Exploring Graph Neural Networks for Molecular Property Prediction

Ramesh Raj S [1] Prof. Tarak K Patra [1]

[1]Indian Institute of Technology, Madras

Introduction

Graph Neural Networks (GNNs) have become powerful tools for modeling molecular structures by leveraging their graph representation. Predicting molecular properties, such as free energy, is critical in computational chemistry, supporting drug discovery, materials science, and reaction modeling. While transformer-based models like ChemBERTa set the benchmark for molecular property prediction, they are computationally expensive. In contrast, GNNs offer a more efficient alternative by explicitly capturing the topology of molecular graphs.

Data

QM9 Dataset: A comprehensive dataset of 134,000 stable organic molecules with up to 9 heavy atoms, sourced from the GDB-17 database. It includes geometric, energetic, electronic, and thermodynamic properties.

SMILES Representation: Molecules in QM9 are represented using SMILES (Simplified Molecular Input Line Entry System), a text-based notation encoding molecular structure.

Methodology

Data Featurization:

RDKit parses SMILES strings and generates molecular graphs.

DeepChem is then used for converting the molecular graph into node and edge features. The `MolGraphConvFeaturizer` from DeepChem is utilized to generate these features.

Node features represent atomic properties (e.g., atom type, hybridization), which are one-hot encoded and concatenated to form feature vectors for each atom (node) in the graph.

Edge features represent bond properties (e.g., bond type, conjugation), which are also one-hot encoded and concatenated for each bond (edge) in the graph.

Edge index stores the graph structure, representing the connectivity between atoms (nodes), indicating the source and target nodes for each bond (edge).

This graph representation, consisting of node features, edge features, and edge indices, serves as the input to the GNN.

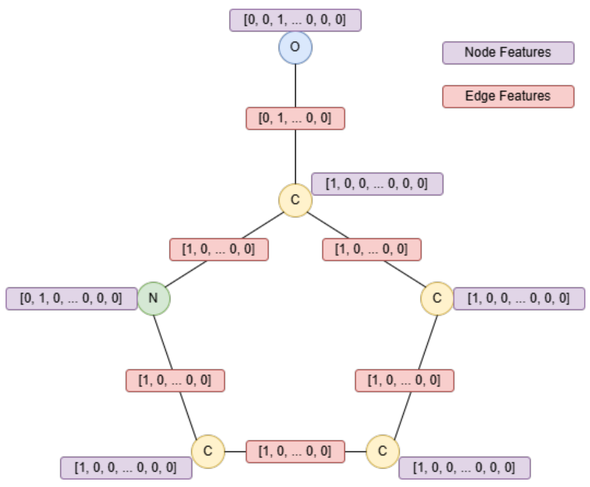

Figure 1. Graph representation of 2-Pyrrolidinone (SMILES: C1CCNC1=0)

Graph Neural Network:

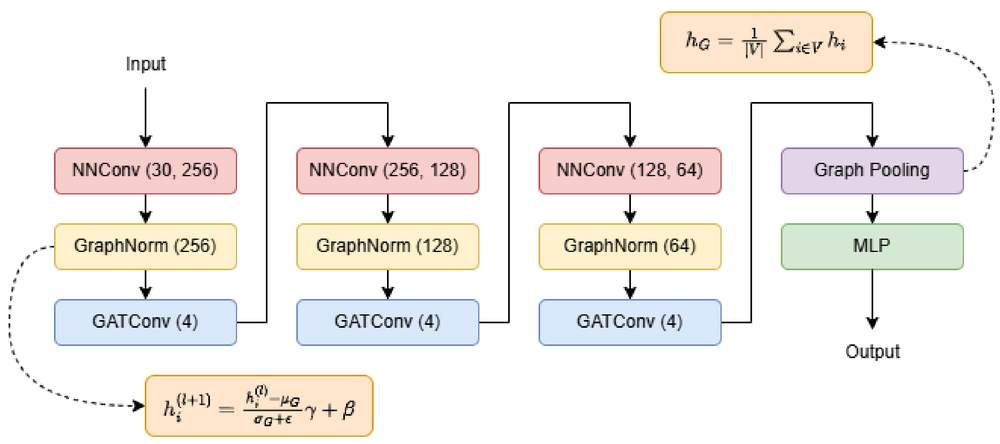

Figure 2. Model Architecture

NNConv (Figure 3): The NNConv layer uses a neural network for message passing between nodes. It leverages the features of neighboring nodes and edges, applying learned transformations to propagate information across the graph.

GATConv (Figure 3): The Graph Attention Network layer introduces attention mechanisms into the graph convolution process. It assigns different attention weights to neighboring nodes, allowing the model to focus on more relevant neighbors when aggregating information.

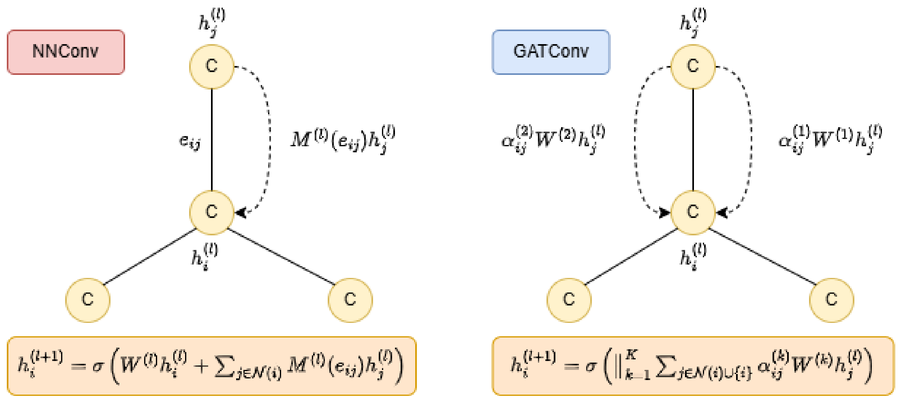

Figure 3. NNConv and GATConv layers

GraphNorm: GraphNorm is applied to normalize the graph's node features. This layer helps stabilize the learning process and improves convergence by reducing internal covariate shift.

Graph Pooling: After the convolutional layers, Graph Pooling (using mean aggregation) is applied to reduce the graph's node features to a fixed-size representation.

MLP for Prediction: The pooled representation is passed through a Multi-Layer Perceptron to predict molecular properties (e.g., free energy).

Methodology

Architectural Variations:

Layer Configurations:

Two-layer architecture with node feature dimensions: 128 → 64.

Three-layer architecture with node feature dimensions: 256 → 128 → 64.

Attention Mechanism: Models are trained with and without attention. Attention-based architectures include 1, 2, or 4 attention heads.

Results

The parallel coordinates plot (Figure 4) visualizes the performance of different GNN architectures. The 3-layer model with 4 attention heads achieved the best predictive performance.

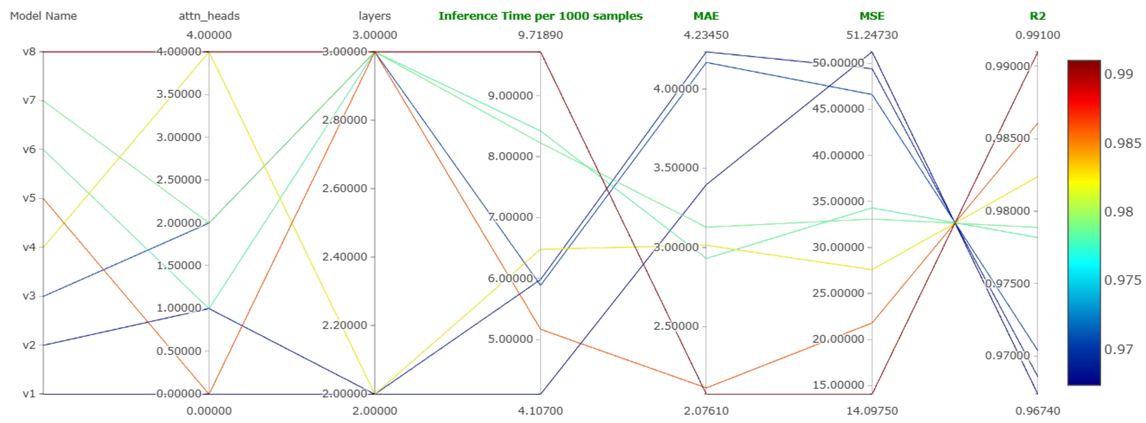

Figure 4. Performance of different GNN Architectures

The R^2 scores (Figure 5) indicate that GNNs achieve predictive performance comparable to ChemBERTa, a transformer-based model optimized for molecular tasks.

While ChemBERTa consistently maintains slightly higher R^2 values, the difference remains minimal, demonstrating that GNNs can effectively capture molecular patterns despite being less parameter-intensive.

As the training data fraction decreases, both models exhibit a decline in performance, but GNNs remain competitive, highlighting their efficiency in learning molecular representations with limited data.

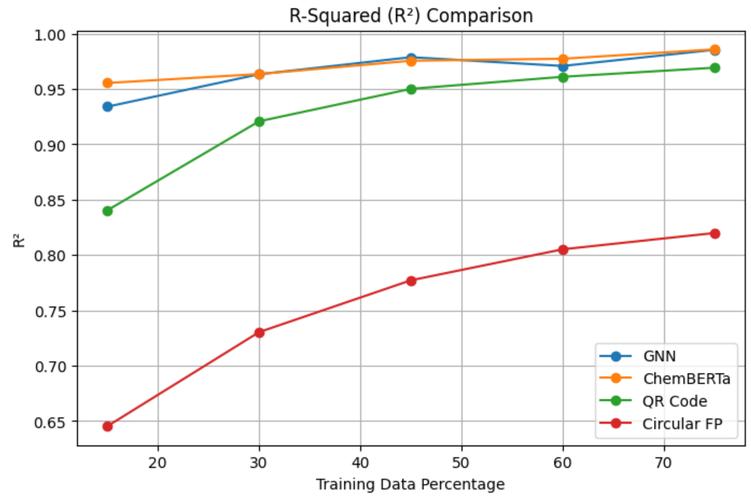

Figure 5. Comparison of R^2 scores across different training data fractions

Conclusion

The architectural variations explored highlight the impact of attention mechanisms and deeper networks in enhancing predictive accuracy. Notably, the 3-layer GNN with 4 attention heads emerged as the most effective configuration.

Despite having fewer parameters than transformer-based models, GNNs effectively learn molecular representations. Their ability to maintain competitive performance even with reduced training fractions underscores their robustness and suitability for molecular modeling tasks.

These findings suggest that GNNs offer a promising alternative to transformer-based models, balancing accuracy and computational efficiency in cheminformatics applications.

References

[1] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural Message Passing for Quantum Chemistry," Proc. ICML, 2017.

[2] M. Simonovsky and N. Komodakis, "Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs," Proc. CVPR, 2017.

[3] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," ICLR, 2018.

[4] T. Cai, S. Luo, K. Xu, D. He, T.-Y. Liu, and L. Wang, "GraphNorm: A Principled Approach to Accelerating Graph Neural Network Training," ICML, 2021.