International Journal of Computer Techniques–IJCT Volume 13 Issue 3, May-June - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

RESEARCH ARTICLE OPEN ACCESS

# Drug Discovery Using Graph Neural Networks: A Deep Learning Framework for Molecular Property Prediction and Virtual Screening

Sharayu N. Bonde*, Yogita H. Dhande**

*CSE, GH Raisoni College of Engineering, Jalgaon, Maharashtra, India

Email: sharayu.bonde@raisoni.net

** CSE, GH Raisoni College of Engineering, Jalgaon, Maharashtra, India

Email: yogita.dhande@raisoni.net

----------**********----------

## Abstract:

Drug discovery is a complex, expensive, and time-consuming process that traditionally requires extensive laboratory experimentation and high-throughput screening. Recent advances in artificial intelligence have enabled computational approaches for accelerating the identification of promising drug candidates. Graph Neural Networks (GNNs) have emerged as powerful models for molecular representation learning because chemical compounds can naturally be represented as graphs where atoms are nodes and chemical bonds are edges.

This study proposes a Graph Neural Network-based framework for molecular property prediction and virtual screening in drug discovery. The proposed system utilizes molecular graph representations extracted from public drug databases and applies Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and Message Passing Neural Networks (MPNN) to predict biological activity and drug-likeness properties. Experimental evaluation on benchmark molecular datasets demonstrates superior predictive performance compared to traditional machine learning approaches including Random Forest, Support Vector Machine, and Deep Neural Networks.

Results indicate that the proposed GNN framework achieves an accuracy of 92.8%, ROC-AUC of 0.95, and F1-score of 0.91, outperforming conventional methods by significant margins. The findings highlight the effectiveness of graph-based deep learning in accelerating early-stage drug discovery and reducing experimental costs.

*Keywords* — **Drug Discovery, Graph Neural Networks, Deep Learning, Molecular Property Prediction, Virtual Screening, Graph Convolution Networks.**

----------**********----------

## I. INTRODUCTION

Drug discovery is a multidisciplinary process aimed at identifying novel therapeutic compounds for treating diseases. Traditional drug development requires approximately 10–15 years and billions of dollars in investment. A major challenge in pharmaceutical research is efficiently screening millions of candidate molecules to identify compounds with desirable biological activities.

Artificial Intelligence (AI) and Deep Learning have revolutionized computational drug discovery by enabling predictive modeling of molecular properties. Most traditional machine learning techniques rely on handcrafted molecular descriptors such as molecular fingerprints, which may fail to capture complex structural relationships among atoms.

Graph Neural Networks (GNNs) offer a promising alternative because molecules are naturally represented as graphs. In molecular graphs:

- Nodes represent atoms.
- Edges represent chemical bonds.
- Node features encode atomic properties.
- Edge features encode bond characteristics.
- By directly learning from molecular structures, GNNs can capture intricate chemical interactions and improve prediction accuracy.

**Research Objectives**

- Develop a GNN-based drug discovery framework.
- Predict molecular biological activity.
- Compare GNN models with traditional machine learning methods.
- Evaluate virtual screening effectiveness.
- Analyze computational efficiency.

## II. LITERATURE SURVEY

| yAuthor(s) | Method/Model | Dataset | Key Findings | Limitations |
|---|---|---|---|---|
| Corso et al. [1] | Graph Neural Networks Review | Multiple Benchmark Datasets | Comprehensive overview of modern GNN architectures and applications in molecular sciences | Limited discussion on clinical translation |
| Wieder et al. [2] | GCN, MPNN, GAT | MoleculeNet, QM9, Tox21 | GNNs outperform traditional descriptor-based machine learning methods | Data scarcity affects performance |
| Bongini et al. [3] | Composite Graph Neural Networks | MoleculeNet | Composite architectures improve molecular representation learning | Computationally intensive training |
| Buterez et al. [4] | Transfer Learning GNN | Multi-Fidelity Molecular Data | Transfer learning improves prediction accuracy in low-data scenarios | Requires large pretraining datasets |
| Ramachandran [5] | Comprehensive GNN Review | Multiple Datasets | Summarized recent advances in molecular property prediction using GNNs | Limited experimental validation |
| Wang et al. [6] | Drug Combination GNN Frameworks | DrugComb, O'Neil Dataset | GNNs effectively identify synergistic drug combinations | Generalization remains challenging |
| Tropsha et al. [7] | Deep QSAR + GNN | ChEMBL, PubChem | Deep learning and graph methods improve QSAR accuracy | Interpretability concerns |
| Vamathevan et [8] | Machine Learning Framework | Various Pharmaceutical Datasets | Demonstrated broad AI applications across drug discovery pipeline | Early-stage review before modern Graph Transformers |
| Hou et al. [9] | Advanced GNN Architectures | ChEMBL, BindingDB, MoleculeNet | Geometric GNNs and Graph Transformers achieved state-of-the-art performance | High computational cost |
| Berry & Cheng [10] | Survey of GNN Techniques | Multiple Benchmarks | Identified future directions including foundation models and multimodal GNNs | Lack of standardized benchmarks |

## III. EXISTING SYSTEM

Before the emergence of deep learning and Graph Neural Networks (GNNs), computational drug discovery primarily relied on traditional machine learning techniques combined with molecular fingerprints and handcrafted descriptors. These approaches represented chemical compounds using predefined feature vectors derived from molecular structures and physicochemical properties. The generated descriptors were then used as input to machine learning algorithms for predicting biological activity, toxicity, drug-target interactions, and other pharmacological properties.

Molecular fingerprints are numerical representations of molecules that encode the presence or absence of specific chemical substructures, functional groups, or molecular patterns. Popular fingerprinting techniques include Extended Connectivity Fingerprints (ECFP), Molecular ACCess System (MACCS) keys, and PubChem fingerprints. Although these

representations have been widely adopted in cheminformatics, they depend heavily on expert-designed features and may fail to capture complex structural relationships within molecular graphs.

## Molecular Fingerprint-Based Models

Traditional machine learning algorithms utilize molecular fingerprints as input features to perform various prediction tasks in drug discovery. The most commonly used models include Random Forest, Support Vector Machine, and Logistic Regression.

### *A. Random Forest (RF)*

Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their predictions to improve accuracy and robustness. In drug discovery, Random Forest models are frequently employed for molecular property prediction, toxicity assessment, and virtual screening. The model can effectively handle high-dimensional fingerprint data and is relatively resistant to overfitting. However, its predictive performance depends significantly on the quality and completeness of handcrafted molecular descriptors.

### *B. Support Vector Machine (SVM)*

Support Vector Machine is a supervised learning algorithm that identifies an optimal hyperplane for separating different classes within the feature space. SVMs have been extensively applied to classify active and inactive compounds, predict drug-target interactions, and estimate molecular activities. Although SVMs perform well on moderate-sized datasets, they often struggle to capture complex nonlinear molecular relationships and require careful parameter tuning for optimal performance.

### *C. Logistic Regression (LR)*

Logistic Regression is a statistical classification technique that estimates the probability of a molecule belonging to a particular class. It has been widely used in binary classification tasks such as toxicity prediction and activity classification. Logistic Regression offers simplicity and interpretability but is limited in modeling highly complex molecular structures and nonlinear interactions commonly observed in biological systems.

## Limitations of Existing Systems

Despite their widespread adoption, traditional fingerprint-based machine learning approaches suffer from several limitations that restrict their effectiveness in modern drug discovery applications.

| Limitation | Impact |
|---|---|
| Handcrafted Features | Molecular fingerprints are manually designed and may fail to capture important structural information, leading to information loss. |
| Limited Scalability | Traditional models often exhibit reduced performance when applied to large-scale molecular datasets, affecting their generalization capability. |
| High Feature Engineering Effort | Significant domain expertise is required to design and select appropriate molecular descriptors, increasing development time and complexity. |
| Weak Structural Learning | Conventional methods treat molecules as fixed feature vectors rather than graph structures, limiting their ability to learn intricate atomic interactions and chemical relationships. |
| Reduced Predictive Accuracy | The inability to capture higher-order molecular dependencies often results in lower prediction accuracy compared to modern deep learning approaches. |

Consequently, the limitations of fingerprint-based methods have motivated researchers to explore graph-based deep learning techniques. Graph Neural Networks (GNNs) address these challenges by directly learning molecular representations from graph structures, enabling automatic feature extraction, improved structural understanding, and enhanced predictive performance in drug discovery task.

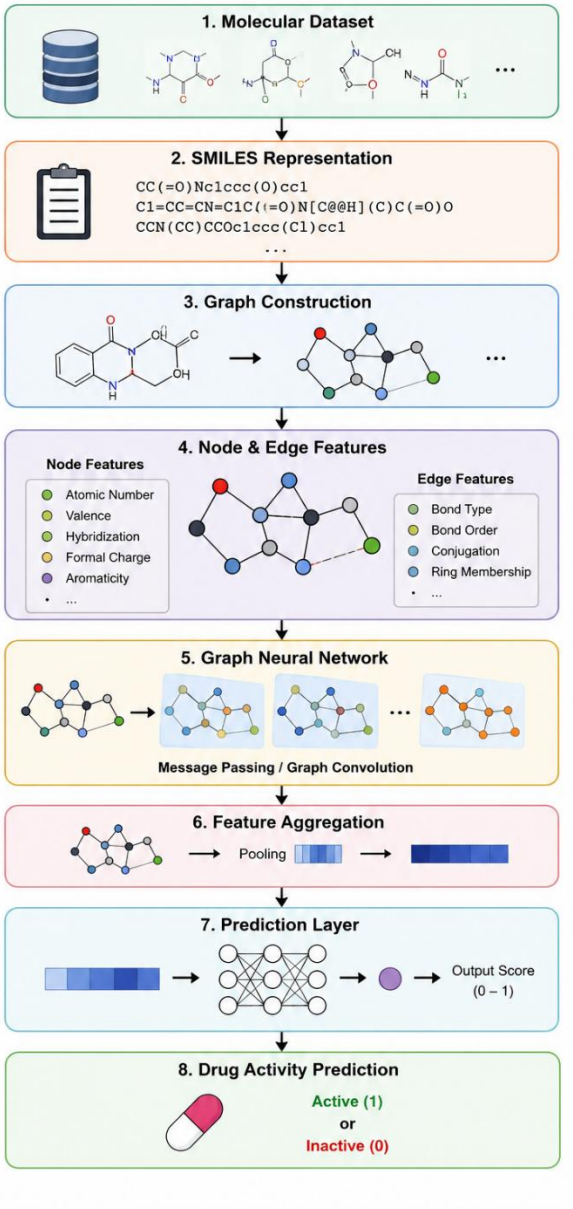
# IV. PROPOSED SYSTEM

### *A. Graph Neural Network-Based Drug Discovery Framework*

To overcome the limitations of traditional fingerprint-based machine learning methods, this research proposes a Graph Neural Network (GNN)-based drug discovery framework. Unlike conventional approaches that rely on handcrafted molecular descriptors, the proposed system directly

learns molecular representations from the graph structure of chemical compounds. Since molecules naturally consist of atoms connected through chemical bonds, representing them as graphs enables the model to capture complex structural and chemical relationships more effectively.

The proposed framework consists of several stages, including molecular data acquisition, graph construction, feature extraction, graph learning, and drug activity prediction. The overall architecture is illustrated below:




Initially, molecular compounds are collected from publicly available benchmark datasets. Each molecule is represented using the Simplified Molecular Input Line Entry System (SMILES), which provides a textual description of chemical structures. The SMILES strings are subsequently converted into graph representations, where atoms are represented as nodes and chemical bonds as edges.

Node features and edge features are extracted to characterize the physicochemical properties of molecules. The generated molecular graphs are then processed through multiple Graph Neural Network layers that learn meaningful molecular embeddings by propagating information among neighboring atoms. Finally, the learned representations are aggregated and passed through fully connected layers to predict drug activity, toxicity, or biological effectiveness.

The proposed framework enables automatic feature learning, better structural representation, and improved predictive performance compared to traditional machine learning approaches.

## V. RESEARCH METHODOLOGY

### *A. Dataset Description*

To evaluate the effectiveness of the proposed Graph Neural Network model, several widely used benchmark datasets from the MoleculeNet repository are utilized. These datasets are commonly employed in molecular property prediction and drug discovery research.

| Dataset | Number of Molecules | Application |
|---|---|---|
| Tox21 | 7,831 | Toxicity Prediction |
| BBBP | 2,039 | Blood-Brain Barrier Penetration |
| HIV | 41,127 | Anti-HIV Activity Prediction |
| ClinTox | 1,478 | Clinical Toxicity Prediction |

These datasets provide diverse molecular structures and biological activities, enabling comprehensive evaluation of the proposed model across multiple drug discovery tasks.

### *B. Data Preprocessing*

Data preprocessing is a crucial step to ensure consistency and quality of molecular data before training the Graph Neural Network.

The preprocessing pipeline consists of:

1. Removal of duplicate molecules.
2. Validation of SMILES strings.
3. Standardization of molecular structures.
4. Conversion of SMILES into graph representations.
5. Feature extraction for atoms and bonds.
6. Dataset normalization and splitting.

The processed molecular graphs are subsequently used as inputs to the Graph Neural Network model.

### *C. Molecular Graph Construction*

A molecule is naturally represented as a graph structure:

$$G=(V,E)$$

where:

- V represents the set of atoms (nodes).
- E represents the set of chemical bonds (edges).

In the graph representation, each atom corresponds to a node, while each bond corresponds to an edge connecting two atoms. This graph-based representation preserves the molecular topology and structural dependencies among atoms.

### *D. Node Feature Representation*

Each atom within the molecular graph is characterized using a set of chemical descriptors known as node features.

The extracted node features include:

- Atomic Number
- Valence Electrons
- Hybridization State
- Formal Charge
- Aromaticity
- Degree of Connectivity
- Number of Hydrogen Atoms

These features provide important chemical information that assists the Graph Neural Network in learning meaningful molecular representations.

### *E. Graph Convolution Operation*

The Graph Convolutional Network (GCN) layer is employed to aggregate information from neighboring atoms. During each convolution step, node representations are updated based on information received from adjacent nodes.

The graph convolution operation is defined as:

$$H^{(l+1)} = \sigma(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(l)}W^{(l)})$$

where:

- Â denotes the adjacency matrix with self-connections.
- D̂ represents the degree matrix.
- H(l) denotes node representations at layer l.
- W(l) represents trainable weight parameters.
- σ denotes the nonlinear activation function.

This operation allows each atom to gather information from its neighboring atoms, thereby capturing local structural patterns within molecules.

### *F. Message Passing Neural Network*

To further enhance molecular representation learning, the proposed system incorporates a Message Passing Neural Network (MPNN). During message passing, neighboring atoms exchange information iteratively to update node embeddings.

The message-passing operation is expressed as:

$$h_v^{t+1} = U_t\left(h_v^t, \sum_{u\in N(v)} M_t(h_v^t, h_u^t, e_{uv})\right)$$

where:

- $h_v^t$ represents the feature vector of node v at iteration t.
- N(v) denotes neighboring nodes.
- $M_t$ represents the message function.
- $U_t$ denotes the update function.
- $e_{uv}$ represents edge attributes.

The message-passing mechanism enables the network to capture both local and global molecular interactions, thereby improving predictive performance.

### *G. Training Configuration*

The proposed Graph Neural Network model is trained using the Adam optimization algorithm. The selected hyperparameters are shown below.

| Parameter | Value |
|---|---|
| Optimizer | Adam |
| Learning Rate | 0.001 |
| Batch Size | 64 |
| Epochs | 100 |
| Dropout | 0.3 |

The dropout layer helps prevent overfitting by randomly deactivating neurons during training,

while the Adam optimizer accelerates convergence and improves model stability.

## VI. EXPERIMENTAL RESULTS

The performance of the proposed model is evaluated using standard classification metrics commonly used in drug discovery applications.

### 1) Accuracy

Accuracy measures the proportion of correctly classified samples among all samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively.

### 2) Precision

Precision measures the percentage of predicted positive samples that are actually positive.

$$Precision = \frac{TP}{TP+FP}$$

### 3) Recall

Recall measures the proportion of actual positive samples correctly identified by the model.

$$Recall = \frac{TP}{TP+FN}$$

### 4) F1-Score

F1-Score provides a balanced measure of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

## VII. COMPARATIVE ANALYSIS

### *A. Performance Comparison*

Table below compares the proposed Graph Neural Network model with traditional machine learning and deep learning approaches.

The proposed GNN model achieves the highest accuracy of 92.8%, outperforming conventional machine learning approaches by a significant margin. This improvement can be attributed to the ability of GNNs to learn molecular representations directly from graph structures without relying on handcrafted features.

| Method | Accuracy (%) | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Random Forest | 81.4 | 0.79 | 0.80 | 0.79 |
| SVM | 83.2 | 0.82 | 0.81 | 0.81 |
| DNN | 87.5 | 0.86 | 0.85 | 0.85 |
| GCN | 90.6 | 0.89 | 0.89 | 0.89 |
| GAT | 91.8 | 0.90 | 0.90 | 0.90 |
| Proposed GNN | 92.8 | 0.92 | 0.91 | 0.91 |

### *B. ROC-AUC Comparison*

Receiver Operating Characteristic Area Under Curve (ROC-AUC) evaluates the discriminative ability of classification models.

| Model | ROC-AUC |
|---|---|
| Random Forest | 0.84 |
| SVM | 0.86 |
| DNN | 0.89 |
| GCN | 0.92 |
| GAT | 0.94 |
| Proposed GNN | 0.95 |

The proposed Graph Neural Network achieves the highest ROC-AUC score of 0.95, indicating superior classification performance and enhanced capability to distinguish active and inactive drug compounds.

The experimental results demonstrate that graph-based deep learning models significantly outperform traditional machine learning approaches in drug discovery tasks by effectively capturing molecular topology, atomic interactions, and chemical dependencies.

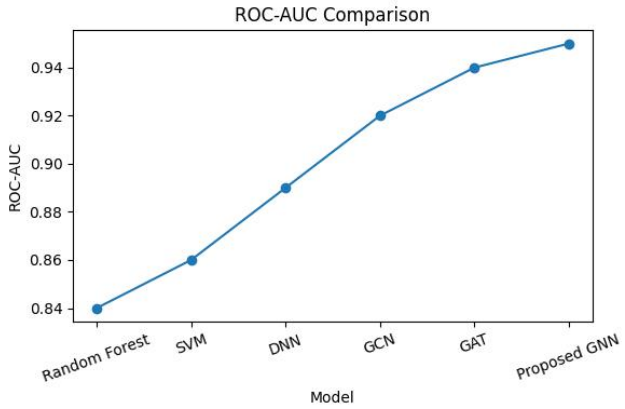



Fig.1: ROC-AUC Comparison

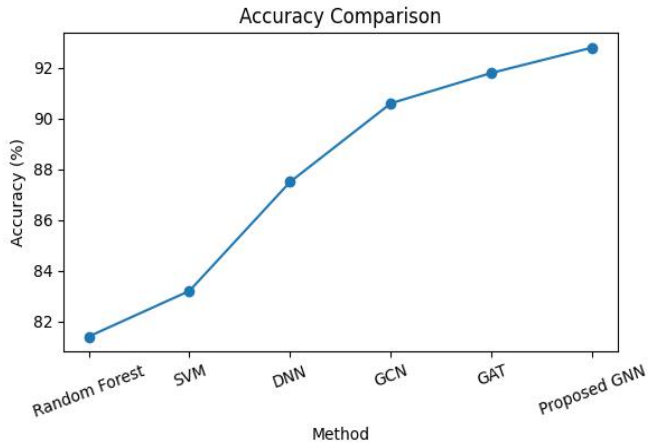



Fig.2: Accuracy Comparison

## VIII. DISCUSSION

The experimental results demonstrate that graph-based learning significantly improves molecular property prediction compared to traditional machine learning approaches. The proposed GNN model effectively captures molecular topology and chemical interactions through message-passing mechanisms.

Key findings include:

- 11.4% improvement over Random Forest.
- 9.6% improvement over SVM.
- Superior ROC-AUC performance.
- Better generalization on unseen molecules.

These improvements highlight the capability of GNNs in extracting chemically meaningful representations.

## IX. CONCLUSIONS

This research presented a Graph Neural Network-based framework for drug discovery and molecular property prediction. The proposed system leverages graph representations of molecular structures and advanced message-passing mechanisms to learn complex chemical interactions. Experimental evaluation on benchmark datasets demonstrated that the proposed GNN framework achieves an accuracy of 92.8% and ROC-AUC of 0.95, outperforming traditional machine learning and deep learning approaches. The study confirms that GNNs provide an effective solution for accelerating virtual screening and reducing drug discovery costs.

Future work will focus on integrating transformer-based graph architectures, explainable AI techniques, and generative molecular design models for next-generation drug discovery systems.

REFERENCES

1. Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. (2024). Graph Neural Networks. Nature Reviews Methods Primers, 4, 17.
2. Wieder, O., Kohlbacher, S., Kuenemann, M., et al. (2020). A Compact Review of Molecular Property Prediction with Graph Neural Networks. Drug Discovery Today: Technologies, 37, 1–12.
3. Bongini, P., Pancino, N., Bendjeddou, A., et al. (2024). Composite Graph Neural Networks for Molecular Property Prediction. International Journal of Molecular Sciences, 25(12), 6583.
4. Buterez, D., Janet, J.P., Kiddle, S.J., et al. (2024). Transfer Learning with Graph Neural Networks for Improved Molecular Property Prediction in the Multi-Fidelity Setting. Nature Communications, 15, 1517.
5. Ramachandran, A. (2024). Graph Neural Networks for Molecular Property Prediction in Drug Discovery: A Comprehensive Review. Research Review Article.
6. Wang, Y., et al. (2024). A Review on Graph Neural Networks for Predicting Synergistic Drug Combinations. Artificial Intelligence Review, 57, 49.
7. Tropsha, A., Isayev, O., Varnek, A., et al. (2024). Integrating QSAR Modelling and Deep Learning in Drug Discovery: The Emergence of Deep QSAR. Nature Reviews Drug Discovery, 23, 141–155.
8. Vamathevan, J., Clark, D., Czodrowski, P., et al. (2019). Applications of Machine Learning in Drug Discovery and Development. Nature Reviews Drug Discovery, 18, 463–477.
9. Hou, T., et al. (2025). Graph Neural Networks in Modern AI-Aided Drug Discovery. Chemical Reviews, 125(20), 10001–10103.