Deep Learning Course MSc Molecular Graph Classification

IOANNIS SAVVAS

mtn2319

Project overview

Classification Target

- Classify molecular structures into Readily Biodegradable (RB) and non-Readily Biodegradable (NRB).
- A substance is considered RB if it can be biodegraded more than 60% in a 28-day window period.

Motivation

- Work on a new dataset with no prior modelling
- Create MPNNs combined with Fingerprints (FP-GNN)
- Compare results of classical ML algorithms – MPNNs – FP-GNN

Dataset

Size and classes

- Dataset 1 (https://zenodo.org/records/3540701): Contains 3192 molecules (2059 NRB - 1133 RB). Has already been used for modelling.
- Dataset 2 (https://zenodo.org/records/8255910): Contains 3703 molecules (1789 NRB - 1918 RB)
- Final Dataset total imbalance ratio --> 55 – 45

Split

- Split into Train – Validation – Test Stratified (Each dataset keeps the same class balance)
- Split ratio: 80 – 10 - 10
- Test set is completely hidden during training-validation procedure

Featurization

Vector Features for Baseline Models and FP-GNN

- MACCS Fingerprints: Fixed-length bit (167). Each bit represents presence or absence of specific predefined substructure or molecular feature.
- Example: Presence of specific atoms (O, N), Structural groups (Carbonyl groups, esters), etc...

Node (Atom) features – Signal

68 Total Features per atom

Atomic Feature	Vector dimension
Symbol	43 (One-Hot)
Adjacent Hydrogens	5 (One-Hot)
Degree	7 (One-Hot)
Formal Charge	1 (Integer)
Radical Electrons	5 (One – Hot)
Hybridization	6 (One – Hot)
Aromaticity	1 (Binary)

MPNN architecture (GAT)

- Due to high number of features per node, Graph Attention Network (GAT: arXiv:1710.10903) was implemented to focus on the most relevant features of neighboring nodes though self-attention.

Message Passing Layer

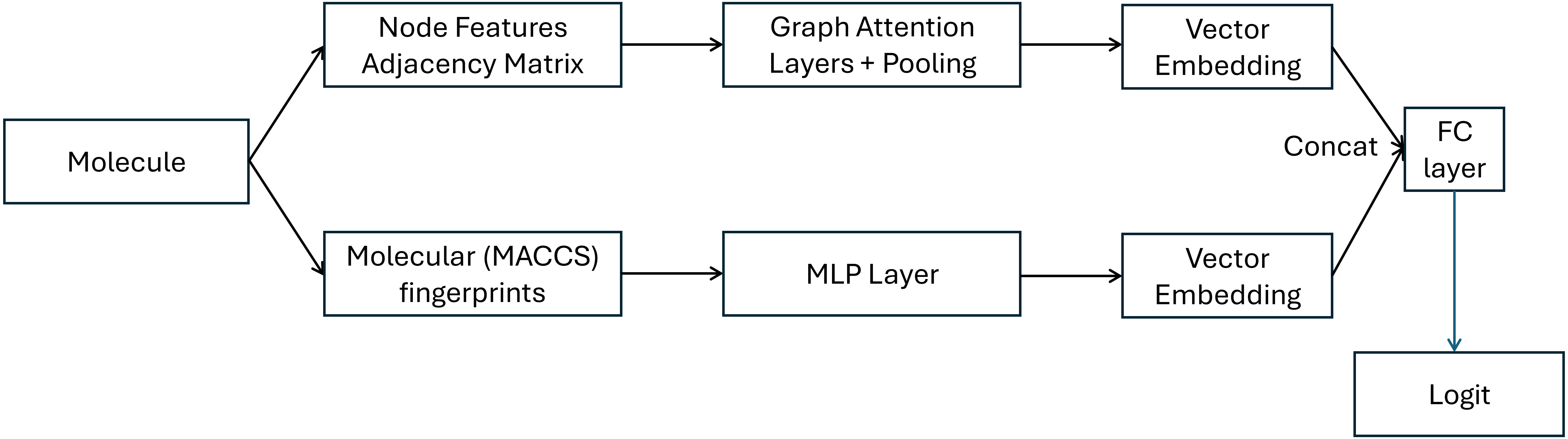
$$h_i^{(l+1)} = \varphi(h_i^{(l)} \bigoplus_{j \in N_i} \alpha(h_i^{(l)}, h_j^{(l)}) \psi(h_j^{(l)}))$$

Graph Pooling (Mean, Permutation Invariant)

$$r_i = \frac{1}{N_i} \sum_{n=1}^{N_i} x_n$$

Fully Connected Layer (MLP)

FP-GNN architecture



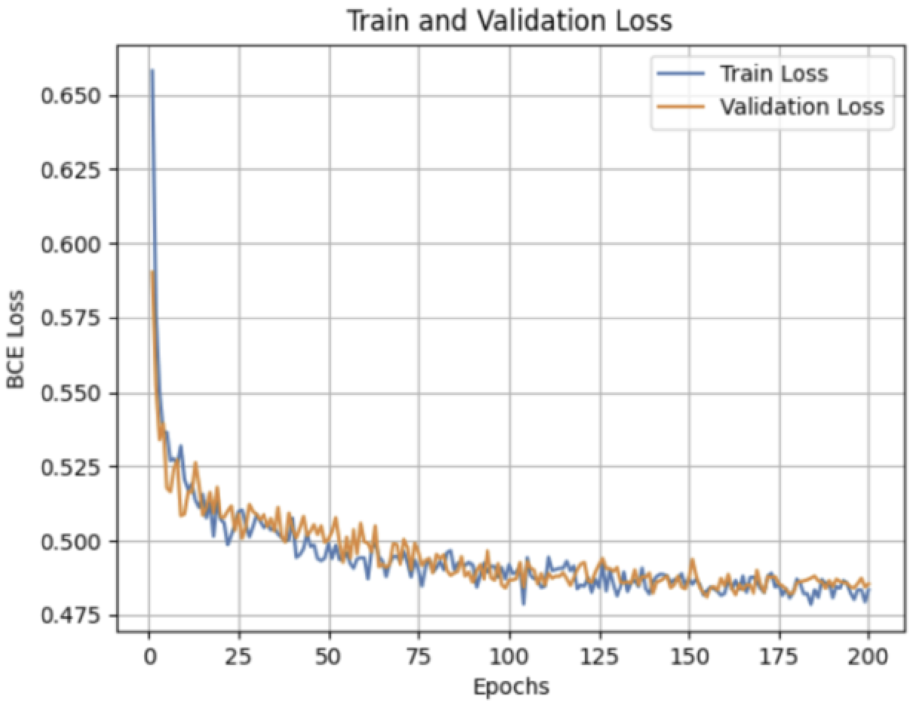
Training and Hyperparameter details

- Models trained on Google Collab T4 GPU
- Loss function : Binary Cross entropy Loss
- Optimizer : AdamW
- Best model selected based on validation loss
- Greedy Search of Hyperparameters

Hyperparameter Name	Range
Batch Size	{32, 64, 128}
Learning Rate	{0.0001, 0.0005, 0.001}
Hidden units	{16, 32, 64}
Attention Heads	{2, 3, 4, 5, 6}
MPNN layers	{2, 3, 4}
Activation function	{ReLU, ELU}
Dropout probability	{0.1, 0.2, 0.3, 0.4}

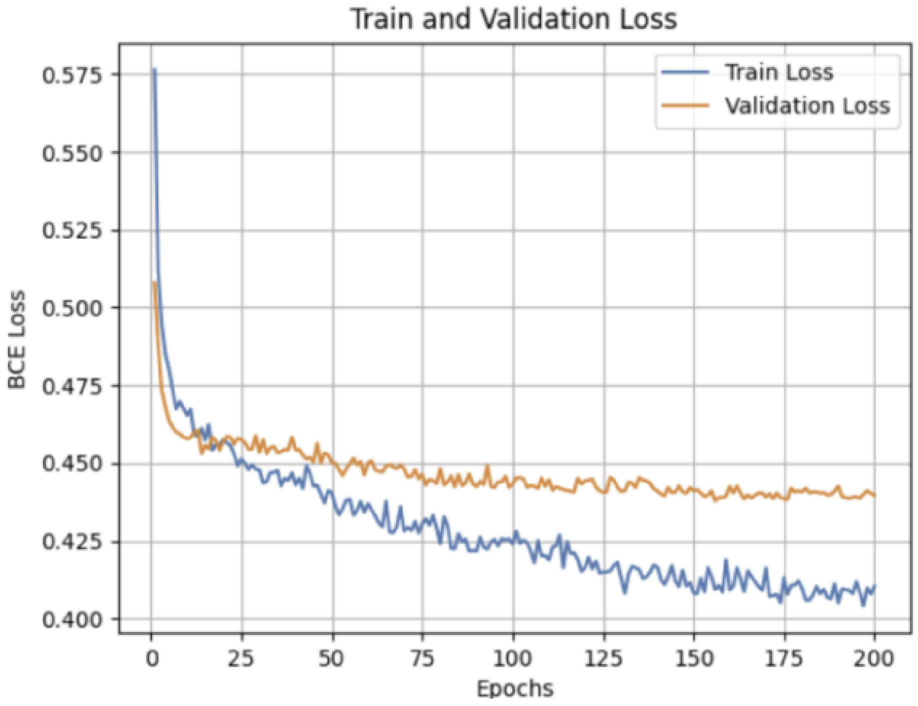
Learning curves

GAT



Best validation loss = 0.481

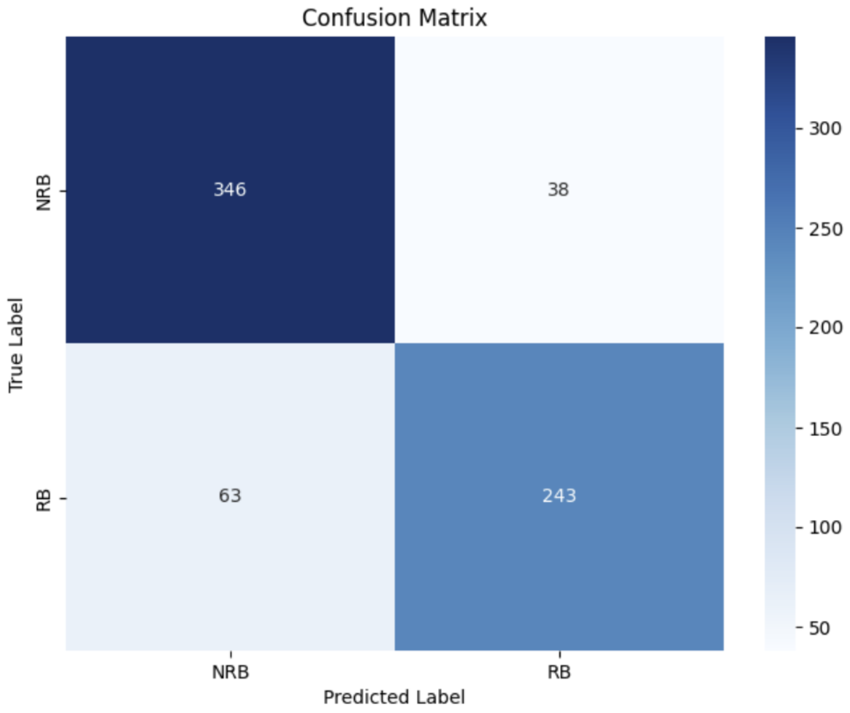
FP-GNN



Best validation loss = 0.439

Test Results and Comparison

FP-GNN Confusion Matrix



Model	Balanced Accuracy
LR	0.809
SVM	0.828
GAT	0.806
FP-GNN	**0.848**

- Better accuracy than SVM in both classes
- 11 more correct predictions in the positive class

Conclusion and Future Work

- Combining GNNs with Molecular Fingerprints maybe can help detect substructure information that MPNNs cannot.
- More molecular fingerprints can be added (ECFP, PubCHEM, Topological) that include more structural motifs.
- Try with more SoTA GNN architectures
- Include-Encode Bond features

Thank you

Any questions?