Molecular Property Prediction with Deep Learning Report

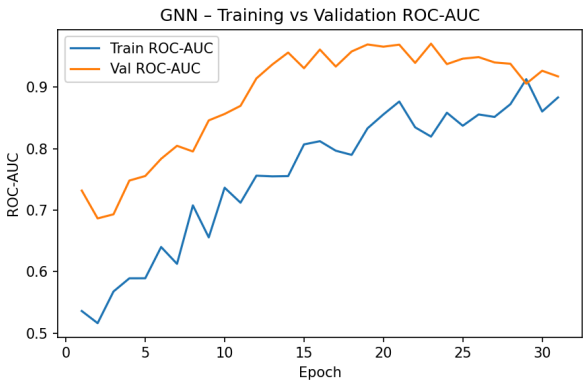
Phineas Wormser

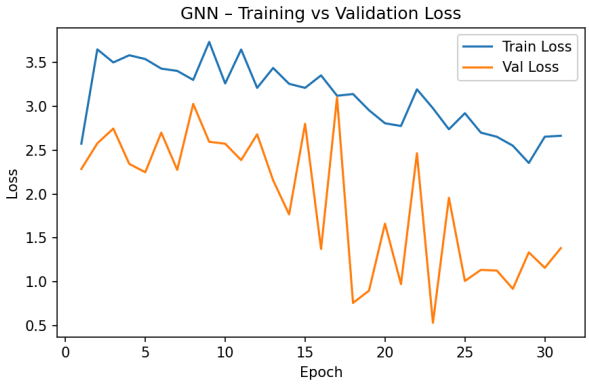
The goal of this project was to predict the antibacterial activity of molecules using machine learning models built on different molecular representations. We explored five main model families. A multilayer perceptron (MLP) trained on Morgan fingerprints, a convolutional neural network (CNN), a recurrent neural network (RNN), and a Transformer trained on SMILES strings, and a graph neural network (GNN) trained on molecular graph representations. The dataset consisted of labeled molecules represented by their SMILES strings, which were featurized using RDKit to produce both 2D fingerprints and graph-based inputs.

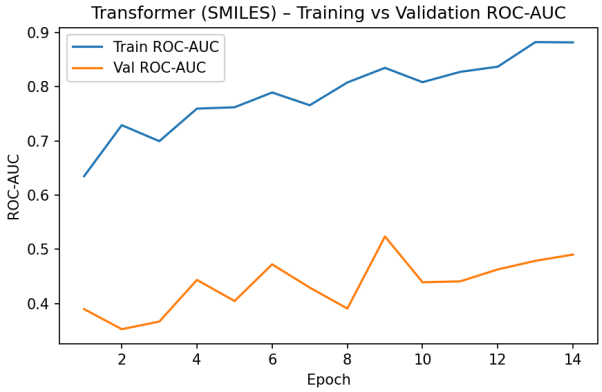
All models were implemented in PyTorch, with PyTorch Geometric used for graph processing. Training employed early stopping, model checkpointing, and binary cross-entropy loss with class weighting to handle dataset imbalance between active and inactive compounds. Evaluation was based primarily on ROC-AUC and Average Precision (AP), as these metrics provide a more reliable measure of model discrimination under class imbalance. The goal was to compare model performance, analyze the impact of architecture and regularization choices, and assess whether more expressive representations (like graphs and sequences) outperform simpler fingerprint-based baselines.

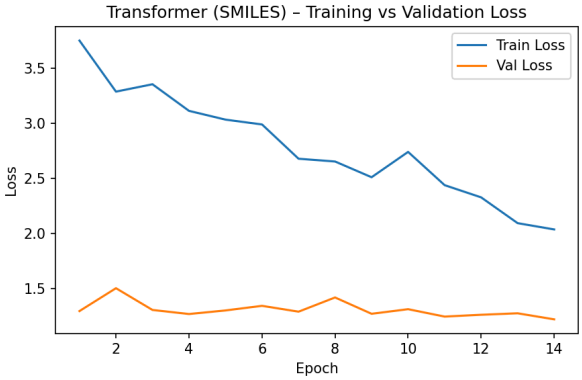
Primary Model Results

Model	ROC-AUC	AP	Accuracy	Precision	Recall
MLP over Morgan	0.622	0.271	0.975	0.250	0.250
CNN over SMILES	0.656	0.233	0.985	1.000	0.125
RNN over SMILES	0.797	0.244	0.985	0.667	0.250
Transformer over SMILES	0.463	0.036	0.983	0.000	0.000
GNN on Molecular Graphs	0.561	0.161	0.947	0.095	0.250









MLP Batch Size

Batch Size	Validation ROC-AUC	Test ROC-AUC	Test AP
8	0.976	0.509	0.222
32	0.997	0.521	0.264
128	0.997	0.639	0.291

MLP Dropout, Weight Decay, and Gradient Clipping Permutations

Dropout	Weight Decay	Gradient Clipping	Validation ROC-AUC	Test ROC-AUC	Test AP
0.0	0	None	0.991	0.499	0.264
0.0	0	1	0.993	0.512	0.263
0.0	0.00001	None	0.996	0.490	0.262
0.0	0.00001	1	0.988	0.522	0.264

0.0	0.0001	None	0.995	0.510	0.222
0.0	0.0001	1	0.960	0.518	0.264
0.1	0	None	0.954	0.528	0.264
0.1	0	1	0.996	0.516	0.263
0.1	0.00001	None	0.964	0.513	0.189
0.1	0.00001	1	0.954	0.538	0.202
0.1	0.0001	None	0.991	0.491	0.262
0.1	0.0001	1	0.988	0.525	0.264
0.3	0	None	0.972	0.542	0.266
0.3	0	1	0.975	0.516	0.263
0.3	0.00001	None	0.998	0.524	0.264
0.3	0.00001	1	0.995	0.489	0.151
0.3	0.0001	None	0.994	0.572	0.271
0.3	0.0001	1	0.974	0.513	0.263
0.5	0	None	0.989	0.596	0.234
0.5	0	1	0.964	0.581	0.268
0.5	0.00001	None	0.991	0.602	0.356
0.5	0.00001	1	0.980	0.559	0.266
0.5	0.0001	None	0.995	0.611	0.316
0.5	0.0001	1	0.995	0.547	0.265

MLP Depth/Width

Hidden Sizes	Validation ROC-AUC	Test ROC-AUC	Test AP
(64,)	0.986	0.517	0.264
(128,)	0.986	0.615	0.220
(256,)	0.996	0.555	0.265

(512, 256)	0.959	0.525	0.263

MLP Learning Rate Schedule

Scheduler	Val ROC-AUC	Test ROC-AUC	Test AP
Constant (none)	0.9888	0.5137	0.2638
ReduceLROnPlateau	0.9785	0.5298	0.2226
Cosine Annealing	0.9776	0.5287	0.2231
Step Decay	0.9785	0.5142	0.2220

MLP Class Imbalance

Configuration	Validation ROC-AUC	Test ROC-AUC	Test AP
posweight	0.9113	0.4920	0.1513
sampler	0.9311	0.5341	0.2019
both	0.9829	0.5067	0.2641

RNN Depth

Layers	ROC-AUC	AP	Precision	Recall
1-layer	0.5719	0.0846	0.000	0.000
2-layer	0.6384	0.0393	0.000	0.000

Transformer Depth

Layers	ROC-AUC	AP	Precision	Recall
1-layer	0.4759	0.1425	0.000	0.000
2-layer	0.4804	0.0217	0.000	0.000
3-layer	0.5668	0.3205	0.273	0.375

GNN Depth

Layers	ROC-AUC	AP	Precision	Recall
1-layer	0.6666	0.0314	0.000	0.000
2-layer	0.7328	0.0478	0.000	0.000
3-layer	0.6558	0.0449	0.000	0.000

CNN Dropout

Dropout	ROC-AUC	AP	Precision	Recall
0.0	0.7103	0.1633	0.030	0.250
0.1	0.7291	0.2328	0.023	0.250
0.3	0.7323	0.0541	0.000	0.000
0.5	0.6349	0.0534	0.000	0.000

CNN L2 Regularization

Weight Decay	ROC-AUC	AP	Precision	Recall
0.0	0.6744	0.1764	0.167	0.125
0.0001	0.7245	0.1838	0.200	0.250
0.001	0.7414	0.1873	0.125	0.125

Transformer Learning Rate Scheduler

Scheduler	ROC-AUC	AP	Precision	Recall
Constant	0.5684	0.2852	0.079	0.375
Step	0.5341	0.0386	0.000	0.000
Cosine	0.4965	0.0499	0.000	0.000

Comparative Analysis

Across the different model families, several hyperparameters had clear and consistent effects on performance. Model depth was one of the most important factors: the 2-layer GNN achieved the best overall ROC-AUC (≈0.73) before performance dropped with a deeper network, suggesting that additional layers began to propagate noise rather than useful signal. Similarly,

the Transformer showed gradual improvement as depth increased from one to three layers, though it still struggled to generalize compared to simpler architectures. Regularization also played a key role, particularly for the CNN, where moderate dropout (0.1–0.3) and L2 weight decay (10^{-3}) improved both stability and ROC-AUC. In contrast, excessive dropout (0.5) sharply reduced precision and recall, indicating underfitting. For the MLP, batch size produced one of the most noticeable differences—larger batches (128) yielded smoother training and higher test ROC-AUC (≈0.64), likely due to more stable gradient estimates.

Certain models clearly benefited more from regularization than others. The CNN over SMILES sequences showed the largest improvement from weight decay and dropout, as these mitigated its tendency to overfit local token patterns. The Transformer and RNN models were less responsive to regularization but benefited from increased model capacity, while the MLP remained stable under a range of hyperparameters, with performance mostly limited by representational power rather than overfitting. Handling class imbalance through weighted sampling and adjusted loss functions improved calibration more than raw ROC-AUC, helping the models produce more meaningful probability outputs. Overall, models trained on molecular fingerprints and SMILES sequences generalized best, while graph-based models (GNNs) underperformed slightly, likely due to limited feature richness and the small dataset size.

Discussion

The results illustrate the tradeoff between model complexity and data scale in molecular property prediction. Deep sequence and graph architectures, which are highly expressive, require large and diverse datasets to reach their potential. In this project, the dataset was relatively small and heavily imbalanced, with far more inactive than active molecules. Under these conditions, simpler architectures like the MLP on Morgan fingerprints and the CNN over SMILES sequences achieved comparable or better generalization than deeper models such as Transformers and GNNs. The more complex models often exhibited high validation AUCs but lower test AUCs and near-zero recall, indicating overfitting to spurious correlations in the training data.

Overfitting was particularly evident in the Transformer and GNN models, where increasing the number of layers did not consistently improve test performance. These architectures can memorize token- or structure-level features that fail to transfer to unseen molecules. The CNN and MLP, on the other hand, benefited from regularization and smaller parameter counts that helped them generalize from limited data. This finding highlights that in low-data, high-imbalance regimes, inductive bias—in the form of handcrafted features like Morgan fingerprints or localized convolutional filters—can outperform more flexible but data-hungry architectures.

Interestingly, traditional 2D molecular representations like fingerprints remained competitive against modern deep learning models. The MLP using Morgan fingerprints performed reliably, achieving moderate AUC and precision-recall values while training quickly and exhibiting minimal instability. While GNNs are theoretically better suited for molecular data, their advantage did not materialize here due to the limited dataset and basic atom/bond featurization.

A 3D EGNN might capture more chemical structure information, but its benefit would likely depend on scaling up the data.

Finally, class imbalance proved to be one of the most significant limitations of this dataset. With fewer than 5% positive samples, models frequently produced high AUCs but poor recall and near-zero precision at standard thresholds. Techniques like weighted loss functions and oversampling helped modestly but did not fully overcome the imbalance. This highlights an important practical lesson: in domains like drug discovery, optimizing for metrics such as precision-recall (AP) and carefully tuning decision thresholds is often more meaningful than raw accuracy or ROC-AUC alone.