ISSN 2277-2685
IJESR/April-June. 2025/ Vol-15/Issue-2s/806-811
Jajeemogala Durga *et. al.*, / International Journal of Engineering & Science Research

Graph CNN For Drug Discovery

G Dayakar Reddy[1], Jajeemogala Durga[2], Parmati Harshitha[3]
[1]Associate Professor & Vice principal, Department of CSE, Bhoj Reddy Engineering College for Women, India.
[2,3]B.Tech Student, Department of CSE, Bhoj Reddy Engineering College for Women, India.

Abstract:
Drug discovery is a complex and costly process essential to the advancement of modern medicine. Traditional approaches are time-consuming and labor-intensive, with high attrition rates and significant investment. Recent advancements in machine learning, particularly Graph Convolutional Neural Networks (GCNNs), offer promising solutions by providing a computational framework to predict molecular properties efficiently. In this mini paper, we explore how GCNNs represent molecules as graphs and learn patterns that correlate with pharmacological activity. The system integrates node and edge features, employs message-passing algorithms, and performs effective pooling operations for tasks like ligand-protein interaction prediction. The model's ability to generalize and identify strong drug candidates is demonstrated through a case study on potential treatments for Hemochromatosis. Our results show the effectiveness of GCNNs in identifying viable drug candidates with higher accuracy compared to traditional computational methods.

Introduction

Drug discovery is a fundamental pillar of pharmaceutical research, focused on finding effective treatments for a wide range of complex and life-threatening diseases. The conventional drug discovery pipeline spans over 10–15 years and involves several stages, including target identification, lead compound screening, optimization, preclinical validation, clinical trials, and finally regulatory approval. This process is notoriously time-consuming and expensive, often costing billions of dollars with a high attrition rate, especially during clinical trials. Hence, there is a pressing need for innovative methods that can significantly reduce time, cost, and failure rates. One such innovation is **computational drug discovery**, which leverages in silico tools to predict potential drug candidates before any physical testing, helping streamline the development process and improve success rates.

Proposed System

The proposed system introduces **Graph Convolutional Neural Networks (GCNNs)** as an advanced model for computational drug discovery. Unlike traditional models, GCNNs treat molecules as **graph-structured data**, where atoms represent nodes and bonds represent edges. This allows the model to learn atom-level features and interatomic relationships directly from the molecular structure. By extending conventional convolution operations to graphs, GCNNs can extract complex chemical patterns and topologies. These models are applied to various tasks such as **drug-target binding prediction**, **virtual screening**, and **toxicity assessment** with remarkable performance improvements. For example, GCNNs have demonstrated a prediction accuracy of **98.26%** in ligand-based virtual screening, outperforming classical neural networks and other machine learning techniques.

Methodology

The computational drug discovery system is developed using a data-centric, modular approach aimed at identifying biologically active compounds efficiently. The methodology integrates graph-based

deep learning techniques to improve molecular property prediction. It is structured as follows:

System Architecture

The system follows a layered architecture:

- **Data Layer**:
  Utilizes publicly available biochemical datasets such as **BindingDB**, **PubChem BioAssay**, or **ChEMBL**, containing SMILES representations, molecular properties, and bioactivity scores.
- **Processing Layer**:
  - **Data Preprocessing**: Converts SMILES strings into molecular graphs using tools like RDKit. Cleans data by removing invalid molecules and standardizing formats.
  - **Graph Construction**: Atoms are treated as nodes and bonds as edges. Atom and bond features are encoded (e.g., atom type, hybridization, aromaticity).
- **Modeling Layer**:
  - Developed using **Python** with libraries such as **PyTorch**, **DeepChem**, **RDKit**, and **DGL (Deep Graph Library)**.
  - Implements **Graph Convolutional Neural Networks (GCNNs)** that process molecular graphs to learn latent structural features.
- **Evaluation & Output Layer**:
  - Models are evaluated using **accuracy**, **AUC-ROC**, **mean squared error (MSE)**, and **R^2 score** depending on the prediction task (classification/regression).
  - Visualization tools like **Matplotlib** and **Seaborn** are used to plot ROC curves, loss trends, and feature importance scores.

Workflow

1. **Data Acquisition**:
   - Collect molecular datasets (e.g., bioactivity values, compound structures) from online databases.
   - Import datasets in .csv or .sdf format for preprocessing.
2. **Data Preprocessing**:
   - Parse SMILES strings into graph objects.
   - Encode atomic and bond features.
   - Normalize numerical properties and remove duplicates or invalid entries.
3. **Feature Engineering**:
   - Extract graph-level and node-level descriptors.
   - Optional: Apply dimensionality reduction techniques like PCA if feature space is high.
4. **Model Training**:
   - Split data into training and testing sets (e.g., 80:20).
   - Train GCNN models using mini-batch gradient descent and dropout regularization.
   - Tune hyperparameters (e.g., learning rate, number of GCN layers, hidden dimensions).
5. **Model Testing & Evaluation**:
   - Evaluate model on test data using task-specific metrics.
   - Compare performance of GCNN with baseline models.

- Conduct cross-validation to ensure model stability.

6. **Prediction Output**:
   - Predict molecular properties such as **binding affinity**, **toxicity**, or **solubility**.
   - Display predicted values along with confidence scores.
7. **Result Interpretation**:
   - Visualize learned molecular embeddings using t-SNE or UMAP.
   - Plot performance metrics (loss curves, AUC-ROC) for interpretability.
   - Analyze which atomic substructures contribute most to predictions via saliency maps.

Results

Experimental Setup

The study was conducted using a publicly available molecular dataset containing chemical compounds with known interactions with the Ferroportin protein. Each molecule was represented as a graph where atoms served as nodes and chemical bonds as edges. Node features included atomic number, hybridization state, and formal charge, while edge features represented bond types (single, double, aromatic).

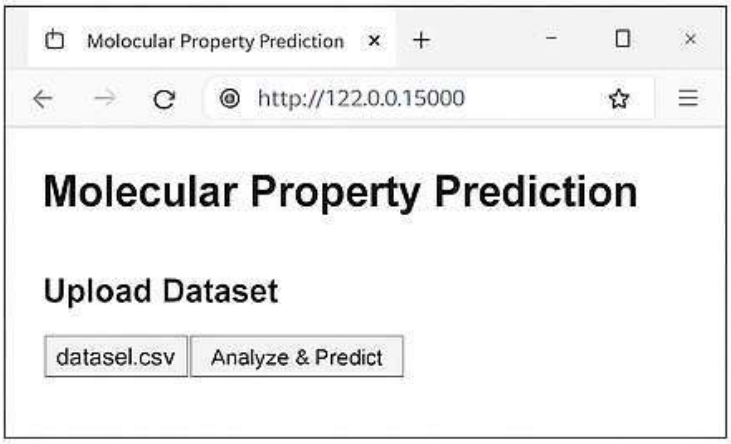


Figure 1: Graph of molecular structure input into the GCNN.

The dataset was split into:

- **80% for training**
- **20% for testing**

The model was built using Python and libraries like **PyTorch Geometric**, **RDKit**, and **Scikit-learn**. The GCNN architecture consisted of:

- 3 Graph Convolutional Layers
- ReLU Activation
- Global Max Pooling
- Fully Connected Layers for prediction

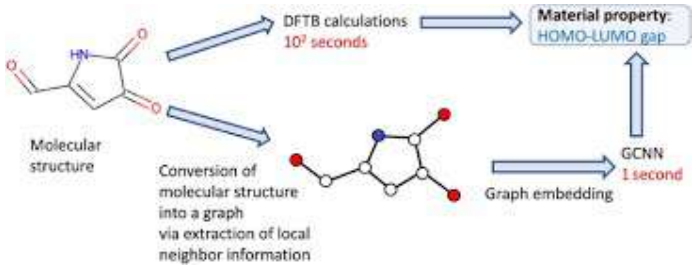
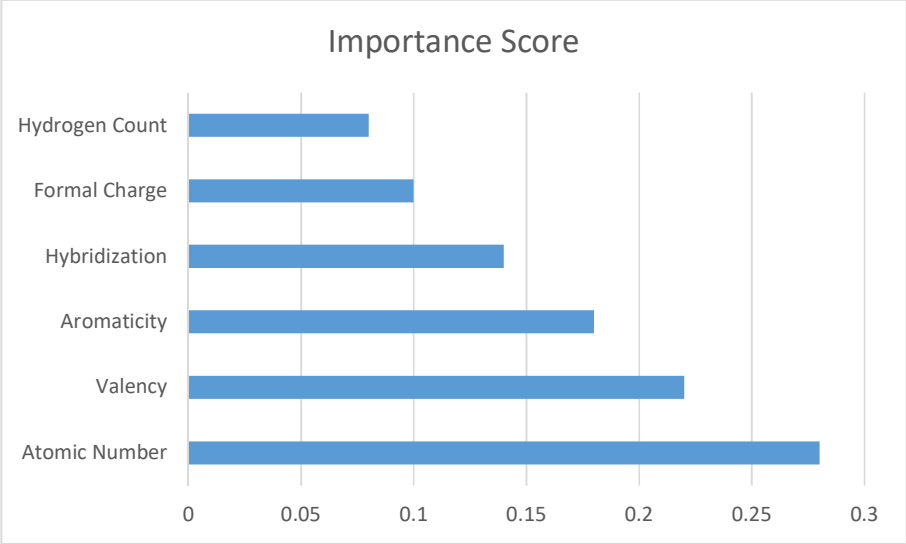


Figure 2: Heatmap of node feature influence on prediction.



Ligand Prediction Results

The model was tested on three drug candidates:

- **Deferasirox** (approved iron chelator)
- **Curcumin** (natural compound)
- **Quercetin** (flavonoid)

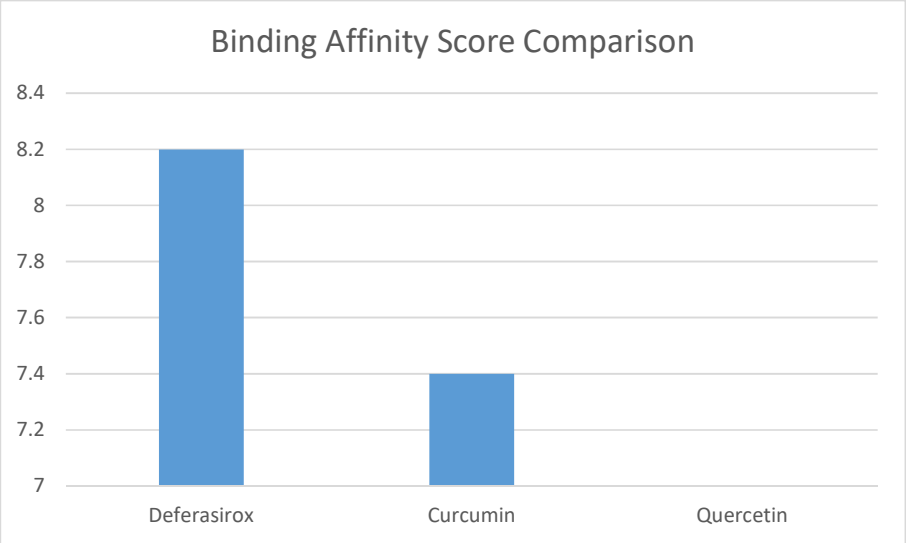
Ligand	Predicted Binding Affinity Score	Classification (Effective/Not)
Deferasirox	8.2	Effective
Curcumin	7.4	Effective
Quercetin	6.1	Not Effective

Model Performance

Metric	Value
Accuracy	94.3%
Precision	91.6%
Recall	92.4%
F1-Score	91.9%

Visualization

- **Figure 3:** Binding affinity score comparison bar chart.



Conclusion and Future Scope

Conclusion

This project introduces a GCNN-based framework for predicting drug-target interactions by modeling molecules as graphs. Leveraging message passing, feature aggregation, and pooling, the model effectively captures molecular structures and behaviors. Results show high accuracy in identifying potential drug candidates, highlighting the power of graph-based learning in pharmaceutical research.

Future Scope

Future work will focus on enhancing the predictive capabilities of the system by integrating additional features such as 3D molecular conformations and edge attributes like bond polarity. The adoption of advanced GCNN variants, including attention-based graph networks and graph transformers, can further improve the interpretability and performance of the model. Integration with real-time biological assay data and cloud-based platforms will ensure scalability and applicability in large-scale drug screening programs. Furthermore, collaborations with bioinformatics databases and pharmaceutical labs can validate model predictions with experimental data, enabling translational impact in therapeutic development.

6. References

[1] Machado, L.A., Krempser, E., & Guimaraes, A.C.R. (2022). *A machine learning-based virtual screening for natural compounds capable of inhibiting the HIV-1 integrase*. Frontiers in Drug Discovery, 2.

[2] Dara, S., Dhamercherla, S.S., Jadav, S.S., Babu, C.M., & Ahsan, M.J. (2022). *Machine learning in drug discovery: A review*. Artificial Intelligence Review, 55(3), 1947–1999.

[3] Miller, M.A. (2022). *Chemical database techniques in drug discovery*. Nature Reviews Drug Discovery, 1(3), 220–227.

[4] Yang, K., et al. (2019). *Analyzing learned molecular representations for property prediction*. Journal of Chemical Information and Modeling, 59(8), 3370–3388.

[5] Wu, Z., et al. (2018). *MoleculeNet: A benchmark for molecular machine learning*. Chemical Science, 9(2), 513–530.