KIMIKA: MOLECULE GENERATOR WITH GNNs

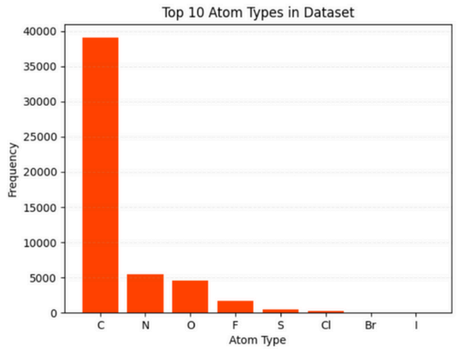
Baručija Emir
Kopić Džana
Sarajlić Amir
Tafro Kanita

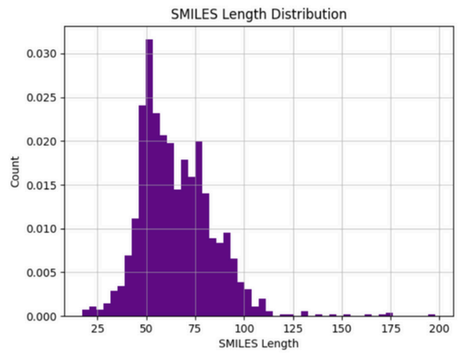
github.com/emirr44/KIMIKA-Molecular-Classification-and-Generation

INTRODUCTION

This repository provides the source code for fine tuning pretrained Graph Neural Network (GraphMVP) for molecular classification and utilizes Variational Autoencoder to generate novel molecules from SMILES strings. The goal is to apply deep learning for drug discovery; we predict β-secretase (BACE1), inhibitor activity (classification), explore new compound structures (generation), and predict HIV activity on SMILES strings (classification). The classification task helps identify potential drug candidates, while the generative model allows de novo design of novel molecules. Screening vast chemical libraries is quite expensive, and AI offers cost-effective way to explore existing and even propose new molecules, learning patterns unseen by humans.

BACE DATASET





The BACE dataset contains 1,513 compounds with binary activity labels (active/inactive). A major challenge is that chemical space is enormous (10^60 possible drug-like molecules), so learning from limited examples is hard.

HIV DATASET

HIV database contains of 41127 SMILES strings wahich were labeled as HIV active (class 1) or HIV inactive (class 0). The dataset is very imbalanced, with 39684 of the instances being in class 0 while only 1443 are labeled 1.

PROBLEM DEFINITION AND MOTIVATION

Molecular Classification - We predict if a molecule inhibits BACE1, a key enzyme in Alzheimer's. Accurate models prioritize lead compounds and cut lab costs. The model was also tested on an HIV database.

Molecular Generation - We use a variational autoencoder (VAE) to map SMILES strings into a continuous latent space, enabling interpolation and optimization for new compounds. VAEs allow gradient-based search and novel molecule generation.

Graph Neural Networks excel at capturing molecular structure by learning atom-bond interactions. GraphMVP uses multi-view self-supervision to integrate 3D knowledge into a 2D GNN, enhancing feature extraction. VAEs outperform traditional methods by learning continuous molecular embeddings.

CONCLUSION

We implemented a pipeline that combines a fine-tuned Graph Neural Network (GraphMVP) with a Variational Autoencoder (VAE) for drug discovery. Our classifier performed strongly on the BACE1 dataset (with a test accuracy of 0.8446) and demonstrated its ability to identify potent inhibitors. The VAE opened up a path toward generating novel compounds by navigating its learned chemical space.

While our models fell short in some aspects , particularly in classifying HIV-active compounds due to dataset imbalance , we're encouraged by these results. This pipeline forms a strong starting point for further exploration and shows the potential for employing deep learning to aid in drug design.

RESULTS

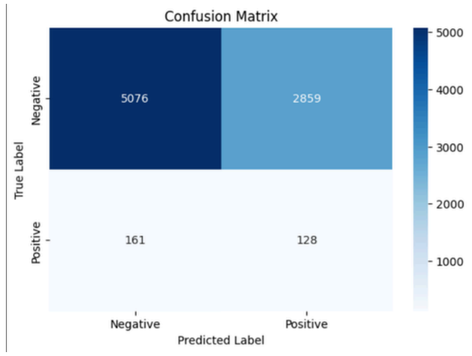
BACE Classification

We trained GraphMVP to predict inhibitor activity against BACE1. With weights, we improved our test accuracy to 0.8446, outperforming the unweighted training (0.7867).

This shows that adding appropriate class weights helps account for data imbalance and lets the classifier learn from underrepresented cases.

Training was smooth and convergence healthy — validation closely follow training, with no major overfitting.

HIV Classification

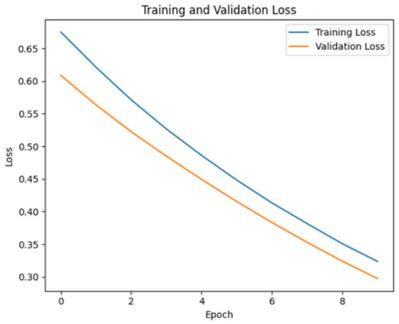


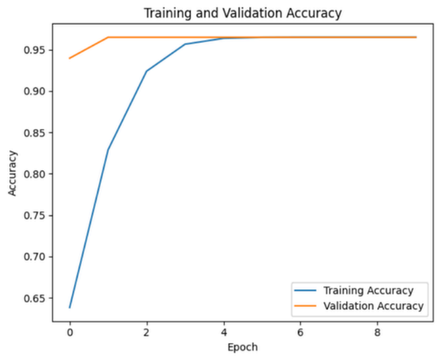
For **Random Forest**, the classifier performed weakly, with an AUC of 0.55 — barely above guessing.

This was due to a heavy class imbalance (289 positives vs 7935 negatives) and poor ability to separate the two groups.

The *confusion matrix* highlights this, with many false negatives.

In contrast, our ***Neural Network*** successfully recognized inactive compounds (with high accuracy), but completely missed the active ones which is a common challenge with imbalanced data.





This signals the need for additional strategies (such as SMOTE, custom-loss functions, or data augmentation) in future work.