Extended connnectivity fingerprint inspired graph neural network that makes use of interim iteration results - Milestone 2

Albert Dinstl

May 2025

Introduction

The goal for this part of the project was to implement and evaluate several baseline experiments on ECFP4 and ECFP6 [6] fingerprints and compare their performance to frozen and trained GNN embeddings.

Experiments

In these experiments, 2048-dimensional ECFP4 and ECFP6 bit vectors, as well as embeddings from untrained, frozen 2- and 3-layer Graph Isomorphism Networks (GIN) [8] and Neural Graph Fingerprints (NGF) [3], were used as inputs to MLP classifiers or regressors. I also conducted experiments in which the GIN and NGF models were trained end-to-end.

Binary Classification on the BACE dataset

The BACE dataset comprises small-molecule inhibitors targeting the human beta-secretase 1 enzyme (BACE-1) [1]. Each sample includes both the molecular graph and its SMILES string, and is labeled according to whether the compound inhibits BACE-1. I trained MLP classifiers on ECFP4 and ECFP6 fingerprints, as well as on frozen, randomly initialized GIN2, GIN3, NGF2, and NGF3 embeddings. Model performance was evaluated using ROC–AUC. I then repeated the experiment with the same architectures trained end-to-end (i.e., allowing GIN2, GIN3, NGF2, and NGF3 to learn during classifier training).

Regression on the Lipohilicity dataset

This task involves predicting the octanol–water distribution coefficient (log D) of each molecule—a key measure of its solubility in nonpolar versus polar solvents. Inputs consist of the molecular graph and SMILES string, and targets are the experimentally measured log D values [9, 5]. I trained MLP regressors on ECFP4 and ECFP6 fingerprints, as well as on frozen, randomly initialized GIN2, GIN3, NGF2, and NGF3 embeddings, evaluating performance via RMSE. These experiments were then repeated with the corresponding GIN and NGF architectures trained end-to-end.

Implementation

To ensure a fair comparison between ECFP and GNN-based methods, I aligned their atom-level feature sets. I examined the PyTorch Geometric documentation to identify the node features automatically extracted from a SMILES string, which include [2]:

- Atomic number
- Chirality
- Degree
- Formal charge
- Number of attatched hydrogen atoms
- Number of radical electrons
- Hybridization
- Aromaticity
- Ring membership

I then passed these same atom invariants to RDKit when computing ECFP bit vectors, ensuring both methods start from identical atomic descriptors. Since all features are categorical, I one-hot encoded each before using them as GNN inputs. For model evaluation, I performed an 80/20 train–test split and stored the resulting subsets in an in-memory dataset. This guaranteed that every model—whether using ECFP fingerprints or GNN embeddings—was trained and tested on exactly the same data partitions. [7]

Results

Table 1: Performance on BACE classification (ROC_AUC) and Lipophilicity regression (RMSE)

Method	ROC_AUC (BACE)	RMSE (Lipophilicity)
ECFP4	0.8877	0.8524
ECFP6	0.8949	0.8460
Frozen GIN (2 layers)	0.7728	1.0649
Frozen GIN (3 layers)	0.7691	1.0962
Frozen NGF (2 layers)	0.6550	1.1875
Frozen NGF (3 layers)	0.6541	1.2240
Trained GIN (2 layers)	0.7828	0.6033
Trained GIN (3 layers)	0.8099	0.5937
Trained NGF (2 layers)	0.8588	0.6454
Trained NGF (3 layers)	0.8787	0.6340

These results clearly highlight the strength of ECFP as a discrete molecular fingerprint

- Classification (BACE): Only the end-to-end–trained NGF embeddings approach ECFP's ROC–AUC, with NGF3 reaching 0.8787 compared to ECFP6's 0.8949.
- Regression (Lipophilicity): ECFP is competitive but is outperformed by trained GNN embeddings. Trained GIN3 achieves the lowest RMSE of 0.5937.

An interesting observation is the small gain from training GIN embeddings on the classification task: frozen GIN2 and GIN3 achieve ROC–AUCs of 0.7728 and 0.7691, while their trained counterparts only improve to 0.7828 and 0.8099. By contrast, NGF networks show a large gap between frozen and trained performance: frozen NGF2/3 embeddings perform at chance-like levels (ROC–AUC =0.655), whereas training improves them to 0.8588 and 0.8787.
This difference is likely due to how NGF constructs its fingerprint with a softmax-based pooling step. In NGF, each atom's contribution to the global fingerprint is weighted by a learnable softmax over output features [3]. When the network is not trained, these softmax weights are effectively random, causing a nearly uniform aggregation of node features. Training refines these softmax distributions, enabling the model to focus on chemically relevant substructures and produce discriminative fingerprints.
Finally, while trained GIN embeddings show only small improvements over their

frozen analogs in classification, they show large gains in regression. This suggests that even untrained GINs capture enough structure to linearly separate labels, but predicting a continuous property requires fine-tuned representations that emerge from training.

Conclusions

When designing a GNN architecture that can approximate ECFP's performance in a frozen state, these results underscore the value of pooling operations that embed inductive biases rather than relying solely on learned parameters. Moreover, any GNN aiming to match ECFP should incorporate multi-scale aggregation—such as Jumping Knowledge [4] or similar mechanisms to integrate information from intermediate layers.

References

[1] URL: `https://drugdesigndata.org/about/grand-challenge-4/bace`.

[2] URL: `https://pytorch-geometric.readthedocs.io/en/latest/_modules/torch_geometric/utils/smiles.html#from_smiles`.

[3] Duvenaud et. al. "Convolutional Networks on Graphs for Learning Molecular Fingerprints". In: *NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2, 07 December 2015* (2015).

[4] Xu et. al. "Representation Learning on Graphs with Jumping Knowledge Networks". In: *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80* (2018).

[5] S. Amézqueta et al. "Chapter 6 - Octanol-Water Partition Constant". In: *Liquid-Phase Extraction*. Ed. by Colin F. Poole. Handbooks in Separation Science. Elsevier, 2020, pp. 183–208. ISBN: 978-0-12-816911-7. DOI: `https://doi.org/10.1016/B978-0-12-816911-7.00006-2`. URL: `https://www.sciencedirect.com/science/article/pii/B9780128169117000062`.

[6] Mathew Hahn David Rogers. "Extended Connectivity Fingerprints". In: *J. Chem. Inf. Model. 2010, 50, 742–754* (2010).

[7] Albert Dinstl. URL: `https://github.com/albertd01/bachelor_thesis`.

[8] Jure Leskovec Keyulu Xu Weihua Hu and Stefanie Jegelka. "How powerful are Graph Neural Networks". In: *Conference paper at ICLR 2019* (2019).

[9] Jian-Bing Wang et al. "*In silico* evaluation of logD$_{7.4}$ and comparison with other prediction methods". In: *Journal of Chemometrics* 29.7 (2015), pp. 389–398.