Lipophilicity Prediction Using Graph Neural Networks

Sravan K Suresh

January 15, 2025

Contents

1 Introduction **2**

2 Dataset **2**
2.1 Overview 2
2.2 Sample Graph 2

3 Model Architecture **3**

4 Training Configuration **4**

5 Results **4**
5.1 Training Metrics 4
5.2 Testing Metrics 4

6 Conclusion **4**

7 References **4**

1 Introduction

Lipophilicity is a critical molecular property influencing a drug's ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profile. This project leverages Graph Neural Networks (GNNs) to predict lipophilicity by encoding molecules as graphs where atoms are nodes and bonds are edges.

2 Dataset

2.1 Overview

The dataset used for this project is `MoleculeNet` from `torch_geometric`. It contains molecular graphs with features representing atoms and edges representing bonds.

- **Dataset type:** `torch_geometric.datasets.molecule_net.MoleculeNet`
- **Number of graphs:** 4200
- **Number of features:** 9
- **Target classes:** 553

2.2 Sample Graph

- **Nodes:** 24
- **Edges:** 54
- **Molecular representation:** `Cn1c(CN2CCN(CC2)c3ccc(Cl)cc3)nc4ccccc14`

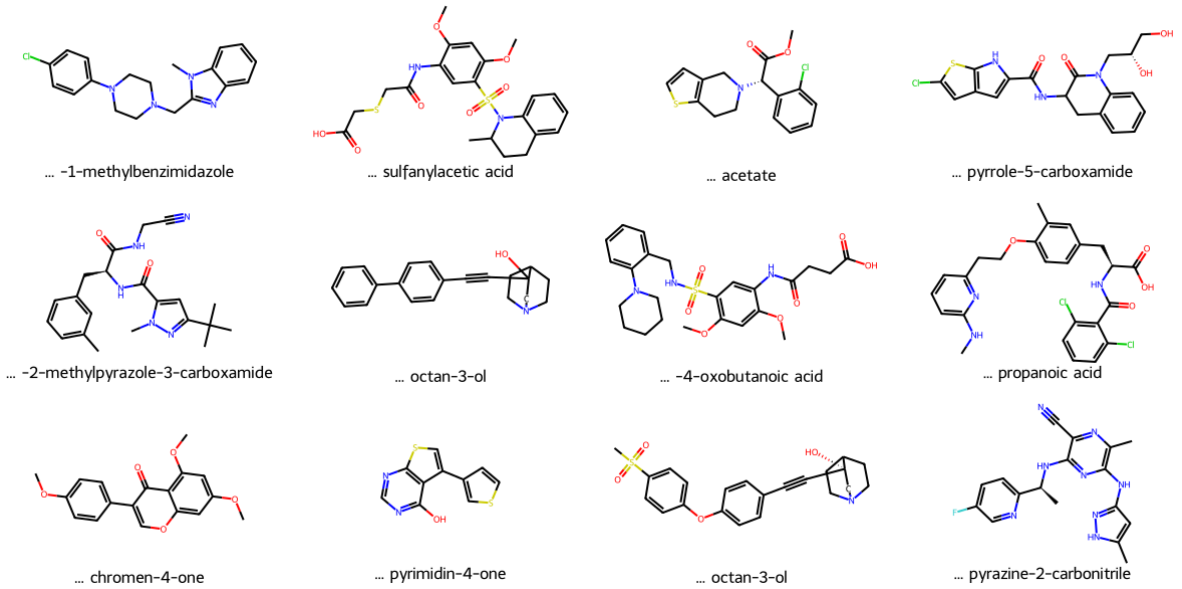

Figure 1: Plot of 12 sample molecules

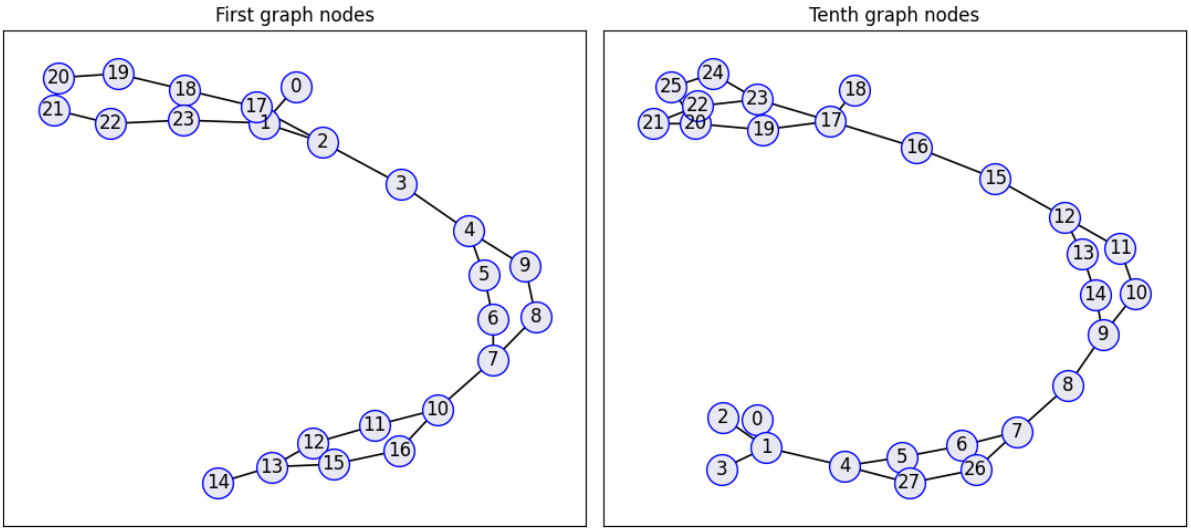

Figure 2: Visualizing two of the Graph nodes using `NetworkX`

3 Model Architecture

The GNN model uses four graph convolution layers, followed by global pooling and a linear regression head.

- **Embedding size:** 64
- **Number of parameters:** 13249
- **Layers:**
  - 1 initial GCNConv layer
  - 3 GCNConv layers
  - 1 linear output layer

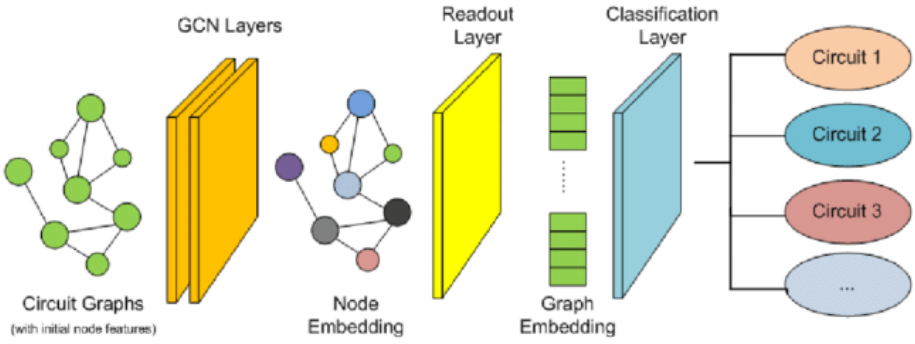

Figure 3: Network architecture of the GNN model

4 Training Configuration

- **Learning rate:** 0.0007
- **Epochs:** 2000
- **Batch size:** 64
- **Training/Test split:** 80/20

5 Results

5.1 Training Metrics

- **Final Train Loss:** 0.02659
- **Final Train R^2 Accuracy:** 98.09%

5.2 Testing Metrics

- **Test R^2 Accuracy:** 50.29%
- **Final Test RMSE:** 0.88335
- **Final Test MAE:** 0.63938

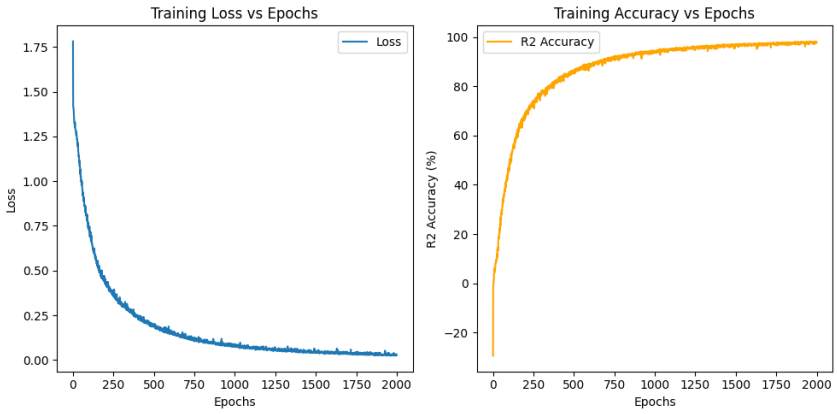

Figure 4: Plots of Training Loss and Accuracy

6 Conclusion

The GNN model demonstrates strong performance on the training set, with an R^2 accuracy of 98.09%. While the test performance is moderate, improvements can be achieved through hyperparameter tuning, dataset augmentation, or advanced GNN architectures such as GraphSAGE or GAT.

7 References

- `torch_geometric` documentation: https://pytorch-geometric.readthedocs.io/
- `MoleculeNet` dataset: https://moleculenet.org/