

Molecular Property Prediction Using Graph Neural Networks

The primary objective of this project was to predict the toxicity of new molecules, a crucial task in drug discovery. The project leveraged the unique structure of molecules, representing them as 2D graphs where nodes are atoms connected by edges symbolizing chemical bonds. The challenge was to develop a model that predicts molecular properties more accurately than a provided baseline, focusing on graph-level prediction using graph neural networks (GNNs).

1) Choosing Graph Attention Network (GAT)

Initially, I explored the Graph Attention Network (GAT). The decision to use GAT was driven by its ability to assign varying levels of importance to nodes. I hypothesized that this attention mechanism would capture the nuances in molecular structures more effectively than traditional GNNs, as different atoms (nodes) and their bonds (edges) play varying roles in determining a molecule's properties.

2) Experimenting with Graph Isomorphism Network (GIN)

Next, I turned to the Graph Isomorphism Network (GIN). This choice stemmed from GIN's unique approach to understanding different graph structures. GIN's architecture is designed to distinguish between different molecular structures, potentially offering better performance in identifying unique molecular features that contribute to toxicity.

3) Developing a Combined GCN and GIN Model

After experimenting with GAT and GIN, I considered combining their strengths. The rationale was that a hybrid model might leverage GCN's effectiveness in capturing global graph structure and GIN's prowess in understanding complex node relationships. The combined model was an attempt to marry these capabilities, aiming for a more robust prediction tool.

4) Implementing GraphSAGE

GraphSAGE stood out as another promising architecture, primarily due to its neighborhood sampling and aggregation strategy. This approach allows the model to learn from local graph structures, potentially capturing essential local interactions among atoms. In molecular graphs, local structural information can be crucial, making GraphSAGE an attractive option.

5) Hyperparameter Tuning for GCN

Despite the various models explored, none significantly outperformed the GCN baseline in terms of ROC AUC scores. This led me to hyperparameter tuning for the GCN model. The process involved adjusting hidden channels, learning rates, and dropout rates to find the most effective combination. The goal was to fine-tune the GCN model to extract the best possible performance, as sometimes, a well-tuned simple model can outperform more complex architectures.

Conclusion

The experimentation journey through different GNN architectures was both challenging and enlightening. Each model offered unique perspectives and tools to tackle the problem, but

ultimately, refining and optimizing the GCN model proved most effective. This experience highlighted the importance of not only model selection but also the fine-tuning of models in achieving optimal results. It was a reminder that in machine learning, especially in a field as nuanced as molecular property prediction, there are no one-size-fits-all solutions, and success often lies in experimentation and optimization.