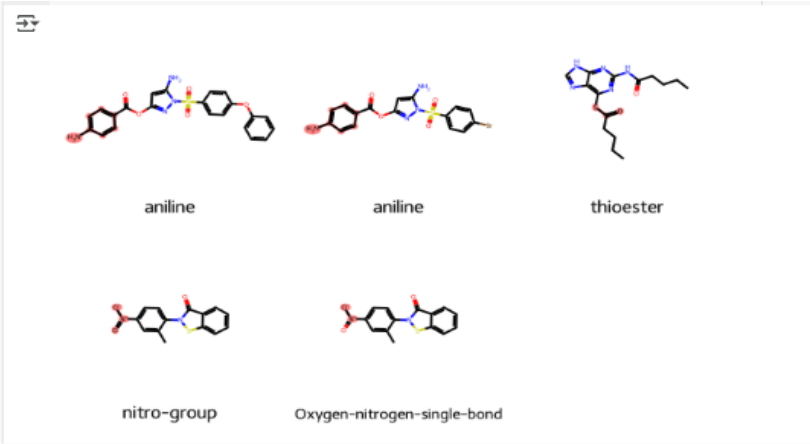


Figure 9: The first 5 identified unwanted substructures

4.4 APPLY AUGMENTATION ON THE DATASETS

The original NS3 target protein bioactivity dataset on DENV-2 collected from CHEMBL database has 236 rows of data. After basic preprocessing it is trimmed down to 192 rows of data. Then after we filtered it for Ro5 properties, the dataset becomes 95 rows of data. Finally We removed unwanted substructures from the dataset and we left with only 37 rows of data. But the size of these datasets is not enough to train our models. So, we need to apply augmentation. And for comparison purpose we augment all of the three datasets after preprocessing the original dataset.

The augmentation technique used employs substructure replacement in SMILES strings using predefined chemical substitutions to generate structurally diverse yet drug-like molecules. It systematically replaces functional groups (e.g., halogens, methoxy, carboxylic acid derivatives, aromatic substituents) in original molecules with bioisosteric or chemically similar alternatives from the SUBSTITUENTS dictionary, by using RDKit's substructure matching and replacement capabilities. Invalid or non-drug-like variants are filtered using chemical validity checks (via Chem.MolFromSmiles) and Lipinski's rule-of-five criteria (molecular weight ≤500, logP ≤5, H-bond donors/acceptors ≤5/10, etc.). Valid augmentations are further diversified by adding Gaussian noise to pIC50 values, ensuring realistic biological activity ranges. Duplicate SMILES are removed, and the final dataset is

subsampled to 5,000 entries, enhancing model training diversity while maintaining pharmacological relevance.

Pseudo Code:

```
# Step 1: Install and import required tools
INSTALL rdkit, pandas, numpy

# Step 2: Load the dataset
LOAD "dataset.csv" INTO dataframe (columns: "smiles", "pIC50")

# Step 3: Define substructure replacement rules
CREATE SUBSTITUENTS dictionary:
   [Cl] → [F], [Br], [I], [C](F)(F)F
   [O][C] → [O][C](F)(F)F, [O][C]C, [C](F)(F)F
   [C](=O)[O] → [C](=O)[N], [C](=O)[S], [C](=O)[O]C
   c1ccccc1 → c1ccc(F)cc1, c1ccc(Cl)cc1, c1ccc(Br)cc1

# Step 4: Augment SMILES strings
FUNCTION augment_smiles(smiles, num_variants=5):
   CONVERT SMILES TO molecule (using RDKit)
   IF invalid molecule: RETURN empty list
   FOR each variant (up to num_variants):
      COPY molecule
      FOR each substructure in SUBSTITUENTS:
         CHECK if molecule contains substructure
         IF YES:
            REPLACE with random choice from SUBSTITUENTS
            CONVERT new molecule TO SMILES
```

ADD to new_smiles list

BREAK to avoid overlapping replacements

IF no replacement found:

GENERATE randomized SMILES (canonical=False, doRandom=True)

RETURN unique SMILES (remove duplicates)

Step 5: Filter drug-like molecules

FUNCTION calculate_lipinski_descriptors(smiles):

CONVERT SMILES TO molecule

IF invalid: RETURN False

CALCULATE molecular properties:

MW (molecular weight), LogP, HBD (H-bond donors), HBA (H-bond acceptors), TPSA, Rotatable Bonds

IF MW ≤ 500 AND LogP ≤ 5 AND HBD ≤ 5 AND HBA ≤ 10 AND TPSA ≤ 140 AND Rotatable Bonds ≤ 10:

RETURN True (drug-like)

ELSE:

RETURN False

Step 6: Generate augmented dataset

INITIALIZE augmented_data = []

FOR each row in df:

GENERATE 5 variants per original SMILES using augment_smiles()

FOR each variant:

IF valid molecule AND passes Lipinski's Rule of Five:

ADD molecule to augmented_data

ADD Gaussian noise (μ=0, σ=0.1) to pIC50 value

CLIP pIC50 to realistic range (3.0–9.0)

Step 7: Post-processing

CONVERT augmented_data TO dataframe

REMOVE duplicate SMILES strings

SUBSAMPLE to 5,000 entries (with replacement)

SAVE dataframe as "dataset_augmented.csv"

4.5 TRAIN THE MODELS

4.5.1 GNN Models Implementation

We trained three types of GNN model with different loss functions and each of them using all the three types of augmented data. So, totally we have nine different GNN models. The main aim of this trained models is to predict PIC50 values for a drug like molecules by using the SMILES notation as an input.

The first GNN model uses MAE as a loss function, the second model uses MSE as a loss function and the last model uses Huber loss as a loss function. Then for each of these models we used all three types of augmented datasets separately for comparison.

Generally, all of the GNN models use a hybrid Graph Neural Network model combining Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) for predicting pIC50 values of a drug candidate for DENV-2 to show how much it is potent against the disease. The model processes graph-structured data from SMILES notation through four layers, where two GCN layers are followed by two GAT layers. The GCN layers aggregate features from local neighborhoods using spectral filtering, while GAT layers introduce multi-head attention mechanisms to weigh the importance of neighboring nodes dynamically. The architecture uses ReLU and ELU activation functions for non-linearity, dropout for regularization, and global mean pooling to condense node embeddings into a graph-level representation. A final linear layer maps the pooled features to a scalar output for the regression target. The model balances

local structural information by GCN and global contextual dependencies by GAT to enhance its ability to capture complex patterns in molecular graphs.

During training, the model minimizes the loss from loss functions employed using the Adam optimizer with a learning rate of 0.000988. The model training process employs early stopping based on validation loss and the R^2 score to prevent overfitting, with a patience threshold of 15 epochs. Evaluation metrics include the loss functions and R^2, providing insights into prediction accuracy and variance explained. The training loop logs losses and R^2 scores, while the best-performing model is saved for final evaluation. Post-training visualization includes plots of training/validation loss and R^2 over epochs, enabling analysis of convergence and model performance. This hybrid approach utilizes the strengths of GCN for structural feature extraction and GAT for attention-based weighting to improve predictive robustness for molecular property prediction (PIC50 value).

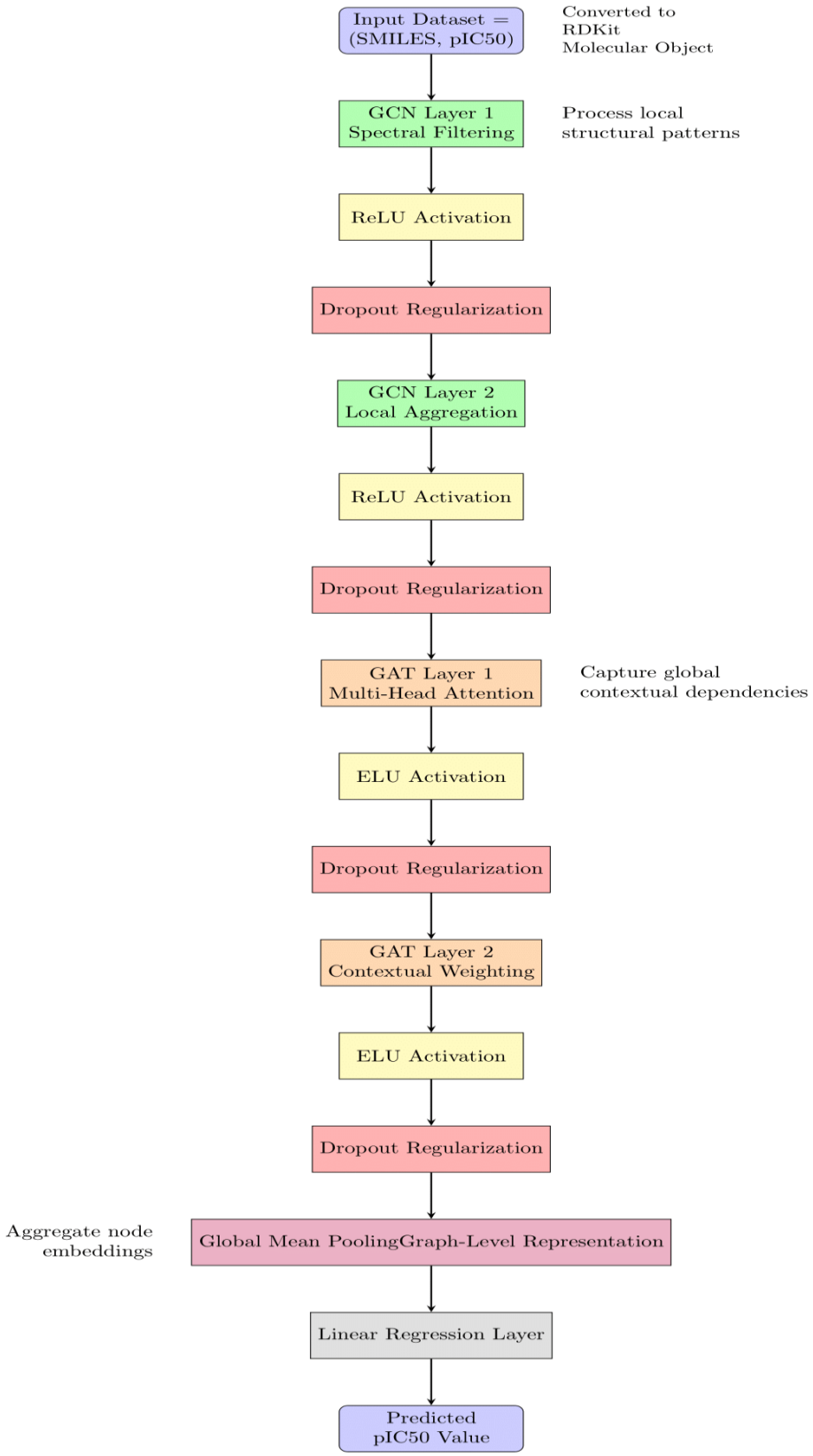


Figure 10: GNN Model Architecture

4.5.2 VGAE Model Implementation

The other type of model used is VGAE model to generate new drug like molecules for drug designing stage of the workflow. And the input dataset used here is the best performing augmented dataset from the three augmented datasets used in the GNN models. The Variational Graph Autoencoder (VGAE) integrates principles from graph neural networks (GNNs) and variational inference to learn probabilistic latent representations of molecular graphs. At its core, the model comprises two components: an encoder that maps molecular graphs into a latent space and a decoder that reconstructs graph structures from these latent embeddings.

The encoder employs four Graph Convolutional Network (GCN) layers to hierarchically extract node-level features and parameterize a probabilistic latent space. The first two stacked GCN layers (conv1, conv2) use ReLU activation and batch normalization to propagate information across the graph's topology, aggregating features from neighboring nodes to build contextualized representations. These layers transform raw node features (e.g., atomic number) through spectral filtering. The "conv1" layer captures local atomic interactions, while the "conv2" layer refines these features by integrating broader neighborhood context (2-hop neighborhoods). The next two GCN layers (conv_mu, conv_logvar) specialize in variational inference. Where "conv_mu" computes the mean (μ) of the latent distribution, and "conv_logvar" computes the log-variance ($\log\sigma^2$) to define uncertainty in the latent space.

During training, latent representations are sampled using the reparameterization trick ($z=\mu+\sigma\odot\epsilon$), enabling gradient flow through stochastic nodes while ensuring the latent space adheres to a prior distribution $N(0,I)$. This probabilistic approach balances reconstruction accuracy and regularization, penalizing deviations from the prior via KL divergence loss.

The decoder reconstructs the graph structure by computing edge probabilities between all pairs of nodes. It uses an inner product operation ($zi \cdot zj$) to measure similarity between latent embeddings zi and zj, followed by a sigmoid activation to output values in [0,1], representing the likelihood of an edge between nodes i and j.

The model trains by learning to compress molecular graphs into compact latent representations and reconstructing them. The process begins with an encoder that maps input graphs (nodes = atoms, edges = bonds) into a probabilistic latent space, where each node is represented by a Gaussian distribution (mean and variance). This probabilistic approach ensures the model captures uncertainty and generalizes well, avoiding overfitting to specific molecular structures. During training, latent embeddings are sampled from these distributions using the reparameterization trick, allowing gradients to flow through the stochastic nodes. The decoder then reconstructs the graph's adjacency matrix by computing similarities between latent node embeddings, predicting edge probabilities via a sigmoid activation. This dual-phase setup ensures the model balances accurate graph reconstruction (via binary cross-entropy loss) and latent space regularization (via KL divergence), which keeps learned distributions close to a standard normal prior.

The model training process uses a combined loss function using the Adam optimizer, with GPU acceleration for efficiency. To prevent overfitting especially on small molecular datasets early stopping halts training if validation performance plateaus, while timestamped checkpoints save the best model. A 5-fold cross-validation strategy ensures robust evaluation by partitioning data into subsets, training on most and validating on the remainder iteratively.

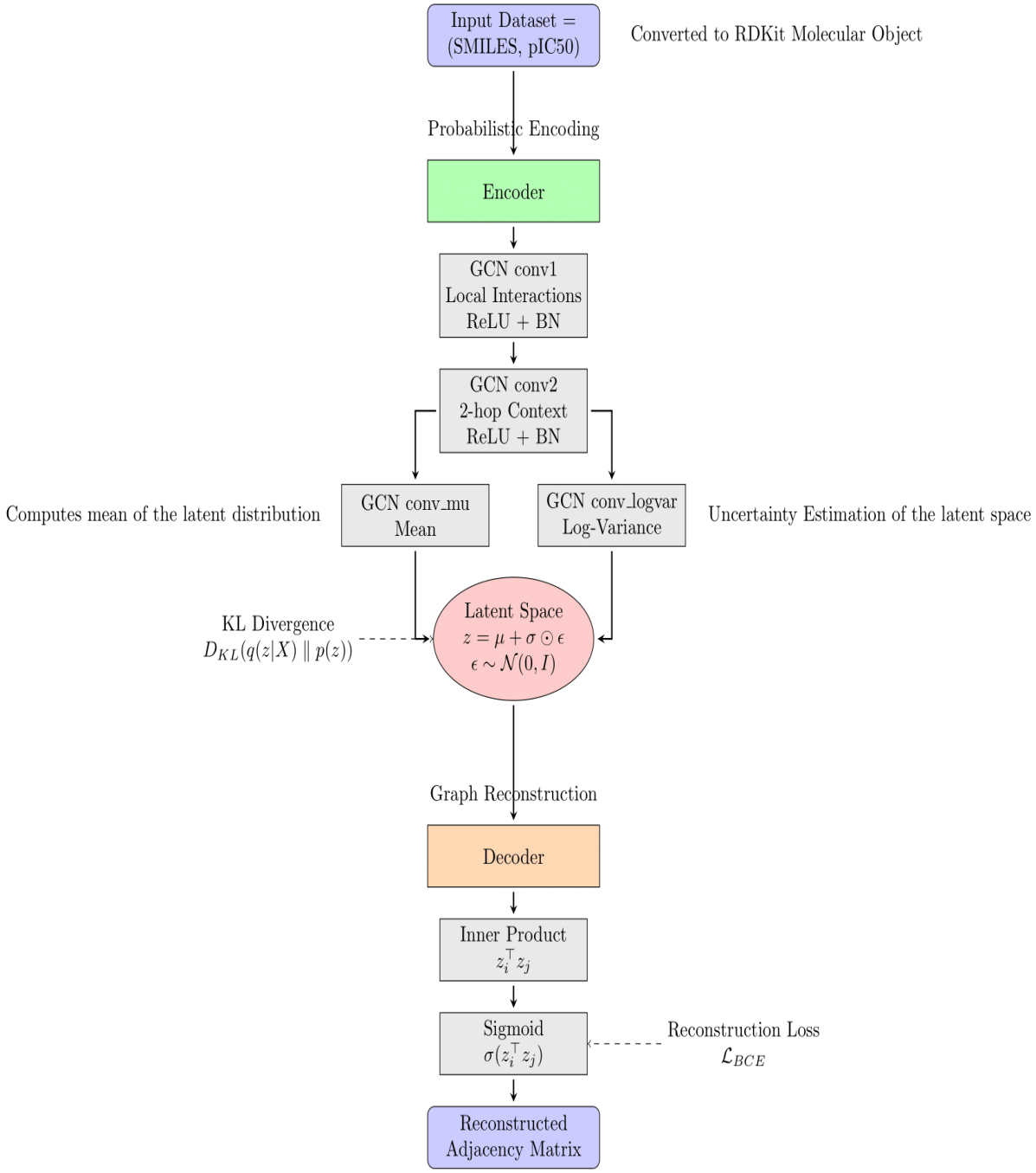


Figure 11: VGAE Model Architecture

4.6 VIRTUAL SCREENING AND DRUG DESIGNING

4.6.1 Virtual Screening

In this stage of the workflow, we use our trained GNN model for drug repurposing using a drug library created from Drug Bank database manually. To create the drug library, we used three filters. The first filter is the drug should be anti-viral. The second filter is the drug should be approved. And the last filter is it should have SMILES notation. With these filters we created a library containing 104 drug candidates. Then we loaded the drug library to our Google-Colab environment. After that we preprocess the drug library by changing SMILES notation into graphical notation. Next, we used our best GNN model to predict PIC50 values for this drug candidates. Finally, we ranked and list out the top 10 by their values.

4.6.2 Drug Designing

The main aim of this stage is to use the VGAE model generate a novel new drug like molecules structure in SMILES notation. First, we loaded our augmented dataset to train the model. Then we use this model to generate latent space. Next, the latent space is used to generate 50 drug-like molecules structures. Finally, we used our GNN model to predict their PIC50 values and rank the top 10 accordingly.

4.7 MOLECULAR DOCKING

Finally, we applied molecular docking on the top 3 drug candidates from both drug repurposing and drug designing stages. In this process the drug candidates are ligands while the target protein is a protein data on of DENV-2 download from Protein Data Bank database identified by "2R69".

The output of molecular docking includes quantitative metrics such as binding energy (kcal/mol) and qualitative visualizations of ligand-protein complexes. Lower binding energy in negative values (<= -5 kcal/mol) indicate stronger interactions or binding affinity. And for this purpose, we used online docking tool called AutoDock vena from SwissDock website.

CHAPTER FIVE: RESULTS AND DISCUSSION

5.1 MODELS PERFORMANCE EVALUATION

5.1.1 GNN Models Performance

From these nine GNN models the best performing model is “Model A.c” or the model that uses MAE as a loss function and the last (clean NS3) augmented dataset for model training with R2 Score of 0.9938 accuracy. So, we used this model to predict PIC50 values of all drug candidates in this research. The model also achieved a test MSE of 0.0057 and MAE of 0.0552, outperforming the baseline model for GNNs, which is DGraphDTA’s. Its performance on the Davis dataset was (MSE = 0.202, MAE ≈ 0.126) and on KIBA dataset (MSE = 0.126, MAE ≈ 0.100).

Model A => “GNN model with MAE loss function”

a. Using the first augmented dataset

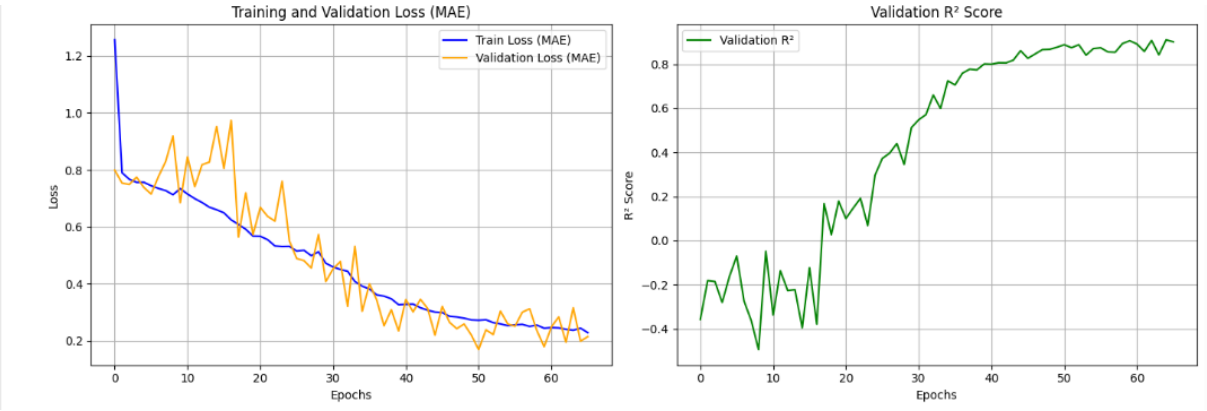


Figure 12: GNN Model A.a performance

b. Using the second augmented dataset

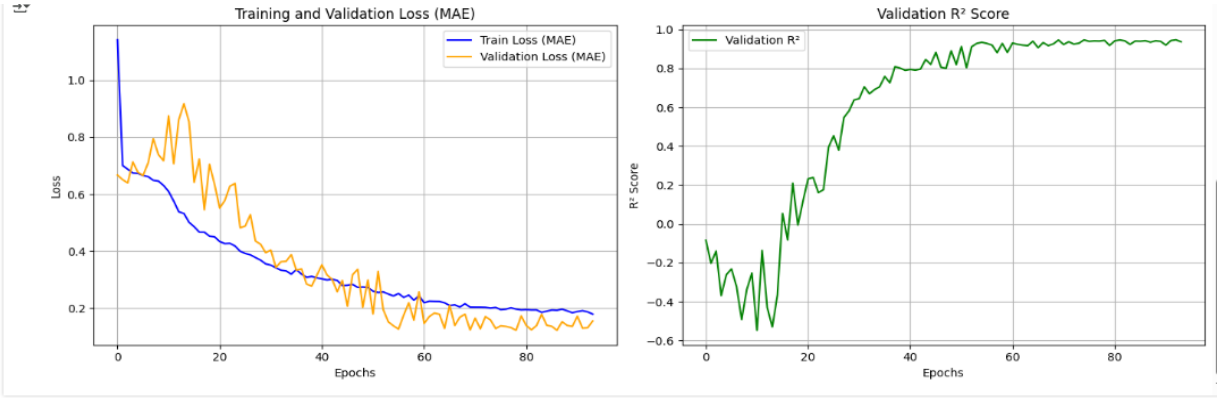


Figure 13: GNN Model A.b performance

c. Using the last augmented dataset

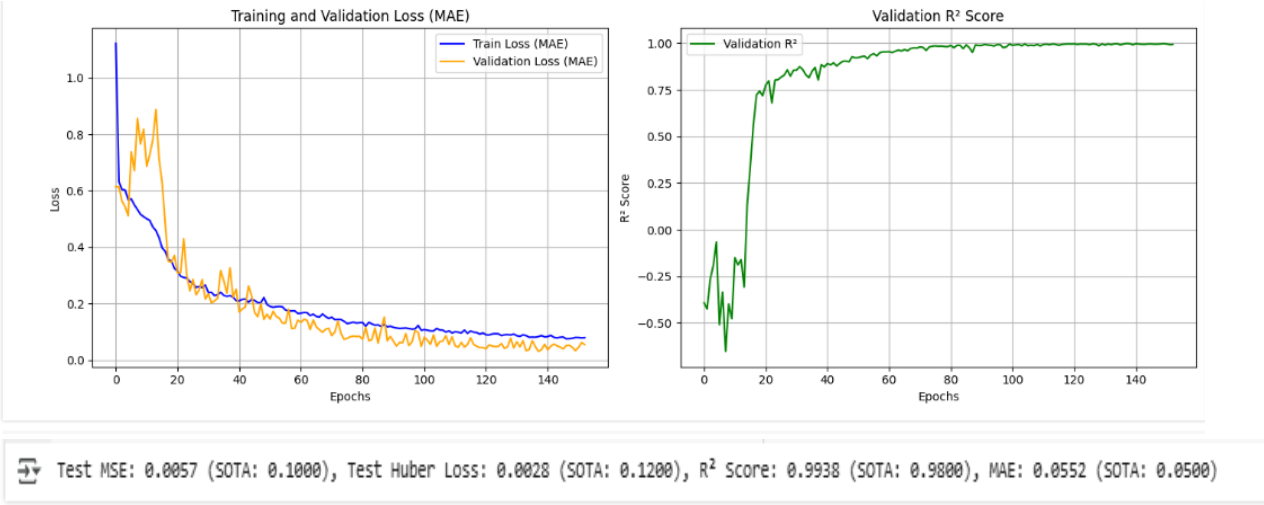


Figure 14: GNN Model A.c performance

Model B => “GNN model with MSE loss function”

a. Using the first augmented dataset

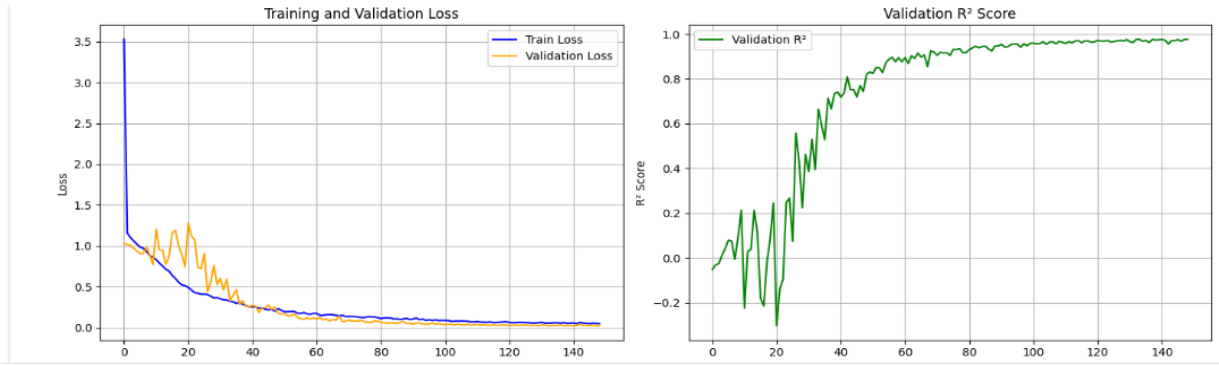


Figure 15: GNN Model B.a performance

b. Using the second augmented dataset

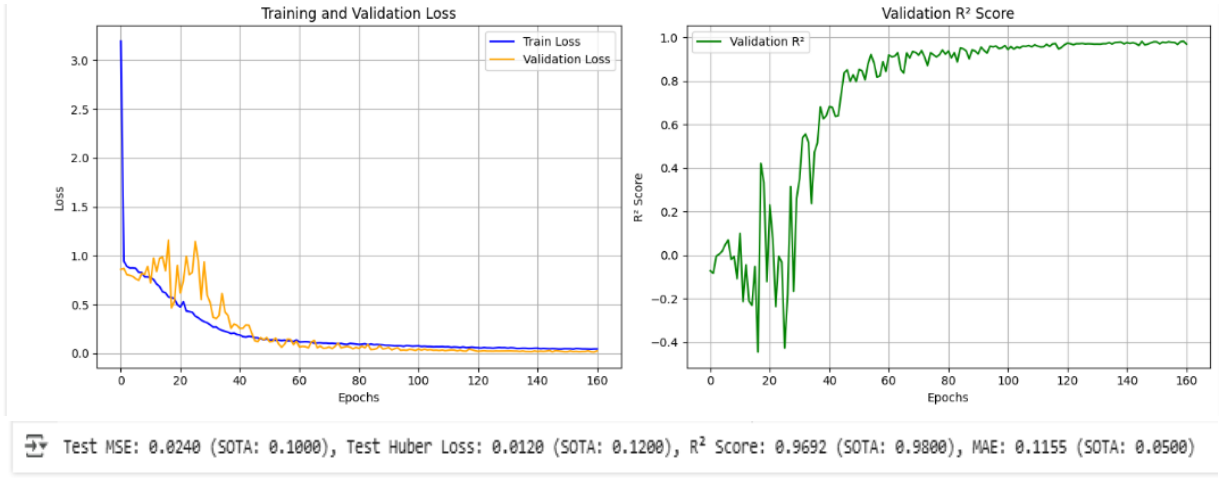
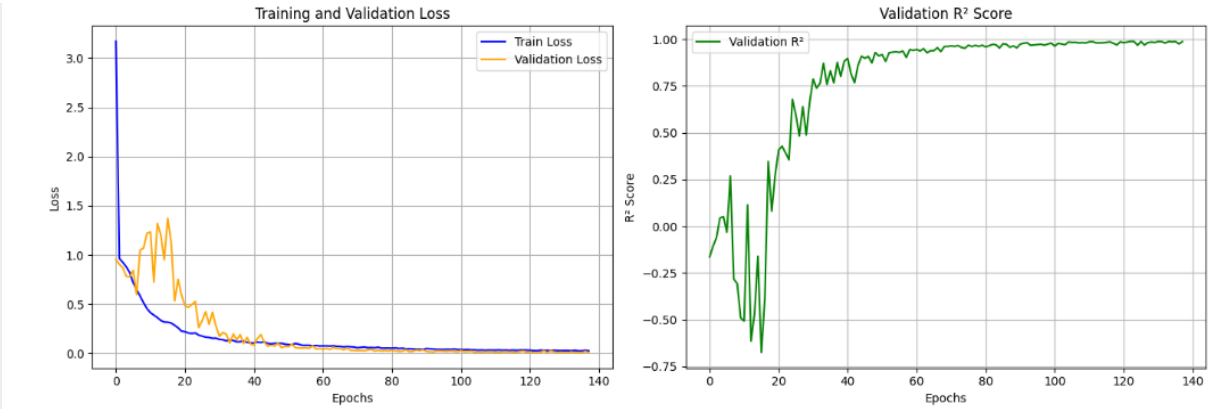


Figure 16: GNN Model B.b performance

c. Using the last augmented dataset

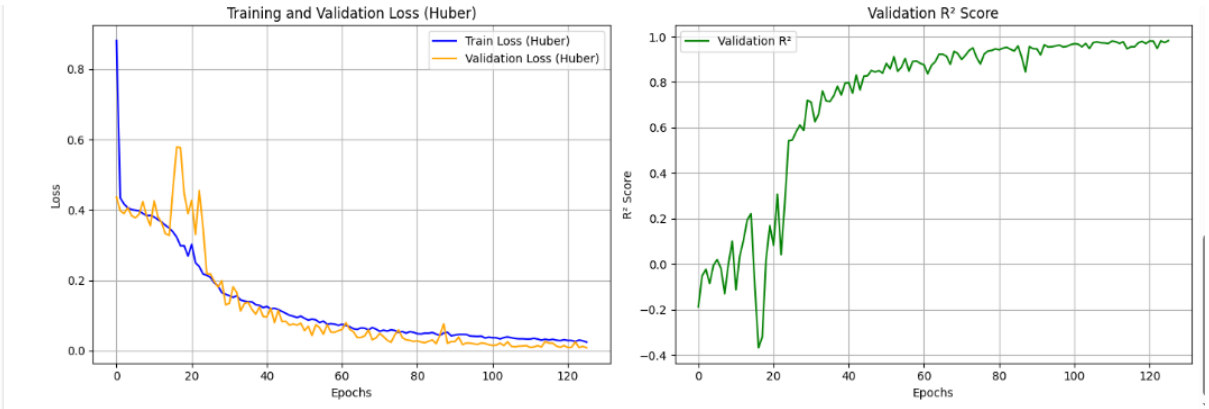


Test MSE: 0.0107 (SOTA: 0.1000), Test Huber Loss: 0.0054 (SOTA: 0.1200), R^2 Score: 0.9882 (SOTA: 0.9800), MAE: 0.0725 (SOTA: 0.0500)

Figure 17: GNN Model B.c performance

Model C => "GNN model with Huber loss function"

a. Using the first augmented dataset



Test MSE: 0.0199 (SOTA: 0.1000), Test Huber Loss: 0.0099 (SOTA: 0.1200), R^2 Score: 0.9803 (SOTA: 0.9800), MAE: 0.0974 (SOTA: 0.0500)

Figure 18: GNN Model C.a performance

b. Using the second augmented dataset

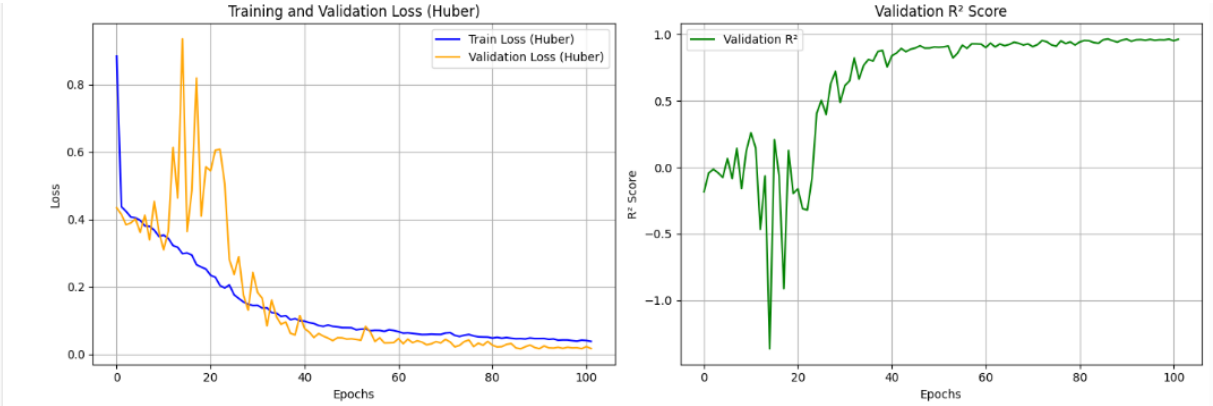


Figure 19: GNN Model C.b performance

c. Using the last augmented dataset

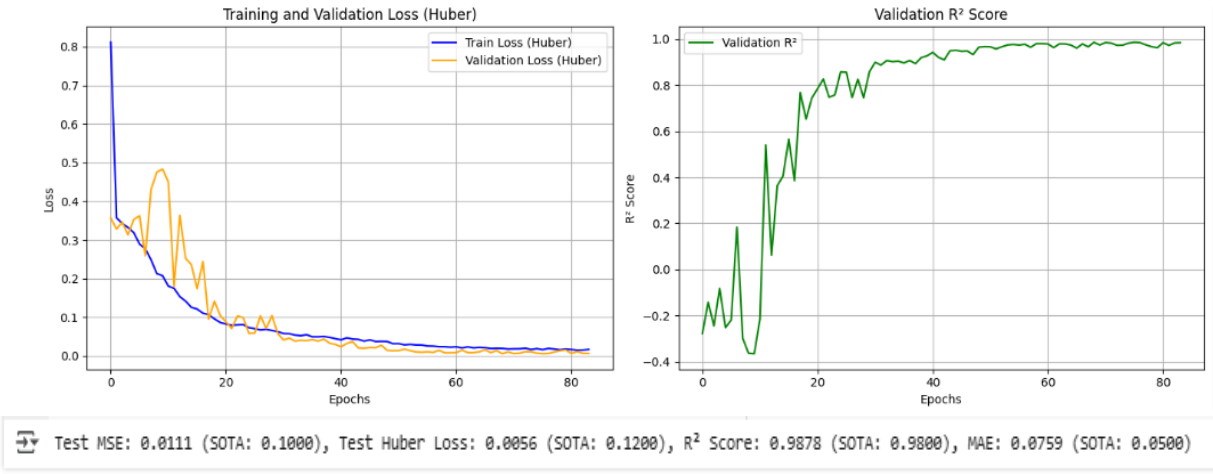


Figure 20: GNN Model C.c performance

5.1.2 VGAE Model Performance

The next figure show performance of the best performing VGAE model after rigorous hyper parameter tuning. This model then used to generate latent space, which would be used for novel drug candidate molecules generation stage.

```
Epoch 83: Train Loss=9.2635, Val Loss=1.7839  EarlyStopping counter: 9/15
Epoch 84: Train Loss=9.2659, Val Loss=1.7811  EarlyStopping counter: 10/15
Epoch 85: Train Loss=9.2620, Val Loss=1.7804  EarlyStopping counter: 11/15
Epoch 86: Train Loss=9.2658, Val Loss=1.7826  EarlyStopping counter: 12/15
Epoch 87: Train Loss=9.2600, Val Loss=1.7815  EarlyStopping counter: 13/15
Epoch 88: Train Loss=9.2631, Val Loss=1.7819  EarlyStopping counter: 14/15
Epoch 89: Train Loss=9.2600, Val Loss=1.7789  EarlyStopping counter: 15/15

Early stopping triggered

Loaded best model: vgae_val-loss_tm1_1.7776_20250518_203356.pt
```

Figure 21: VGAE Model performance

5.2 TOP DRUG CANDIDATES PRECICTED BY THE GNN MODEL

These are the top 10 drugs suggested for repurposing using the GNN model.

A) Top 10 Drug Candidates Predicted from the Drug library

Top 10 Compounds by Predicted pIC50:

	Drug_Name	SMILES	predicted_pIC50
14	Imiquimod	CC(C)CN1C=NC2=C1C1=C(C=CC=C1)N=C2N	5.981026
71	Atazanavir	COC(=O)N[C@H](C(=O)N[C@@H](CC1=CC=CC=C1)[C@@H]...	5.803333
38	Tenofovir alafenamide	CC(C)OC(=O)[C@H](C)N[P@](=O)(CO[C@H](C)CN1C=NC...	5.698824
84	Dequalinium	CC1=CC(N)=C2C=CC=CC2=[N+]1CCCCCCCCCC[N+]1=C(C)...	5.630970
25	Nevirapine	CC1=C2NC(=O)C3=C(N=CC=C3)N(C3CC3)C2=NC=C1	5.625618
97	Indinavir	CC(C)(C)NC(=O)[C@@H]1CN(CC2=CN=CC=C2)CCN1C[C@@...	5.620689
39	Daclatasvir	COC(=O)N[C@@H](C(C)C)C(=O)N1CCC[C@H]1C1=NC=C(N...	5.598548
88	Nitazoxanide	CC(=O)OC1=CC=CC=C1C(=O)NC1=NC=C(S1)[N+]([O-])=O	5.487456
41	Ritonavir	CC(C)[C@H](NC(=O)N(C)CC1=CSC(=N1)C(C)C)C(=O)N[...	5.482825
57	Tipranavir	[H][C@@](CC)(C1=CC(NS(=O)(=O)C2=NC=C(C=C2)C(F)...	5.462296

Figure 22: Top 10 Drug Candidates Predicted from the Drug library

These are top 10 drug-like molecules generated by the VGAE model and the ranked by the GNN model.

B) Top 10 Drug Candidates Predicted from the Generated Molecules

	SMILES	predicted_pIC50
0	C.F.F.FC12OC34N5NC13C524.FON1N(F)N2N(F)N12.ONC...	7.006371
1	C.F.F.FNF.FON1N(F)N2N(F)N12.N1N2C34ON5C13C524....	6.978085
2	C.FC1OC2(F)NNC12.FON1N(F)N2N(F)N12.ONC12CC(F)(...	6.591246
3	C.FC1OC2(F)NNC12.FN(F)N1N(F)N2ON21.ONC12CC(F)(...	6.572318
4	C.FC1OC2(F)C(F)NC12.FN(F)N1N(F)N2ON21.ONC12CC(...	6.569020
5	C.FC1OC2(F)NNC12F.FON1N(F)N2N(F)N12.ONC12CC(F)...	6.484778
6	C.FN1NC2NOC21F.FNF.FON1N(F)N2N(F)N12.NC1NC2(F)...	6.473773
7	C.FC1OC2(F)NNC12F.FN(F)N1N(F)N2ON21.ONC12CC(F)...	6.465849
8	CC1(NO)C2C3(F)NC21N(NF)O3.F.F.FC12OC34N5NC13C5...	5.069014
9	CC1(NO)CC2(F)NC1N(N(F)F)O2.FC1OC2(F)NNC12.FON1...	4.693545

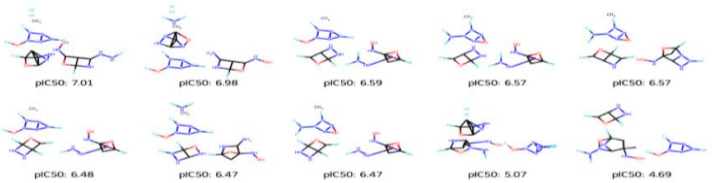


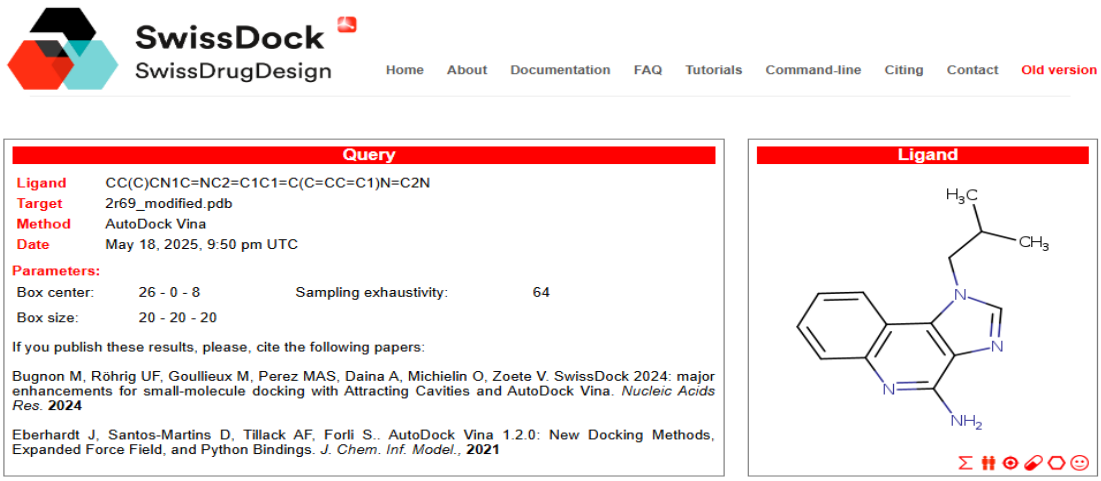
Figure 23: Top 10 Drug Candidates Predicted from the Generated Molecules

5.3 Docking results both types of Drug Candidates

As we can see from the docking results from the top drug candidates their best binding affinities are within the acceptable range i.e. <= -5 Kcal/mol. Therefore, we can say the drug candidates show a promising inhibitory characteristic against DENV-2 disease.

Docking result for the top Drug Candidates Predicted from the Drug library:

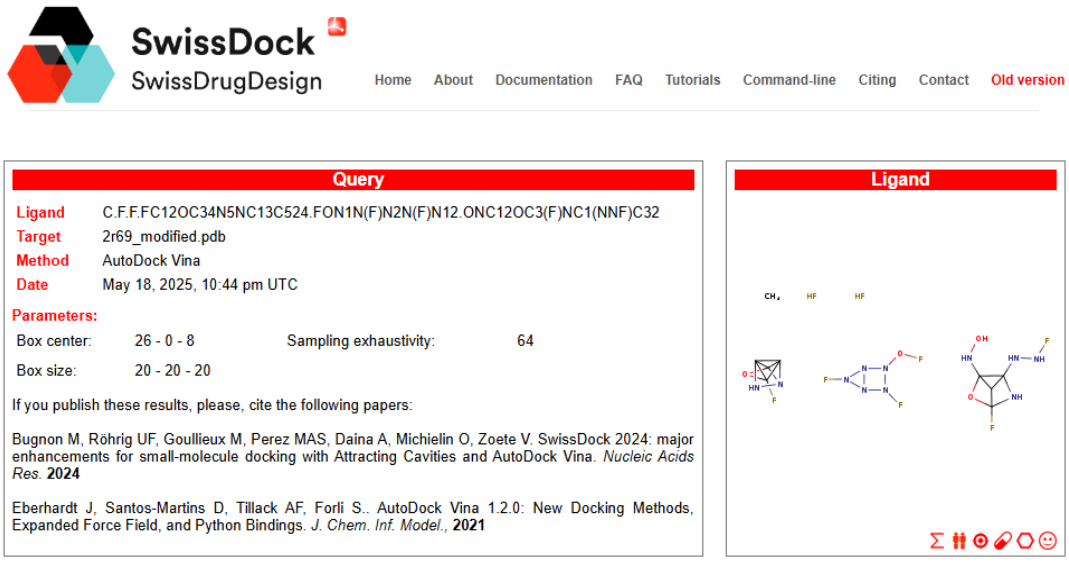
Figure 24: Docking result for the top Drug Candidate Predicted from the Drug library



Model	Calculated affinity (kcal/mol)
1	-6.028
2	-5.652
3	-5.624
4	-5.525
5	-5.490
6	-5.484
7	-5.460
8	-5.459
9	-5.380
10	-5.311
11	-5.288
12	-5.284
13	-5.257
14	-5.251
15	-5.216
16	-5.163
17	-5.153
18	-5.150
19	-5.088
20	-5.065

Docking result for the top Drug Candidates Predicted from the Generated Molecules:

Figure 25: Docking results for the top Drug Candidate Predicted from the Generated Molecules



Model	Calculated affinity (kcal/mol)
1	-5.764
2	-5.726
3	-5.490
4	-5.424
5	-5.413
6	-5.281
7	-5.189
8	-5.147
9	-5.097
10	-5.060
11	-5.049
12	-4.993
13	-4.939
14	-4.921
15	-4.892
16	-4.862
17	-4.832
18	-4.825
19	-4.813
20	-4.794

5.4 RESEARCH QUESTIONS DISCUSSION

1. What are the key molecular targets for DENV-2 that can be used for drug discovery efforts?

The key molecular targets for DENV-2 drug discovery identified in this study are the NS3 target protein from CHEMBL and 2R69 (crystal structure of the Dengue Virus Serotype 2) from Protein Data Bank. These targets were prioritized for their conserved functional domains and validated roles in the viral life cycle, making them ideal for structure-based drug design and inhibition strategies.

2 How can AI-driven computational techniques be employed to identify and design effective antiviral compounds targeting DENV-2?

The techniques were employed through Graph Neural Networks (GNNs) and Variational Graph Autoencoders (VGAEs). Hybrid GNN architectures combining Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) predicted pIC50 values for compounds by analyzing molecular graphs derived from SMILES notation. These models captured local (atomic interactions) and global (molecular topology) features to prioritize compounds with high inhibitory potential. Additionally, VGAEs generated novel drug-like molecules by learning probabilistic latent representations of molecular graphs. And molecular docking is employed to validate binding affinities of top candidates against DENV-2.

3 Which existing bioactive molecules or drugs have the potential to be repurposed for treating infections caused by DENV-2?

Existing bioactive molecules repurposed for DENV-2 were identified through virtual screening of FDA-approved antivirals from DrugBank database by using the best performing GNN model to predict their PIC50 values. A library of 104 drugs was manually created for this purpose. Top-ranked drugs were validated via molecular docking, showing binding affinities $\leq$ -5 kcal/mol against NS3, indicating potential repurposing opportunities for DENV-2.

4 What predictive models and AI algorithms are most effective in identifying drug candidates with high affinity and specificity for DENV-2 targets?

The most effective predictive model from the nine different GNN models has R^2 score of 0.9938 in pIC50 prediction. The model uses a hybrid GCNGAT architecture with MAE loss function by processing graph-based molecular representations in order to capture complex drug-target interactions. For de novo drug design, the best performing VGAE model is used after rigorous hyperparameter tuning.

CHAPTER SIX: CONCLUSION AND FUTURE WORKS

In this concluding chapter, we revisit important aspects of our study, synthesizing the findings and methodologies used to evaluate various deep learning models for predicting potential drug candidates against DENV-2 disease.

6.1 CONCLUSIONS

Our investigation reviewed the effectiveness of AI-driven computational approaches leveraging Graph Neural Networks (GNNs) and Variational Graph Autoencoders (VGAES) for drug discovery by targeting the DENV-2 NS3 protein, which is a critical component of the dengue virus lifecycle. Through a meticulously designed in silico pipeline including dataset curation from ChEMBL, DrugBank, and PDB; molecular augmentation to address data scarcity; and rigorous model training with hybrid GNN architectures (GCN-GAT) optimized via MAE loss, the study demonstrated robust bioactivity prediction (R^2 score of 0.9938) and successful identification of high-affinity drug candidates. The integration of GNN-based virtual screening for repurposing FDA-approved antivirals and VGAE-driven de novo molecule generation enabled the discovery of novel inhibitors, validated by molecular docking with binding energies $\leq$ -5 kcal/mol, underscoring their potential to disrupt NS3 function. These findings affirm the power of graph-based deep learning in capturing complex molecular interactions, while highlighting the necessity of iterative preprocessing (e.g., Lipinski's Rule of Five filtering, PAINS and other unwanted compounds removal) and multimodal validation to ensure pharmacological relevance. By bridging AI-driven predictive modeling with structure-based validation, this work not only advances computational strategies for DENV-2 but also establishes a scalable framework for accelerating antiviral discovery against other challenging diseases.

6.2 FUTURE WORKS

Looking forward, this study identifies several avenues for further research to advance AI-driven antiviral drug discovery for DENV-2 and other pathogens. Addressing data scarcity remains pivotal, as the current work relied on curated, filtered datasets with limited size after Lipinski's

Rule of Five and PAINS filtering. Expanding the bioactivity dataset through integration of additional sources to prioritize high-quality compounds will enhance model robustness and generalizability across dengue serotypes. Additionally, refining the generative model (VGAE) to incorporate pharmacokinetic constraints and synthesizability metrics during molecule generation will ensure novel compounds meet therapeutic benchmarks.

REFERENCES

Naithani, U., & Guleria, V. (2024). Integrative computational approaches for discovery and evaluation of lead compound for drug design. Frontiers in Drug Discovery, 4, Article 1362456. https://doi.org/10.3389/fddsv.2024.1362456

Pinzi, N., & Rastelli, G. (2024). How drug repurposing can advance drug discovery: challenges and opportunities. Frontiers in Drug Discovery, 4, Article 1460100. https://doi.org/10.3389/fddsv.2024.1460100

Gautam, S., Thakur, A., Rajput, A., & Kumar, M. (2024). Anti-Dengue: A machine learning-assisted prediction of small molecule antivirals against dengue virus and implications in drug repurposing. https://doi.org/10.3390/v16010045, Viruses (Published by MDPI)

Halder, S. K., Ahmad, I., Shathi, J. F., et al. (2024). A comprehensive study to unleash the putative inhibitors of serotype 2 of dengue virus: Insights from an in silico structure-based drug discovery. https://doi.org/10.1007/s12033-022-00582-1, Molecular Biotechnology (Published by Springer Nature)

Gallo, F. N., Marquez, A. B., Fidalgo, D., et al. (2024). Antiviral drug discovery: Pyrimidine entry inhibitors for Zika and dengue viruses. https://doi.org/10.2139/ssrn.4778126, SSRN (Social Science Research Network, now owned by Elsevier)

Huang, D., Yang, M., Wen, X., Xia, S., & Yuan, B. (2024). AI-driven drug discovery: Accelerating the development of novel therapeutics in biopharmaceuticals. https://doi.org/10.60087/jklst.vol3.n3.p.206-224, Journal of Knowledge, Logic, and Scientific Thought (EuroPub)

Duan, F.-L., Duan, C.-B., Xu, H.-L., Zhao, X.-Y., Sukhbaatar, O., Gao, J., Zhang, M.-Z., Zhang, W.-H., & Gu, Y.-C. (2024). AI-driven drug discovery from natural products. https://doi.org/10.1016/j.aac.2024.06.003, Antimicrobial Agents and Chemotherapy (Published by Elsevier)

Jayatunga, M. K. P., Ayers, M., Bruens, L., Jayanth, D., & Meier, C. (2024). How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. https://doi.org/10.1016/j.drudis.2024.104009, Drug Discovery Today (Published by Elsevier)

Roney, M. (2024). Artificial intelligence in anti-dengue drug development. https://doi.org/10.1016/j.ipha.2024.01.006, International Journal of Pharmaceutical Analysis (Published by Elsevier)

Lee, D. (2023). AI-driven drug discovery and development for precision medicine. https://africansciencegroup.com/index.php/AJAISD, African Journal of AI and Sustainable Development (AJAISD)

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Rupp, M., & Friederich, P. (2022). Graph neural networks for materials science and chemistry. Communications Materials, 3(1), 93. https://doi.org/10.1038/s43246-022-00315-6

Bender, A., Cortes-Ciriano, I., & Nicolaou, C. A. (2021). Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discovery Today, 26(2), 511–524. https://doi.org/10.1016/j.drudis.2020.12.009

Gentile, F., Agrawal, V., Hsing, M., Ton, A. T., Ban, F., Norinder, U., … Cherkasov, A. (2020). Deep Docking: A deep learning platform for augmentation of structure based drug discovery. Journal of Chemical Information and Modeling, 60(9), 4149–4160. https://doi.org/10.1021/acs.jcim.0c00484

Zhang, Q., Lee, J., Bai, Y., Tsujimoto, N., Yang, W., Huh, S., … Zhavoronkov, A. (2022). AlphaFold accelerated AI drug discovery platform identifies novel hit compounds for novel targets. ChemRxiv. https://doi.org/10.26434/chemrxiv-2022-fp8zv

Sarkar, C., Das, B., Rawat, V. S., Wahlang, J. B., Nongpiur, A., Tiewsoh, I., Lyngdoh, N. M., Das, D., Bidarolli, M., & Sony, H. T. (2023). Artificial intelligence and machine learning technology-driven modern drug discovery and development. https://doi.org/10.3390/ijms24032026, International Journal of Molecular Sciences (Published by MDPI)

Acharjee, P. B., Ghai, B., Elangovan, M., Bhuvaneshwari, S., Rastogi, R., & Rajkumar, P. (2023). Exploring AI-driven approaches to drug discovery and development. https://doi.org/10.58414/SCIENTIFICTEMPER.2023.14.4.48, The Scientific Temper (Published by Connect Journals)

Costa, R. A., Rocha, J. A. P., Pinheiro, A. S., et al. (2022). A computational approach applied to the study of potential allosteric inhibitors protease NS2B/NS3 from Dengue Virus. https://doi.org/10.3390/molecules27134118, Molecules (Published by MDPI)

Nascimento, I. J. dos S., Rodrigues, É. E. S., Silva, M. F., & Araújo-Júnior, J. X. (2022). Advances in computational methods to discover new NS2B-NS3 inhibitors useful against dengue and Zika viruses. https://doi.org/10.2174/1568026623666221122121330, Current Topics in Medicinal Chemistry (Published by Bentham Science Publishers)

Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. H. M., & Ahsan, M. J. (2022). Machine learning in drug discovery: A review. https://doi.org/10.1007/s10462-021-10058-4, Artificial Intelligence Review (Published by Springer Nature)

Kaushik, A. C., & Raj, U. (2020). AI-driven drug discovery: A boon against COVID-19? https://doi.org/10.1016/j.aiopen.2020.07.001, AI Open (Published by Elsevier)

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., & Wei, Z. (2020). Drug–target affinity prediction using graph neural network and contact maps. RSC Advances, 10(20701-20712). https://doi.org/10.1039/d0ra02297g, RSC Advances (Published by the Royal Society of Chemistry)

Kazakova, R. R., & Masson, P. (2022). Quantitative measurements of pharmacological and toxicological activity of molecules. *Chemistry, 4*(4), 1466–1474. https://doi.org/10.3390/chemistry4040097

Pérez, J., Díaz, C., Salado, I. G., Pérez, D. I., Peláez, F., Genilloud, O., & Vicente, F. (2013). Evaluation of the effect of compound aqueous solubility in cytochrome P450 inhibition assays. *Advances in Bioscience and Biotechnology, 4*, 628–639. https://doi.org/10.4236/abb.2013.45083

Lin, J., Sahakian, D. C., de Morais, S. M. F., Xu, J. J., Polzer, R. J., & Winter, S. M. (2003). *The role of absorption, distribution, metabolism, excretion and toxicity in drug discovery*. *Current Topics in Medicinal Chemistry*, 3(10), 1125–1154. doi:10.2174/1568026033452096

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2024). The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research, 52 (D1), D1180–D1192. Advance online publication. https://doi.org/10.1093/nar/gkad1004

The UniProt Consortium. (2025). UniProt: The Universal Protein Knowledgebase. Nucleic Acids Research, 52 (D1), D1072–D1081. https://doi.org/10.1093/nar/gkad1004

Wishart, D. S., Feunang, Y. D., Anjum, A., Djoumbou, Y., Lo, K., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeed, S., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, M., Arndt, D., Zhou, Y., Liu, N., Paramandhak, K., & Greiner, R. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research, 46 (D1), D1072–D1081. https://doi.org/10.1093/nar/gkx1037

Lipinski, C. A. (2004). Lead- and drug-like compounds: The rule-of-five revolution. Drug Discovery Today: Technologies, 1 (4), 337–341. https://doi.org/10.1016/j.ddtec.2004.11.007

Brenk, R., Schipani, A., James, D., Krasowski, A., Gilbert, I. H., Frearson, J., & Wyatt, P. G. (2008). Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem, 3*(3), 435–444. https://doi.org/10.1002/cmdc.200700139

Baell, J. B., & Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry, 53*(7), 2719–2740. https://doi.org/10.1021/jm901137j

Lok, S. M., Kostyuchenko, V. K., Nybakken, G. E., Holdaway, H. A., Battisti, A. J., Sukupolvi-Petty, S., Sedlak, D., Fremont, D. H., Chipman, P. R., Roehrig, J. T., Diamond, M. S., Kuhn, R. J., & Rossmann, M. G. (2007). Crystal structure of Fab 1A1D-2 complexed with E-DIII of Dengue virus at 3.8 angstrom resolution (PDB ID: 2R69). RCSB Protein Data Bank. https://doi.org/10.2210/pdb2R69/pdb

Bugnon, M., Röhrig, U. F., Goullieux, M., Perez, M. A. S., Daina, A., Michielin, O., & Zoete, V. (2024). SwissDock 2024: Major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. Nucleic Acids Research, 52(W1), W324–W332. https://doi.org/10.1093/nar/gkae300

Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. Journal of Chemical Information and Modeling, 61(8), 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

APPENDIXES

APPENDIX A: SAMPLE CODE FOR THE GNN MODEL

```
# Define the GNN model

import torch
from torch.nn import Linear
import torch.nn.functional as F
from torch_geometric.nn import GCNConv, GATConv, global_mean_pool

class HybridGCNGAT(torch.nn.Module):
    def __init__(self, hidden_dim=128, gat_heads=4):
        super().__init__()

        # First two layers: GCNConv
        self.conv1 = GCNConv(graphs[0].num_node_features, hidden_dim)
        self.conv2 = GCNConv(hidden_dim, hidden_dim)

        # Next two layers: GATConv
        self.conv3 = GATConv(hidden_dim, hidden_dim, heads=gat_heads)
        self.conv4 = GATConv(hidden_dim * gat_heads, hidden_dim,
heads=1)

        # Final linear layer for regression
        self.lin = Linear(hidden_dim, 1)

    def forward(self, data):
        x, edge_index, batch = data.x, data.edge_index, data.batch

        # First GCN layer with ReLU activation
        x = F.relu(self.conv1(x, edge_index))
        x = F.dropout(x, p=0.5, training=self.training)  # Optional
dropout

        # Second GCN layer with ReLU activation
        x = F.relu(self.conv2(x, edge_index))
        x = F.dropout(x, p=0.5, training=self.training)  # Optional
dropout

        # First GAT layer with ELU activation
        x = F.elu(self.conv3(x, edge_index))
        x = F.dropout(x, p=0.5, training=self.training)  # Optional
dropout
```

```
        # Second GAT layer with ELU activation
        x = F.elu(self.conv4(x, edge_index))

        # Global pooling (mean over all nodes in each graph)
        x = global_mean_pool(x, batch)

        # Final linear layer for regression output
        return self.lin(x).squeeze()
```

APPENDIX B: SAMPLE CODE FOR THE VGAE MODEL

```
#Define the Encoder and Decoder for the VGAE model

class VariationalGCNEncoder(torch.nn.Module):
    def __init__(self, in_channels, hidden_channels, out_channels):
        super().__init__()
        self.conv1 = GCNConv(in_channels, hidden_channels)
        self.conv_mu = GCNConv(hidden_channels, out_channels)
        self.conv_logstd = GCNConv(hidden_channels, out_channels)

    def forward(self, x, edge_index):
        x = F.relu(self.conv1(x, edge_index))
        return self.conv_mu(x, edge_index), self.conv_logstd(x,
edge_index)


class InnerProductDecoder(torch.nn.Module):
    def forward(self, z, edge_index):
        return (z[edge_index[0]] * z[edge_index[1]]).sum(dim=1)
```

```
# Step 3: Load the dataset

from rdkit import Chem
from torch_geometric.data import Data

def smiles_to_graph(smiles, target):
    mol = Chem.MolFromSmiles(smiles)
    if mol is None:
        return None

    # Node features (atom-level)
    node_features = []
    for atom in mol.GetAtoms():
        features = [
            atom.GetAtomicNum(),       # Atomic number
            atom.GetDegree(),          # Degree of the atom
            atom.GetHybridization().real,  # Hybridization type
            atom.GetIsAromatic()       # Is the atom aromatic?
        ]
        node_features.append(features)

    # Edge indices (bond-level)
    edge_indices = []
```

```
    for bond in mol.GetBonds():
        i = bond.GetBeginAtomIdx()
        j = bond.GetEndAtomIdx()
        edge_indices.append([i, j])
        edge_indices.append([j, i])  # Add reverse edges

    edge_index = torch.tensor(edge_indices,
dtype=torch.long).t().contiguous()
    x = torch.tensor(node_features, dtype=torch.float)
    y = torch.tensor([target], dtype=torch.float)

    return Data(x=x, edge_index=edge_index, y=y)
    import pandas as pd

# Load dataset
df = pd.read_csv("/content/4-Clean_NS3_data (1)_augmented.csv")

# Keep only SMILES and pIC50 columns
df = df[["smiles", "pIC50"]].dropna()

# Convert SMILES to graphs
graphs = []
for _, row in df.iterrows():
    data = smiles_to_graph(row["smiles"], row["pIC50"])
    if data is not None:
        graphs.append(data)

print(f"Number of valid graphs: {len(graphs)}")
```