Predicting Molecular Solubility Using Machine Learning and Graph Neural Networks

Arnav Jayswal

Overview

This project aims to predict the aqueous solubility (log mol/L) of organic molecules from their SMILES representations. We compare traditional machine learning models using engineered molecular fingerprints and physicochemical descriptors with graph neural networks (GNNs) that learn directly from molecular graph structures.

Data Preparation

We used a dataset of SMILES strings paired with experimentally measured solubility values. Data processing was done via RDKit:

- **Fingerprint Features:** Generated 4096-bit Morgan fingerprints capturing substructure information.
- **Physicochemical Descriptors:** Calculated molecular weight, LogP, number of hydrogen bond donors/acceptors, TPSA, rotatable bonds, and aromatic rings.
- **Graph Features:** Parsed SMILES into molecular graphs, extracting atom and bond features, including ring counts, atom type entropy, hybridization states, formal charges, and aromaticity.

These features enabled us to build *basic* and *enhanced* representations for traditional ML models, and graph-structured input for GNNs.

Models

Traditional Machine Learning

We trained Random Forest and XGBoost regressors under two setups:

- *Basic:* Using only Morgan fingerprints.
- *Enhanced:* Combining fingerprints with physicochemical and graph-derived descriptors.

Hyperparameters were tuned via 5-fold cross-validation with `GridSearchCV`. Evaluation used RMSE, R^2, and MAE metrics on a held-out test set.

Graph Neural Network

We implemented a GNN with the following architecture:

- Three stacked Graph Attention (GAT) convolutional layers with batch normalization and ReLU activations.
- Global add pooling to aggregate node features into a graph-level representation.
- Fully connected layers with dropout to predict solubility.

The model was trained using the AdamW optimizer with MSE loss, employing learning rate scheduling and early stopping to prevent overfitting.

Results

Model	RMSE	R^2	MAE
Random Forest (Basic)	1.15	0.72	0.86
XGBoost (Basic)	1.15	0.72	0.86
Random Forest (Enhanced)	0.82	0.86	0.57
XGBoost (Enhanced)	0.77	0.87	0.55
GNN (Graph-based)	**0.69**	**0.90**	—

Table 1: Performance comparison on test data. GNN outperforms traditional models, especially when enhanced features are used.

Key Insights

- Inclusion of physicochemical and graph-based descriptors substantially improves traditional ML models.
- GNNs effectively capture molecular structure directly from graph data, leading to superior predictive accuracy.
- Attention mechanisms in GNN layers allow the model to weigh atomic interactions contextually.
- Proper training strategies such as learning rate scheduling and dropout were critical for the GNN's performance.