COM S 573: Project Proposal

1 Project Title

Graph Neural Networks for Molecular Property Prediction in Drug Discovery

2 Team Members

- Rahman Abdul Rafi
- Samarth Begari
- Toral Chauhan

3 Project Details

3.1 Project Objective

The objective of this project is to build, train, and evaluate Graph Neural Network (GNN) models to predict key molecular properties relevant to early-stage drug discovery, such as toxicity, bioactivity, and ADME-related properties. By developing accurate predictive models, we aim to enable faster triage of candidate compounds and improve the efficiency of the drug discovery pipeline.

The core problem addressed is the limitation of traditional QSAR (Quantitative Structure–Activity Relationship) models, which rely heavily on handcrafted descriptors and often fail to generalize across diverse regions of chemical space. To overcome this, we propose an end-to-end learning approach that directly processes raw molecular graphs, with two primary tasks: (a) classification of toxicity and bioactivity in a binary or multi-label setting, and (b) regression of physicochemical properties such as solubility and lipophilicity.

This problem is important because experimental wet-lab assays are both expensive and time-consuming, creating a bottleneck in drug discovery. Accurate *in-silico* prediction methods can substantially reduce costs and accelerate timelines by prioritizing promising compounds for testing, while also enhancing patient safety through early identification of toxic liabilities.

Machine learning, and GNNs in particular, are well-suited to this challenge. Molecules are naturally represented as graphs, with atoms as nodes and bonds as edges. GNNs exploit this structure through message passing to learn rich and expressive molecular representations, offering clear advantages over fixed fingerprints and traditional descriptor-based models. This makes GNNs a powerful tool for advancing predictive modeling in drug discovery.

3.2 Datasets

For this project, we propose to use publicly available molecular property datasets from MoleculeNet, a widely adopted benchmark suite for machine learning in chemistry and drug discovery.
These datasets are curated from *experimental bioassay results, physical chemistry databases, and toxicology reports*, and are accessible through **DeepChem** and **PyTorch Geometric** dataset loaders, ensuring reproducibility and consistency with prior research.

We plan to select two classification datasets and one regression dataset to capture different aspects of molecular property prediction:

- **Tox21 (Toxicology):** A multi-label classification dataset where chemical compounds are tested across 12 toxicity-related tasks. Data originates from the U.S. Toxicology in the 21st Century initiative, with compounds assayed for effects such as nuclear receptor signaling and stress responses.
- **BBBP (Blood-Brain Barrier Penetration):** A binary classification dataset that identifies whether compounds can cross the blood-brain barrier, a critical property for central nervous system (CNS) drugs. Data is collected from PubChem BioAssay records.
- **ESOL (Solubility Dataset):** A regression dataset containing experimentally measured water solubility values (logS) for small organic molecules, curated from physical chemistry literature.

All three datasets are curated from experimental assays and published literature, ensuring their biological and chemical relevance.

Features and Labels: Molecules will be represented as graphs derived from SMILES strings.

- **Node (atom) features:** atom type (Z), degree, valence, aromaticity, formal charge, hybridization, chirality.
- **Edge (bond) features:** bond type (single/double/triple/aromatic), conjugation, ring membership, stereo configuration.
- **Labels:**
  - Tox21: 12 binary labels (toxicity tasks).
  - BBBP: 1 binary label (permeable vs. non-permeable).
  - ESOL: 1 continuous value (logS).

Data Splits: We will use the official MoleculeNet splits when available, and apply *scaffold-based splitting* otherwise. Scaffold splitting ensures that molecules in the test set belong to chemical scaffolds not present in the training set, which evaluates the model's ability to generalize to novel chemotypes—an essential requirement for drug discovery.
The datasets are of moderate size, making them computationally feasible for a semester project while still providing sufficient diversity for benchmarking. A summary of the datasets is shown in Table 1.

Table 1: Summary of Proposed selected datasets and splits.

Dataset	**Task Type**	**Total Compounds**	**Train / Val / Test**
Tox21	Multi-label classification	∼7,800	6,200 / 800 / 800
BBBP	Binary classification	∼2,050	1,640 / 205 / 205
ESOL	Regression (solubility)	1,128	904 / 112 / 112

3.3 Machine Learning Algorithm

For this project, we will employ Graph Neural Networks (GNNs) as the primary machine learning approach. GNNs are particularly well-suited for molecular property prediction because molecules are naturally represented as graphs, with atoms as nodes and chemical bonds as edges. Through message passing mechanisms, GNNs propagate information across the molecular graph to learn expressive and context-dependent representations of atoms and substructures.

We will evaluate several widely used GNN architectures:

- **Graph Convolutional Networks (GCN)** – a baseline model that aggregates neighbor features in a convolutional manner.
- **Graph Isomorphism Networks (GIN)** – a highly expressive variant designed to capture complex molecular structures.
- **Graph Attention Networks (GAT)** – which introduce attention mechanisms to assign different importance to neighboring atoms/bonds.

3.4 Expected Outcomes

The expected outcome of this project is a set of GNN models capable of accurately predicting molecular properties across multiple benchmark datasets (Tox21, BBBP, and ESOL). Specifically, we anticipate:

- **Improved predictive performance** compared to baseline models that rely on fixed fingerprints, measured by ROC-AUC for classification tasks and RMSE/MAE for regression tasks.
- **Demonstration of generalization** through scaffold-based splits, showing that GNNs can predict properties of molecules with previously unseen chemical backbones.
- **Interpretability insights**, where we identify atoms, bonds, or substructures that contribute most strongly to predictions, providing meaningful connections to known chemical and biological knowledge.
- **Reproducible pipeline**, including data preprocessing, model training, and evaluation, which can serve as a foundation for future work in molecular machine learning.

Overall, the project aims to validate the effectiveness of GNNs for molecular property prediction and demonstrate their potential to accelerate early-stage drug discovery by reducing experimental costs and enabling more informed compound selection.

References

1. Fang, Z., Zhang, X., Zhao, A., Li, X., Chen, H., & Li, J. (2025). Recent developments in GNNs for drug discovery. *arXiv preprint arXiv:2506.01302*. Available at: https://arxiv.org/abs/2506.01302 :contentReference[oaicite:0]index=0.

2. Yao, R., Shen, Z., Xu, X., Ling, G., Xiang, R., Song, T., Zhai, F., & Zhai, Y. (2024). Knowledge mapping of graph neural networks for drug discovery: A bibliometric and visualized analysis. *Frontiers in Pharmacology*, 15:1393415. doi: https://doi.org/10.3389/fphar.2024.1393415 :contentReference[oaicite:1]index=1.

3. Besharatifard, M., & Vafaee, F. (2024). A review on graph neural networks for predicting synergistic drug combinations. *Artificial Intelligence Review*, 57, 49. doi: https://doi.org/10.1007/s10462-023-10669-z :contentReference[oaicite:2]index=2.

4. Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2), 513–530. doi: https://doi.org/10.1039/C7SC02664A.

5. Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Central Science*, 3(4), 283–293. doi: https://doi.org/10.1021/acscentsci.6b00367.

6. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4), 688–702.e13. doi: https://doi.org/10.1016/j.cell.2020.01.021.

7. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., & Song, C. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1, 57–81. doi: https://doi.org/10.1016/j.aiopen.2021.01.001.