Journal of Computer Science

Research Article

An Integrated Framework to Predict the Strength of New SMILE Using Graph Attention Network

Sandhi Kranthi Reddy and S. V. G. Reddy

Department of CSE, GST, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India

Article history
Received: 24-04-2025
Revised: 11-06-2025
Accepted: 24-06-2025

Corresponding Author:
Sandhi Kranthi Reddy
Department of CSE, GST, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India
Email: ksandhi@gitam.in

Abstract: Machine Learning (ML) and Deep Learning (DL) have significantly advanced various fields, including healthcare, finance, autonomous systems, and scientific research. In healthcare, these technologies have been widely applied to disease prediction, such as cancer and diabetes. However, the challenge of drug resistance persists, creating a need for more effective drugs. Developing new drugs is a complex, expensive, and time-intensive process, requiring innovative approaches to enhance efficiency. To address this, GAT-PDE (Graph Attention Network-Based Framework for Predicting Drug Efficacy) is proposed to predict the efficacy of the new drugs/SMILES for specific diseases. The framework incorporates Pharmacophore fingerprints, Jaccard coefficient, quartile analysis, and Graph Attention Networks (GATs) to improve drug efficacy predictions. The Jaccard coefficient assesses molecular similarity between a reference drug and a database of one million compounds using pharmacophore fingerprints. Avapritinib, a proven drug for gastrointestinal stromal tumours (GIST), serves as the reference compound. Quartile analysis categorizes molecules based on Jaccard coefficient, generating labelled data. A GAT model is trained on this data, achieving 88% accuracy in predicting drug efficacy, demonstrating its potential for predicting efficacy of a new drug.

Keywords: Machine Learning, Deep Learning, Graph Attention Network, Pharmacophore Fingerprints, Jaccard Coefficient, Quartile Analysis, Avapritinib, Gastrointestinal Stromal Tumours

Introduction

Artificial Intelligence (AI), particularly Machine Learning (ML) and Deep Learning (DL) are transforming various domains, including healthcare, finance, autonomous systems, and scientific advancements. These AI-driven technologies enable computers to learn from vast amounts of data, recognize patterns, and make intelligent decisions with minimal human intervention (Mian *et al.*, 2024). ML works well on Euclidean data like Tabular format, Image data etc., while DL, a subset of ML, works on both Euclidean data and non-Euclidean data like Molecular Graphs, Networks etc. Their applications encompass customized recommendations, fraud detection, medical diagnostics, and self-driving cars (Sarker 2021a-b). AI is continuously evolving, reshaping problem-solving strategies, boosting efficiency, and fostering innovation across diverse domains.

ML and DL have revolutionized healthcare by enhancing disease prediction, early diagnosis, and treatment planning, ultimately improving patient outcomes (Gandhi & Gandhi, 2022; Rahman *et al.*, 2024). Despite these advancements, developing effective new drugs remains a significant challenge (Biala *et al.*, 2023). Many chronic and life-threatening diseases, including various cancers, continue to demand novel therapeutic solutions as resistance to existing treatments escalates, diminishing their efficacy (Saeed *et al.*, 2023; Kamrani *et al.*, 2023). This increasing resistance highlights the critical need for innovative drug discovery approaches to identify and develop more effective and sustainable drug candidates (Garg *et al.*, 2024).

Developing new drugs is an extremely intricate, expensive, and time-intensive process that requires multiple critical stages to identify effective treatments for specific diseases (Hughes *et al.*, 2011). Traditional methods for identifying potential drug candidates from vast chemical libraries are not only resource-intensive but also highly inefficient, often leading to high attrition rates (Parasrampuria *et al.*, 2018). Consequently, innovative computational strategies are essential to accelerate and enhance the drug discovery process.

Advancements in ML and DL are transforming drug discovery, particularly in the early stages of identifying

promising drug candidates. These technologies enable the rapid analysis of massive biological datasets, improve drug efficacy predictions, and optimize lead compounds for further development (Kumar & Roy, 2025). By optimizing these processes, ML and DL significantly expedite drug discovery, reduce costs, and enhance the success rate of developing effective treatments (Lv *et al.*, 2023; Visan *et al.*, 2024).

A crucial aspect of modern drug discovery is leveraging reference drugs with known efficacy to identify new candidates (Boniolo *et al.*, 2021). However, traditional computational methods often struggle to capture the intricate structural and functional relationships between molecules, limiting their predictive accuracy (Krzywanski *et al.*, 2024). To overcome this challenge, Graph Neural Networks (GNNs) have emerged as a powerful solution by utilizing graph-based representations, where atoms serve as nodes and chemical bonds as edges, to uncover deep structural and functional patterns essential for accurate drug/SMILE efficacy prediction (Khemani *et al.*, 2024). By leveraging reference drugs as a foundation, GNN-based models provide a scalable, data-driven strategy for evaluating new drug candidates. Their ability to enhance predictive accuracy, reduce development costs, and increase the success rate of novel therapeutics makes them an essential tool in addressing drug resistance and tackling complex diseases.

GNNs can be used for three main types of prediction tasks (Zhang *et al.*, 2021): graph-level, node-level, and edge-level.

- A node-level task is used to predict the property of node or classify the node.
- An edge-level task is used to predict the existence of an edge between two nodes.
- A graph-level task, is used to predict the property of whole graph or classify the whole graph.

To fully capture molecular functionalities, a graph-level task is the most suitable choice and among various GNN architectures, Graph Attention Networks (GATs) have gained significant attention for graph-level tasks (Vrahatis *et al.*, 2024). They enhance traditional GNN models by incorporating an attention mechanism, which allows the network to assign different importance weights to neighbouring nodes. This feature is particularly beneficial in drug discovery, where the complete molecular graph is utilized to capture structural and functional relationships, leading to more accurate predictions of drug efficacy.

In this research paper, we propose and explore the application of Graph Attention Network-Based Framework for Predicting Drug Efficacy (GAT-FDE) to assess the efficacy of the new drug for targeting PDGFR using Avapritinib as a reference compound which has demonstrated significant efficacy in patients with PDGFRA mutant gastrointestinal stromal tumors (GISTs). This approach corresponds to ligand-based virtual screening (LBVS), a widely used technique in the lead identification phase of drug discovery, where new compounds are selected based on their similarity to known active ligands.

Literature Review

Chang *et al.* (2018) developed CDRscan, a convolutional neural network (CNN)-based deep learning model to predict anticancer drug responses using genomic profiles of 787 cancer cell lines and structural data from 244 drugs. The model achieved high prediction accuracy and identified novel drug repurposing opportunities.

Zhu *et al.* (2021) developed DLEPS, a deep learning system based on deep neural networks that predicts drug effectiveness by using changes in gene expression from diseased tissues. After training on over one million gene profiles linked to thousands of molecules, the system predicted gene changes in new, unseen data, with a strong correlation score of 0.74.

Gaudelet *et al.* (2021) discuss how graph-based machine learning is being used in drug discovery to better understand biological molecules and integrate various types of biological data. They highlight its usefulness in identifying drug targets, designing new drugs, and finding new uses for existing medicines.

Li *et al.* (2022) developed a new graph neural network that analyzes how atoms interact in molecules at multiple levels. This approach simplifies the process and performs well in predicting how well drugs bind to proteins for COVID-19 drug design, highlighting the role of GNNs in protein-ligand interaction prediction.

Xie *et al.* (2022) developed EPL-GNN, a deep learning model based on graph neural networks to predict lung cancer patients' response to immune checkpoint inhibitors using H&E biopsy images. Tested on 583 patients, the model outperformed existing biomarkers, highlighting the role of GNNs in predicting lung cancer patient responses to these inhibitors and guiding treatment options.

Budak *et al.* (2023) used a graph neural network to identify FDA-approved drugs that could be repurposed for COVID-19 treatment. By analyzing drug structures and binding strengths, they found several kinase inhibitors and antiviral drugs, originally used for lung cancer and other diseases, as promising candidates for faster COVID-19 therapies, highlighting the role of GNNs in predicting drug repurposing.

Saihood *et al.* (2024) developed MS-GNN-ALCFF, a graph neural network model using a multi-side graph construction layer and attention-based fusion to classify lung nodules using 3D CT images. Tested on large datasets, it showed strong and consistent accuracy, highlighting the role of GNNs in lung cancer detection.

Vaida *et al.* (2025) developed M-GNN, a graph neural network model by integrating metabolomics and demographic data to improve early detection of lung cancer. Using 800 samples, their model achieved high accuracy, highlighting the role of GNNs in lung cancer prediction.

Ali *et al.* (2025) evaluated six deep learning models, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), 3-stacked GRU, 3-stacked GRU with Attention (GRU-ATTN), Transformer networks, and Autoencoder architectures to predict how drugs change gene expression in cells. Among which, the 3-stacked GRU with Attention has outperformed other models with 79% accuracy. This highlights the effectiveness of attention mechanisms in capturing critical patterns.

The reviewed literature highlights the increasing importance of machine learning, particularly graph neural networks (GNNs), in advancing drug discovery and improving disease prediction by effectively capturing complex molecular interactions and biological data. This has led to improvements in drug efficacy prediction using gene expression profiles, drug target identification, protein-ligand interaction prediction, drug repurposing, and patient-specific treatment response. These findings indicate that GNN-based models, including our proposed model (GAT-PDE), hold a strong potential for accurately predicting the efficacy of new drugs, thereby supporting faster and more cost-effective drug development with promising clinical applications.

Methods

Molecular Fingerprints

Molecular fingerprints are essential computational representations of chemical structures that are widely used in cheminformatics for similarity searching, clustering, and predictive modelling in drug discovery. They encode the chemical properties of molecules as bit vectors or numerical arrays, where each bit represents the presence ("1") or absence ("0") of a particular structural element. By generating and utilizing these numerical representations, we can perform efficient computational and analytical modeling, enabling rapid and accurate comparisons between compounds.

Many types of fingerprints exist to digitally represent chemical structures for diverse cheminformatics applications. For example, MACCS fingerprints efficiently encode specific chemical properties, making them highly useful in virtual screening and QSAR studies to quickly identify and optimize potential drug leads. Topological fingerprints offer a detailed view of molecular structure by mapping the interrelationships among atoms, while Morgan fingerprints (or circular fingerprints) capture the local chemical environment around each atom by considering neighboring atoms within a set radius. Avalon fingerprints generate unique signatures by recording the presence or absence of particular substructural elements, and atom-pair fingerprints document pairwise atomic interactions, facilitating efficient similarity searches based on spatial configurations. Additionally, path-based fingerprints (like those produced by RDKit and Daylight) and torsion-based methods provide valuable insights into sequential and conformational features. Pharmacophore fingerprints encode key functional features that directly relate a compound's structure to its biological activity (Muegge & Mukherjee, 2016). Among all these techniques, pharmacophore fingerprints have been used in our work.

Pharmacophore Fingerprints represent molecule with high bit length of 39,972 focusing on key features that directly influence a compound's biological activity (Yang *et al.*, 2022). By emphasizing these critical functional attributes, pharmacophore fingerprints enable more effective virtual screening and lead optimization, significantly enhancing the ability to identify bioactive molecules with diverse scaffolds for drug discovery.

Similarity Metrics

In drug discovery, similarity metrics are essential to evaluate how similar different compounds are, which helps in identifying promising leads.

Common similarity metrics widely used in drug discovery are Cosine Similarity, Dice Coefficient and Jaccard/Tanimoto Similarity. Cosine similarity measures the cosine of the angle between two molecular fingerprint vectors, emphasizing the pattern of features rather than their magnitude. Dice coefficient evaluates similarity by calculating twice the number of shared features divided by the sum of features in both fingerprints, highlighting the degree of overlap. Jaccard similarity determines similarity by dividing the number of shared features by the total number of unique features across both fingerprints, providing a ratio of intersection over union (Willett, 2009). Among all these metrics, Jaccard similarity is used in our work.

Jaccard Similarity measures similarity by dividing the number of shared features by the total number of unique features (i.e., the union) present in the compared fingerprints, providing a clear and interpretable ratio of structural overlap. Its straightforward calculation and sensitivity to key functional characteristics make it particularly effective for our virtual screening and lead identification efforts (Bajusz *et al.*, 2015).

Formula to find Jaccard similarity between binary molecular fingerprints is

$$T(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Where:

A and B are two molecular fingerprints

Quartile Analysis

Quartile analysis is a statistical technique that divides a dataset into four equal segments, providing insights

into how key molecular properties are distributed (Goswami & Chakrabarti, 2012; Akshara & Jain, 2024). In our work, it is used to transform continuous scores into discrete labels: 0 for the first quartile, 1 for the second, 2 for the third, and 3 for the fourth, assigning each compound to its corresponding quartile. This process is essential for supervised learning in drug discovery, as it allows Graph Neural Networks to differentiate between varying similarity levels and generate more precise predictions.

Graph Attention Networks

Graph Attention Networks (GATs) enhance traditional GNNs which operate based on message passing and aggregation. by introducing an attention mechanism, which assigns varying importance (weights) to neighboring nodes when updating a node's representation. This allows the model to focus more on critical molecular interactions, improving prediction accuracy (Lavecchia, 2024).

At each hidden layer of GAT, the following steps are performed.

Linear Transformation

Each node's feature vector undergoes a learnable linear transformation using a weight matrix W. This transformation allows the model to project the input features into a new feature space, making them more suitable for learning complex relationships.

For a given node i with an initial feature vector h_i, the transformation is defined as:

$$h_i^{'} = wh_i \tag{2}$$

Where:

h_i represents the input feature vector of node i.

w is the learnable weight matrix applied to all nodes in the graph.

$h_i^{'}$ is the transformed feature vector of node i.

Compute Attention Scores

In GATs, the attention mechanism determines the importance of neighboring nodes when updating a node's representation. This is done by computing an attention score between a node i and its neighbor j.

The attention score between nodes i and j is computed as:

$$e_{ij} = LeakyReLU\left(a^T\left[h_i^{'}||h_j^{'}\right]\right) \tag{3}$$

Where:

a is a learnable attention vector (initialized randomly) and a^T is its transpose.

$||$ is concatenation, $h_j^{'}$ and $h_j^{'}$ are transformed feature vector of nodes i and j.

LeakyReLU is an activation function that assigns small non-zero values to negative inputs, ensuring the model continues learning effectively.

Softmax Normalization

Attention scores are normalized using the softmax function. It ensures that all attention coefficients sum to 1, allowing the model to focus on more relevant neighbors.

The normalized attention score between nodes i and j is computed as:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in N(i)}\exp(e_{ik})} \tag{4}$$

Where:

$N(i)$ is a Set of neighbors of node i (including itself).

Weighted Aggregation

Neighbouring node features are aggregated using the assigned attention weights. Weighted aggregation for node i is computed as:

$$h_i^{''} = \sum_{j\in N(i)}\alpha_{ij}h_j^{'} \tag{5}$$

Final Aggregation and Transformation

It is computed as:

$$h_i^{'''} = w^{'}h_i^{''} \tag{6}$$

Where:

$w^{'}$ is a learnable weight matrix for the final transformation (same for all nodes).

Finally, the updated node representations are passed to the next layer.

This approach enables GATs to efficiently capture complex molecular interactions, making them well-suited for drug efficacy prediction.

GAT-Based Framework for Predicting Drug Efficacy (GAT-PDE)

The Schematic representation of the Graph Attention Network-Based Framework for Predicting Drug Efficacy (GAT-FDE) is depicted in Figure 1.

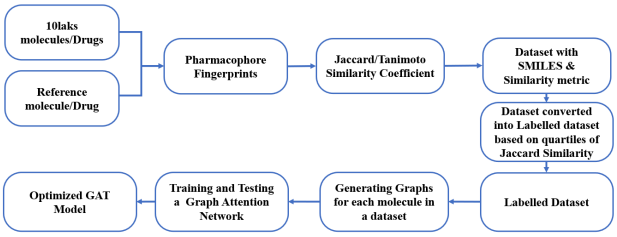

Fig. 1: Schematic representation of the Graph Attention Network-Based Framework for Predicting Drug Efficacy (GAT-FDE)

The proposed method includes the following steps.

DOI: 10.3844/jcssp.2025.2220.2228

Identifying Reference Drug

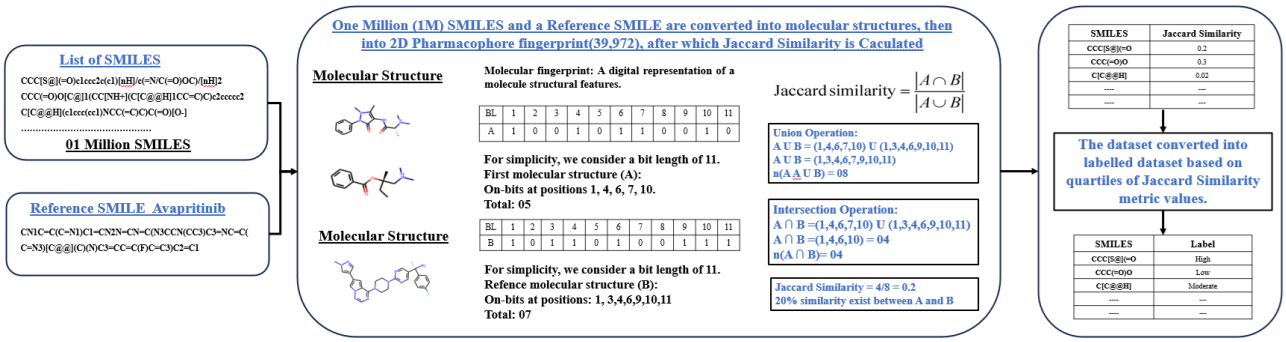

Fig. 2: Process of Converting SMILES to Pharmacophore Fingerprints and Computing Jaccard Similarity

Avapritinib has been identified as a reference drug, as it is a highly selective and potent tyrosine kinase inhibitor (TKI) that has demonstrated significant efficacy in patients with PDGFRA mutant GISTs, leading to improved progression-free survival (PFS) and objective response rates that effectively targets PDGFR mutations (Dhillon, 2020; Li *et al.*, 2023; Teuber *et al.*, 2024).

Molecular Fingerprints Generation

The Pharmacophore molecular fingerprints of length 39972 are generated for all the molecules (One Million) including the reference drug Avapritinib.

Similarity Metrics Calculation

Jaccard/Tanimoto Coefficient is calculated for all one million molecules with respect to the Avapritinib.

Converting Dataset into Labelled Dataset

The dataset is converted into labelled dataset based on quartiles of Jaccard Similarity metrics.

Generating Graphs

Graphs are generated for each molecule in the dataset.

GAT Model

Graph Attention Network algorithm is trained and evaluated the performance of the Model.

Results and Discussion

As per the process of GAT-Based framework for predicting drug efficacy, the SMILES representation of Avapritinib and one million molecules from a drug database are first converted into molecular structures using the RDKit Python library as depicted in Figure 2. Each molecular structure is then converted into a pharmacophore fingerprint of length 39,972 bits. The same process is applied to the reference SMILES, Avapritinib. For better understanding, the figure presents an example where molecular structures A and B are represented with a fingerprint length of 11 bits. Union and intersection operations are performed on fingerprints A and B, resulting in 08 elements in the union set and 04 in the intersection set. Finally, the Jaccard similarity is computed as 0.2.

The dataset consists of two columns: SMILES and the Jaccard similarity metric as depicted in Figure 3.

SMILES	Tanimoto_values
CCC[S@](=O)c1ccc2c(c1)[nH]/c(=N/C(=O)OC)/[nH]2	0.054
CCC(=O)O[C@]1(CC[NH+](C[C@@H]1CC=C)C)c2ccccc2	0.013
C[C@@H](c1ccc(cc1)NCC(=C)C)C(=O)[O-]	0.014
C[C@H](Cc1ccccc1)[NH2+][C@@H](C#N)c2ccccc2	0.004
C[C@@H](CC(c1ccccc1)(c2ccccc2)C(=O)N)[NH+](C)C	0.002
Cc1c(c(=O)n(n1C)c2ccccc2)NC(=O)[C@H](C)[NH+](C)C	0.029
c1ccc(cc1)[C@@H](C(=O)[O-])O	0.006

Fig. 3: Dataset of One Million SMILES with Jaccard Similarity

Figure 4 illustrates the summary statistics of Jaccard Similarity, showing its distribution across key metrics, including mean, variability, and percentile values, which highlights a generally low similarity range with a maximum value of approximately 0.40.

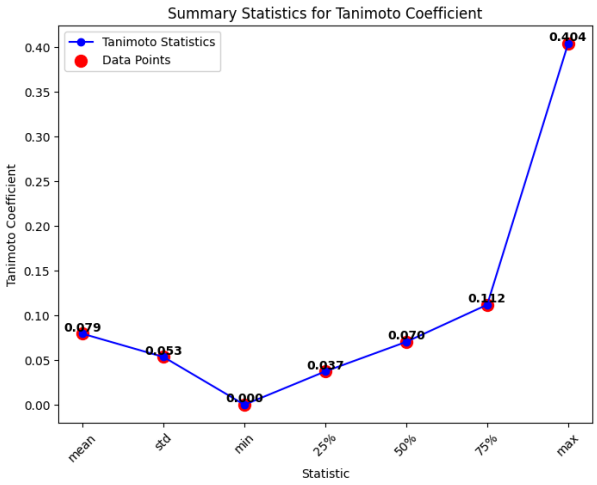

Fig. 4: Summary Statistics of Jaccard Similarity

Quartile analysis is used to divide the dataset into four equal groups based on Jaccard similarity values. These groups are labeled as low (0), medium (1), high

DOI: 10.3844/jcssp.2025.2220.2228

(2), and very high (3). This method ensures a balanced classification, making it easier to analyze patterns in molecular similarity. Each SMILES entry has a similarity value, which determines its category. If the value falls in the lowest range, it is classified as low (0), while higher values are placed into medium (1), high (2), or very high (3) categories. This process is applied to all SMILES, ensuring that each one is grouped according to its similarity score. Figure 5 depicts the dataset consisting of three columns SMILES, Jaccard Similarity and Label.

SMILES	Tanimoto_values	label
CCC[S@](=O)c1ccc2c(c1)[nH]/c(=N/C(=O)OC)/[nH]2	0.05	1
CCC(=O)O[C@]1(CC[NH+](C[C@@H]1CC=C)C)c2ccccc2	0.01	0
C[C@@H](c1ccc(cc1)NCC(=C)C)C(=O)[O-]	0.01	0
C[C@H](Cc1ccccc1)[NH2+][C@@H](C#N)c2ccccc2	0.00	0
C[C@@H](CC(c1ccccc1)(c2ccccc2)C(=O)N)[NH+](C)C	0.00	0
Cc1c(c(=O)n(n1C)c2ccccc2)NC(=O)[C@H](C)[NH+](C)C	0.03	0
c1ccc(cc1)[C@@H](C(=O)[O-])O	0.01	0
CC[C@](C)(C[NH+](C)C)OC(=O)c1ccccc1	0.00	0
COc1cc(c(c2c1OCO2)OC)CC=C	0.01	0
Cc1ccccc1NC(=O)[C@H](C)[NH+]2CCCC2	0.01	0
CC(=O)Oc1ccccc1C(=O)[O-]	0.01	0
C[NH+]1[C@@H]2CC[C@H]1CC(C2)OC(=O)[C@H](CO)c3c...	0.02	0
c1cc(ccc1[C@@H](CC(=O)[O-])C[NH3+])Cl	0.01	0
c1cc(ccc1C(=O)[O-])N[C@@H]2[C@@H]([C@H]([C@@H]...	0.09	2
C[C@@H](c1ccc2c(c1)nc(o2)c3ccc(cc3)Cl)C(=O)[O-]	0.06	1

Fig. 5: Dataset of One Million SMILES with Jaccard Similarity after labeling

The bar chart in Figure 6 represents the number of SMILES in each category, showing how they are distributed across the four similarity levels: low(0), medium(1), high(2), and very high(3).

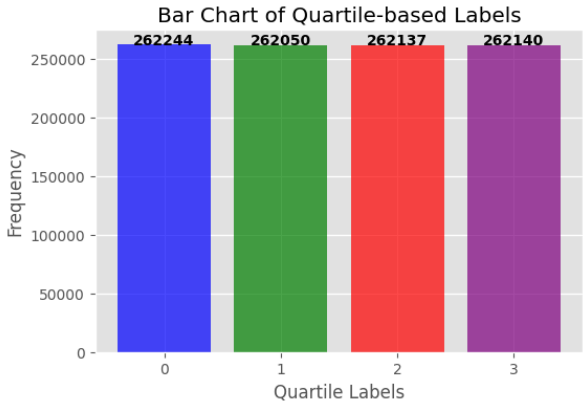

Fig. 6: Distribution of SMILES using Quartile Analysis based on Jaccard Similarity

Once the dataset is converted into categorical form, each SMILES string is transformed into a graph using the from_smiles() function from the Torch Geometric library.

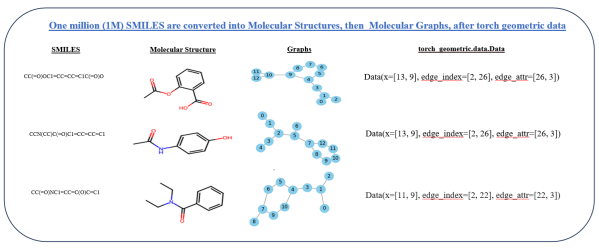

Fig. 7: Process of Converting a SMILES into Graphs

Figure 7 illustrates the process of converting a SMILES representation into a graph and then into a Torch Geometric object. This process is repeated for all one million SMILES, generating corresponding graph-based Torch Geometric objects.

The dataset is divided into training and testing sets, with 70% used for training and 30% for testing. The Graph Attentive Network (GAT) is trained on the training dataset using the AttentiveFP model from the Torch Geometric library. Figure 8 illustrates the process of training the GAT algorithm and validating the GAT model. After training, the model is tested on the test dataset and evaluated using performance metrics.

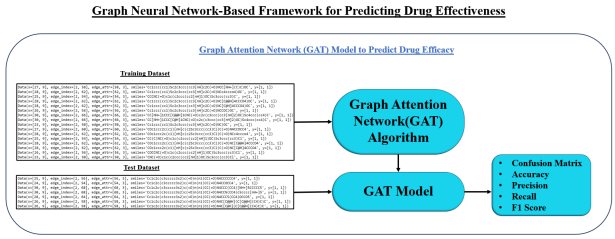

Fig. 8: Training and Validation Process of GAT Model

The GAT model was trained and tested for ten epochs. Throughout the training process, both training and testing accuracy gradually increased, while training and testing loss progressively decreased. Training accuracy improved from 80.45% to 87.28%, and test accuracy increased from 86.30% to 88.30%, as depicted in Figure 9. Similarly, training loss decreased from 0.4508 to 0.2995, while test loss reduced from 0.3236 to 0.2750, as depicted in Figure 10. These results demonstrate that the model effectively learned and improved its performance over time. Table 1 summarizes these trends over ten epochs.

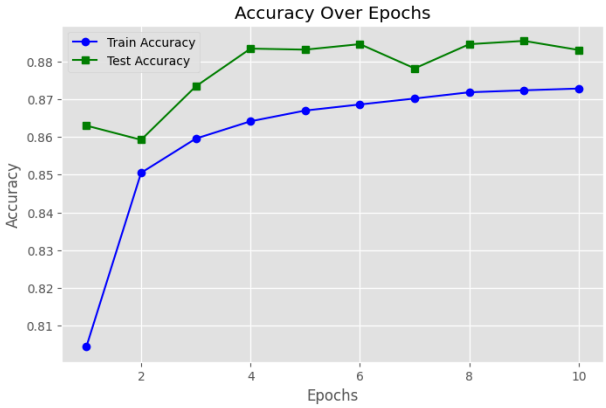

Fig. 9: Training and Testing Accuracy over 10 Epochs of GAT Model

Table 1: Accuracy and Loss of Training and Testing

Accuracy	Range	Loss	Range
Training Accuracy	80.45% to 87.28%	Training Loss	0.45 to 0.29
Testing Accuracy	86.30% to 88.30%	Testing Loss	0.32 to 0.27

Fig. 10: Training and Testing Loss over 10 Epochs of GAT Model

The confusion matrix for the test dataset (314,572 samples) is depicted in Figure 11. It displays the number of correct and incorrect predictions for each class, helping to assess the model's performance and identify misclassifications. Key metrics such as accuracy, precision, recall, and F1-score can be calculated from it for a more detailed evaluation as depicted in Figure 12.

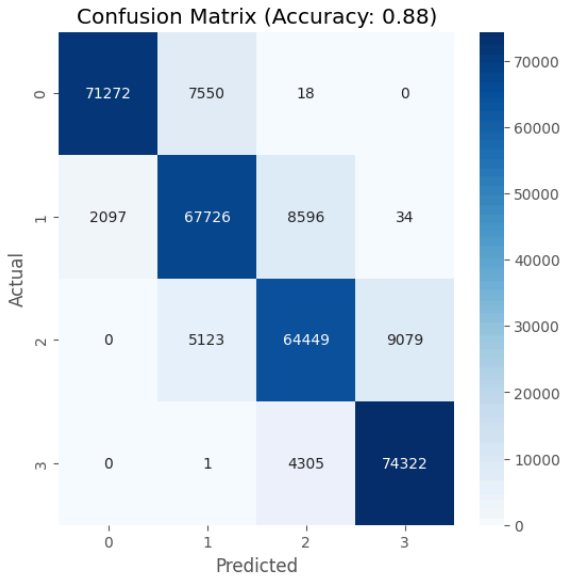

Fig. 11: Confusion Matrix of GAT Model for Test Dataset

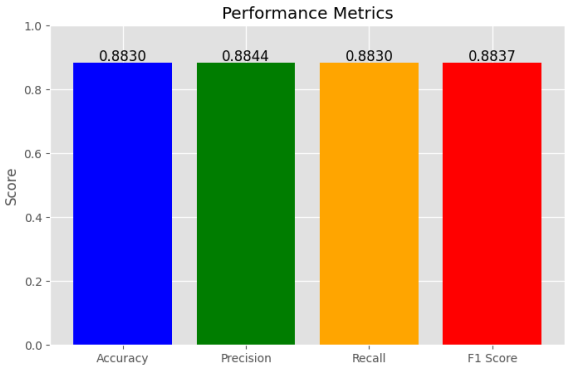

Fig. 12: Performance Metrics of GAT Model for Test Dataset

The proposed GAT-PDE model achieved an accuracy of 88.3%, and other performance metrics support its robustness in predicting drug efficacy.

Conclusion

Machine Learning (ML) and Deep Learning (DL) have significantly advanced various fields, including healthcare. However, the challenge of drug resistance persists which continues to demand novel drugs. Addressing the proposed GAT-PDE framework demonstrates a promising approach for predicting drug efficacy by leveraging Graph Attention Networks (GATs), pharmacophore fingerprints, Jaccard similarity, and quartile-based labelling. By utilizing Avapritinib as a reference drug, the model effectively categorizes molecular similarities and achieves high accuracy in drug efficacy prediction. The extension of this framework to NSCLC drug discovery, integrating molecular fingerprints and deep learning, further highlights its adaptability and potential impact. With an accuracy of 88% at 10 epochs, the framework shows promise for identifying effective drugs targeting PDGFR in NSCLC, and further training could enhance its predictive performance. This approach has the potential to accelerate drug discovery, reduce costs, and improve treatment outcomes for resistant diseases.

Future Scope

The current framework uses a single reference compound, Avapritinib, to evaluate molecular similarity and predict efficacy for the PDGFR target. Expanding this approach to include multiple targets and multiple reference drugs could offer a more comprehensive strategy, particularly for complex diseases like NSCLC. Integrating the framework into existing drug repurposing platforms may also help identify new uses for approved drugs, reducing both time and cost. Furthermore, collaborating with clinical researchers to test the model's predictions in laboratory or clinical settings would provide crucial validation and help refine the framework for real-world application.

Acknowledgement

We would like to thank GITAM (Deemed to be University) for conducting reviews to assess research progress and DRC members (Dr. Prem Kumar Singh, K Prasad Rao & Dr. M Rama Narasinga Rao) for providing valuable feedback and suggestions, which helped in the development of this work.

We thank the publisher for giving us the opportunity to share our work and contribute to the field through this publication, and also extend special thanks to the editorial team for reviewing and editing the article.

Authors' Contributions

Sandhi Kranthi Reddy: Developed the proposed framework, including its implementation,

experimentation, and analysis of results.

S. V. G. Reddy: Provided overall supervision, technical guidance, and critical feedback to refine the methodology and article.

Ethics

This article is original and has not been published elsewhere. All authors have read and approved the final version of the manuscript, and there are no ethical issues with this research.

References

Akshara, R., & Jain, A. (2024). Transforming E-commerce with a Novel Multifaceted Data-Decision Framework. *International Journal of Electronics and Communication Engineering*, *11*(9), 120–134. https://doi.org/10.14445/23488549/ijece-v11i9p112

Ali, S. J., Omer, M., Le, D. T., Raza, S. M., & Choo, H. (2025). Deep Learning for Drug Response Prediction with Gene Expression Data. *2025 International Conference on Information Networking (ICOIN)*. https://doi.org/10.1109/icoin63865.2025.10992878

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, *7*(1). https://doi.org/10.1186/s13321-015-0069-3

Biala, G., Kedzierska, E., Kruk-Slomka, M., Orzelska-Gorka, J., Hmaidan, S., Skrok, A., Kaminski, J., Havrankova, E., Nadaska, D., & Malik, I. (2023). Research in the Field of Drug Design and Development. *Pharmaceuticals*, *16*(9), 1283. https://doi.org/10.3390/ph16091283

Boniolo, F., Dorigatti, E., Ohnmacht, A. J., Saur, D., Schubert, B., & Menden, M. P. (2021). Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opinion on Drug Discovery*, *16*(9), 991–1007. https://doi.org/10.1080/17460441.2021.1918096

Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., Kim, T. S., Jung, J., & Shin, J.-M. (2018). Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-27214-6

Dhillon, S. (2020). Avapritinib: First Approval. *Drugs*, *80*(4), 433–439. https://doi.org/10.1007/s40265-020-01275-2

Gandhi, V. C., & Gandhi, P. P. (2022). A Survey - Insights of ML and DL in Health Domain. *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. https://doi.org/10.1109/icscds53736.2022.9760981

Garg, P., Malhotra, J., Kulkarni, P., Horne, D., Salgia, R., & Singhal, S. S. (2024). Emerging Therapeutic Strategies to Overcome Drug Resistance in Cancer Cells. *Cancers*, *16*(13), 2478. https://doi.org/10.3390/cancers16132478

Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Regep, C., Liu, G., Hayter, J. B. R., Vickers, R., Roberts, C., Tang, J., Roblin, D., Blundell, T. L., Bronstein, M. M., & Taylor-King, J. P. (2021). Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab159

Goswami, S., & Chakrabarti, A. (2012). Quartile clustering: a quartile based technique for generating meaningful clusters. *ArXiv:1203.4157*.

Hughes, J., Rees, S., Kalindjian, S., & Philpott, K. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, *162*(6), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x

Kamrani, A., Hosseinzadeh, R., Shomali, N., Heris, J. A., Shahabi, P., Mohammadinasab, R., Sadeghvand, S., Ghahremanzadeh, K., Sadeghi, M., & Akbari, M. (2023). New immunotherapeutic approaches for cancer treatment. *Pathology - Research and Practice*, *248*, 154632. https://doi.org/10.1016/j.prp.2023.154632

Khemani, B., Patil, S., Kotecha, K., & Tanwar, S. (2024). A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-023-00876-4

Krzywanski, J., Sosnowski, M., Grabowska, K., Zylka, A., Lasek, L., & Kijo-Kleczkowska, A. (2024). Advanced Computational Methods for Modeling, Prediction and Optimization—A Review. *Materials*, *17*(14), 3521. https://doi.org/10.3390/ma17143521

Kumar, V., & Roy, K. (2025). Embracing the changes and challenges with modern early drug discovery. *Expert Opinion on Drug Discovery*, *20*(4), 419–431. https://doi.org/10.1080/17460441.2025.2481259

Lavecchia, A. (2024). Advancing drug discovery with deep attention neural networks. *Drug Discovery Today*, *29*(8), 104067. https://doi.org/10.1016/j.drudis.2024.104067

Li, J., Zhang, X., Deng, Y., Wu, X., Zheng, Z., Zhou, Y., Cai, S., Zhang, Y., Zhang, J., Tao, K., Cui, Y., Cao, H., Shen, K., Yu, J., Zhou, Y., Ren, W., Qu, C., Zhao, W., Hu, J., … Shen, L. (2023). Efficacy and Safety of Avapritinib in Treating Unresectable or Metastatic Gastrointestinal Stromal Tumors: A Phase I/II, Open-Label, Multicenter Study. *The Oncologist*, *28*(2), 187-e114. https://doi.org/10.1093/oncolo/oyac242

Li, X.-S., Liu, X., Lu, L., Hua, X.-S., Chi, Y., & Xia, K. (2022). Multiphysical graph neural network (MP-GNN) for COVID-19 drug design. *Briefings in Bioinformatics*, *23*(4). https://doi.org/10.1093/bib/bbac231

Lv, Q., Zhou, F., Liu, X., & Zhi, L. (2023). Artificial intelligence in small molecule drug discovery from 2018 to 2023: Does it really work? *Bioorganic Chemistry*, *141*, 106894. https://doi.org/10.1016/j.bioorg.2023.106894

Mian, S. M., Khan, M. S., Shawez, M., & Kaur, A. (2024). Artificial Intelligence (AI), Machine Learning (ML) & Deep Learning (DL): A Comprehensive Overview on Techniques, Applications and Research Directions. *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*. https://doi.org/10.1109/icscss60660.2024.10625198

Muegge, I., & Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, *11*(2), 137–148. https://doi.org/10.1517/17460441.2016.1117070

Parasrampuria, D. A., Benet, L. Z., & Sharma, A. (2018). Why Drugs Fail in Late Stages of Development: Case Study Analyses from the Last Decade and Recommendations. *The AAPS Journal*, *20*(3). https://doi.org/10.1208/s12248-018-0204-y

Rahman, A., Debnath, T., Kundu, D., Khan, Md. S. I., Aishi, A. A., Sazzad, S., Sayduzzaman, M., & Band, S. S. (2024). Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health*, *11*(1), 58–109. https://doi.org/10.3934/publichealth.2024004

Saeed, R. F., Awan, U. A., Saeed, S., Mumtaz, S., Akhtar, N., & Aslam, S. (2023). *Targeted Therapy and Personalized Medicine*. 177–205. https://doi.org/10.1007/978-3-031-27156-4_10

Saihood, A. A., Hasan, M. A., Shnawa, S. M., Fadhel, M. A., Alzubaid, L., Gupta, A., & Gu, Y. (2024). Multiside graph neural network-based attention for local co-occurrence features fusion in lung nodule classification. *Expert Systems with Applications*, *252*, 124149. https://doi.org/10.1016/j.eswa.2024.124149

Sarker, I. H. (2021a). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Sarker, I. H. (2021b). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00592-x

Teuber, A., Schulz, T., Fletcher, B. S., Gontla, R., Mühlenberg, T., Zischinsky, M.-L., Niggenaber, J., Weisner, J., Kleinbölting, S. B., Lategahn, J., Sievers, S., Müller, M. P., Bauer, S., & Rauh, D. (2024). Avapritinib-based SAR studies unveil a binding pocket in KIT and PDGFRA. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-023-44376-8

Vaida, M., Wu, J., Himdiat, E., Haince, J.-F., Bux, R. A., Huang, G., Tappia, P. S., Ramjiawan, B., & Ford, W. R. (2025). M-GNN: A Graph Neural Network Framework for Lung Cancer Detection Using Metabolomics and Heterogeneous Graph Modeling. *International Journal of Molecular Sciences*, *26*(10), 4655. https://doi.org/10.3390/ijms26104655

Visan, A. I., & Negut, I. (2024). Integrating Artificial Intelligence for Drug Discovery in the Context of Revolutionizing Drug Delivery. *Life*, *14*(2), 233. https://doi.org/10.3390/life14020233

Vrahatis, A. G., Lazaros, K., & Kotsiantis, S. (2024). Graph Attention Networks: A Comprehensive Review of Methods and Applications. *Future Internet*, *16*(9), 318. https://doi.org/10.3390/fi16090318

Willett, P. (2009). Similarity methods in chemoinformatics. *Annual Review of Information Science and Technology*, *43*(1), 1–117. https://doi.org/10.1002/aris.2009.1440430108

Xie, C., Vanderbilt, C., Feng, C., Ho, D., Campanella, G., Egger, J., & Fuchs, T. (2022). *Computational biomarker predicts lung ICI response via deep learning-driven hierarchical spatial modelling from H&E*.

Yang, W., Zou, J., & Yin, L. (2022). *Compound Property Prediction Based on Multiple Different Molecular Features and Ensemble Learning*. 57–69. https://doi.org/10.1007/978-981-19-8300-9_7

Zhang, X.-M., Liang, L., Liu, L., & Tang, M.-J. (2021). Graph Neural Networks and Their Current Applications in Bioinformatics. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.690049

Zhu, J., Wang, J., Wang, X., Gao, M., Guo, B., Gao, M., Liu, J., Yu, Y., Wang, L., Kong, W., An, Y., Liu, Z., Sun, X., Huang, Z., Zhou, H., Zhang, N., Zheng, R., & Xie, Z. (2021). Prediction of drug efficacy from transcriptional profiles with deep learning. *Nature Biotechnology*, *39*(11), 1444–1452. https://doi.org/10.1038/s41587-021-00946-z