Datasets and Challenges in GNN-Driven Drug Discovery

Anonymous

Abstract

This review surveys the data resources, benchmarking practices, and key experimental challenges facing GNN modeling for drug discovery.

1 Key Datasets for GNN Evaluation

- **MoleculeNet:** Aggregates 17 datasets for classification and regression.
- **QM9:** 134,000 molecules with quantum properties.
- **ChEMBL:** Bioactivity data for more than 2 million compounds.

Dataset	# Molecules	Primary Task
MoleculeNet	50,000+	Classification
QM9	134,000	Regression (atomic)
ChEMBL	2,000,000+	Bioactivity

Table 1: Key datasets in molecular property prediction

2 Benchmarking Challenges

- **Data Quality:** Inconsistent labels, irrelevant descriptors.
- **Class Imbalance:** Scarcity of actives in screening sets.
- **Dataset Splits:** Scaffold-based for generalization.

3 Experimental Protocols

- Model comparison : standardized training and reporting.
- Metrics : AUROC, MAE/RMSE tailored to task.
- Reproducibility : code and data sharing.

4 Future Directions

- Larger, richer datasets.
- Benchmarks for interpretability.
- Scaling to ultra-large chemical spaces.

5 References

Comprehensive list available upon request.