Background

In molecular property prediction, it is crucial to develop effective molecular representations—commonly known as molecular fingerprints—to serve as inputs for machine learning models. Two dominant paradigms have emerged in this area: Extended Connectivity Fingerprints (ECFP) and Graph Neural Networks (GNN).

Extended Connectivity Fingerprint (ECFP):
ECFP generates fixed-size binary vectors by iteratively aggregating local neighborhood information of atoms. In each iteration, the algorithm combines an atom's features with those of its neighbors, hashing the result to create discrete structural identifiers that progressively capture larger substructures. These identifiers set corresponding bits in the fingerprint, yielding a representation that encodes the presence or absence of specific substructures.[1]

Graph neural network (GNN):
GNNs learn molecular representations in an end-to-end differentiable manner. They propagate and transform node features through a series of message-passing layers and then aggregate the resulting information with a global pooling operation at the readout layer.

The key difference between these approaches is that ECFP relies on fixed, rule-based, and non-differentiable operations like hashing and bit setting, whereas GNNs leverage continuous, differentiable functions that allow the model to optimize the feature extraction process during training.

Motivation

Research interest and questions

This thesis explores how to integrate the discrete and iterative approach of ECFP into the design of Graph Neural Networks aiming at producing high quality fingerprints for downstream tasks like molecular property prediction. This research seeks to answer the following questions:

- How can discrete iterative methods like ECFP be integrated into the continuous and differentiable nature of Graph Neural Networks?
- What benefits arise from leveraging the aggregation of intermediate representations in Graph Neural networks?

Goal of the Project

The primary goal of this project is to investigate how the concept of aggregating interim iteration results—as employed in ECFP—can be integrated into the design of Graph Neural Networks.

Specifically, the project aims to develop a GNN architecture that generates molecular embeddings by explicitly incorporating and aggregating intermediate node-level embeddings and combining them with concatenation.

Approach

The initial phase of the project is to randomly initialize a frozen Graph Isomorphism Network (GIN) [1] and use it as a function to obtain graph embeddings of molecular graphs. These molecular embeddings will be evaluated in a downstream task against traditional ECFP bit vectors. For both molecular representations an MLP will be trained and its ROC-AUC will be compared to one another. As another baseline a frozen Neural Graph Fingerprint (NGF) [4] model will be used to also train an MLP classifier.

In the second phase of the project, the goal is to design an enhanced GIN that leverages mechanisms like Jumping Knowledge concatenation [3] to incorporate interim iteration results in the final graph embedding of a molecule. It will be assessed how close such a frozen enhanced GIN's embeddings come to ECFP when training an MLP on a downstream task with these embeddings.

Single-Layer GNN for Node-Level Embeddings (ECFP First Iteration Anal

The initial phase of the project is to design and randomly initialize a single-layer GNN that produces graph embeddings similar to the output of the first iteration of ECFP. In this stage, each atom (node) is represented by a common set of features, and its embedding is updated by aggregating information from its immediate neighbors—mirroring the way ECFP updates an atom’s identifier with local structural information. The goal for this part of the project will be to design a network that can produce the same number of uniquely distinguishable embeddings as ECFP in its first iteration. The quality of these embeddings will be asessed in a downstream task and their accuracy will be compared to ECFP2 bit vectors.

Multi-Layer GNN for Molecular Fingerprinting Using Interim Iteration Results

In the second phase, the approach will be extended to a multi-layer GNN that explicitly captures and aggregates the node embeddings generated at each intermediate layer. This design will allow for the dissection of contributions from each iteration, ultimately producing a global molecular fingerprint. The final molecular-level embedding will be evaluated on downstream tasks, such as molecular property prediction, and its performance will be compared against traditional ECFP bit vectors. This comparative analysis aims to assess whether incorporating interim iteration results can enhance the quality and interpretability of the learned molecular representations. [2,3]

Challenges to consider

- A differentiable neighbourhood aggregation method
  In ECFP, the local neighborhood of each atom is aggregated by hashing the atom's feature array along with its neighbors' features into a new substructure identifier. This process, while effective for generating fixed binary fingerprints, is inherently non-differentiable. For a GNN to support end-to-end learning via gradient descent, the neighborhood aggregation must be implemented using differentiable operations. [4]
- Duplicate removal:
  In traditional ECFP fingerprint generation, if the same identifier is produced for different atoms or substructures, the fingerprint simply marks the corresponding bit as active, effectively collapsing duplicates into a single indicator. This non-differentiable process is acceptable for fixed, binary fingerprints. However, in a GNN, every operation must be differentiable to allow for gradient-based learning. Therefore, the GNN must employ differentiable aggregation methods that can handle duplicate or highly similar intermediate representations without disrupting the learning process.
- Transforming graph embeddings into fixed size fingerprints
  When using concatenation for neighborhood aggregation, the resulting graph-level embedding may vary in length from one molecule to another, as it depends on the number of nodes (atoms). However, downstream tasks such as classification require input vectors of a consistent dimensionality. To address this, a differentiable operation must be applied to transform the variable-length, concatenated embeddings into fixed-size neural fingerprints.

Milestones

Literature

[1] David Rogers, Mathew Hahn, Extended Connectivity Fingerprints, J. Chem. Inf. Model. 2010, 50, 742–754
[2] Keyulu Xu, Weihua Hu, Jure Leskovec and Stefanie Jegelka, How powerful are Graph Neural Networks? Published as a conference paper at ICLR 2019
[3] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, Stefanie Jegelka, Representation Learning on Graphs with Jumping Knowledge Networks
[4] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, Ryan P. Adams, Convolutional Networks on Graphs for Learning Molecular Fingerprints, NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2, 07 December 2015