Assistant in Computational Biology

Postdoctoral Research

Roozbeh H. Pazuki (He/Him)

January 3, 2025

GNNs Embeddings

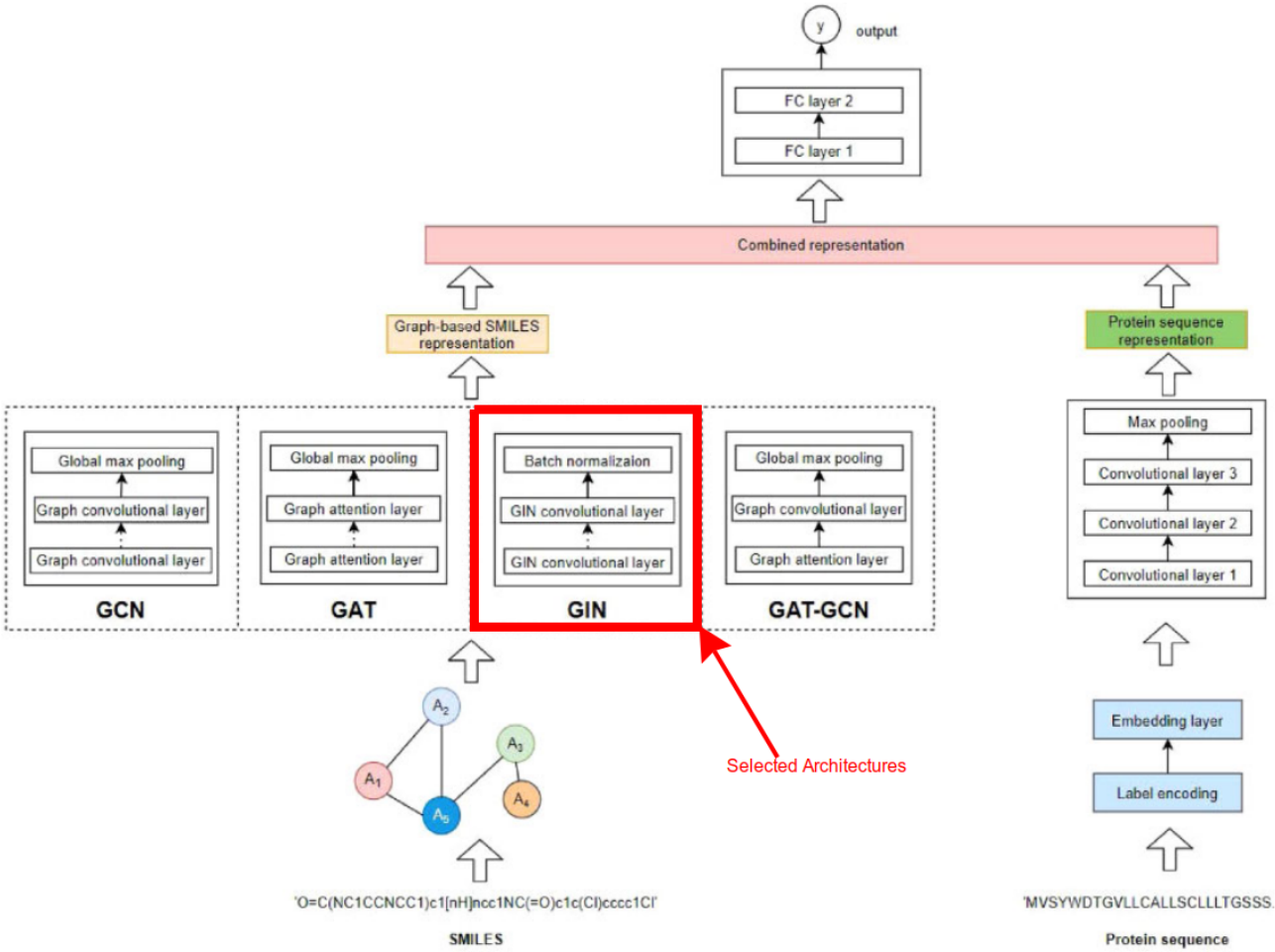
Details

- **Method Name:** GraphDTA.
- **Publication Year:** 2021.
- **Publication Title:** "GraphDTA: predicting drug–target binding affinity with graph neural networks" [1].
- **Architectures:** Graph Isomorphism Convolution Network.
- **File Name:** "drug_GraphDTA_GINConvNet_label.csv".

GNNs Embeddings

Original Model

Embedding representations are extracted using the graph isomorphism network (GIN) architecture from [1]. Here, the drug's SMILES alongside 1D convolutional mapping of protein sequences are embedded, and downstream, the drug-target affinity's value is estimated. We trained the model by using the paper's supplementary material.

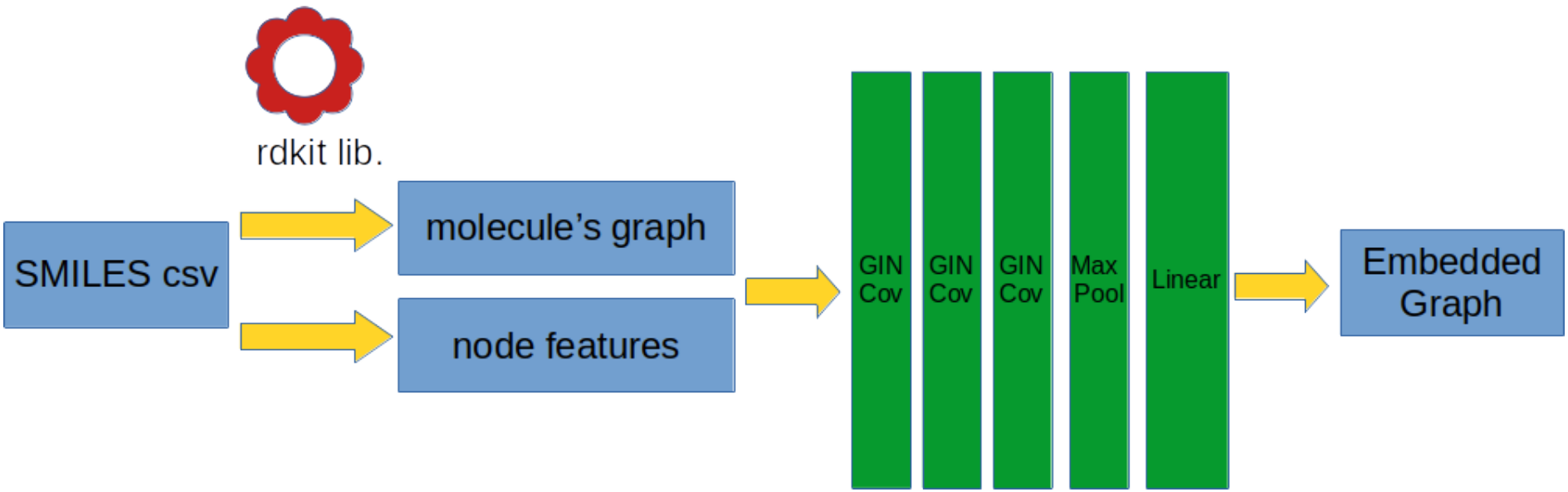


GNNs Embeddings

Prepration

Using the "rdkit library", a custom code sanitises and converts SMILES to their graph representations. For each node's atom, the node features were also extracted. Then, by feeding the augmented partial network from [1], we only use the embedded graphs output.

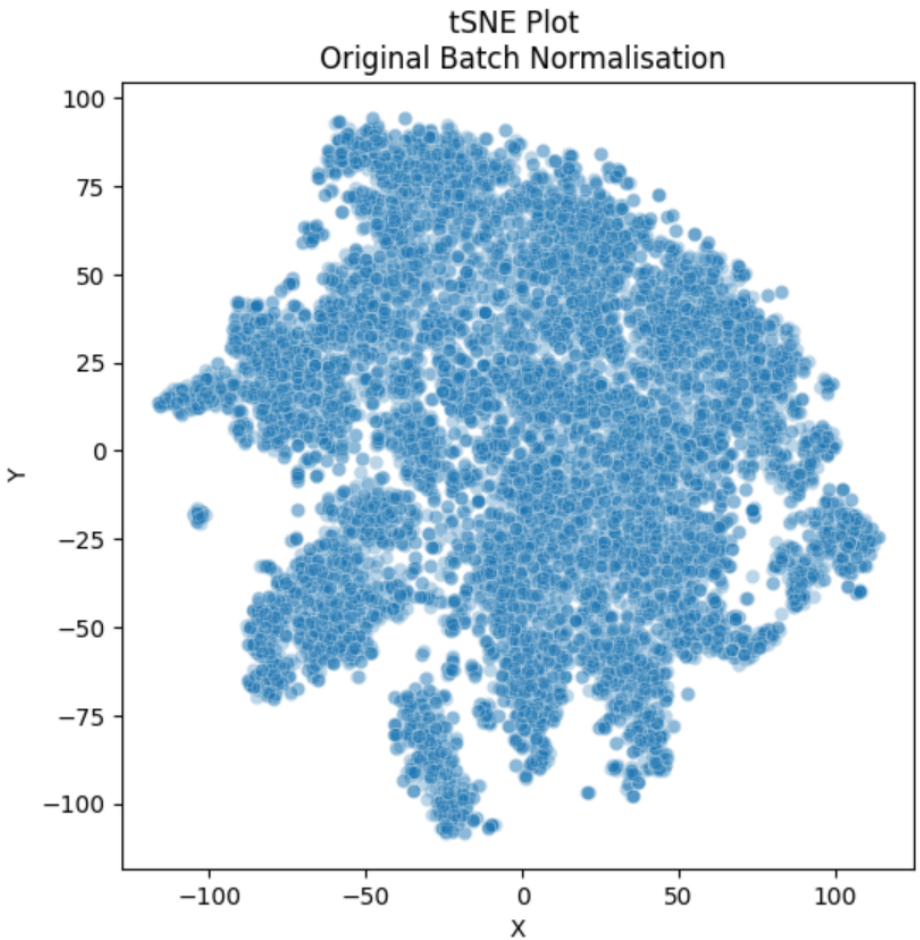
Note: Ten SMILES string had "explicit valence" error, so, are missed on the final embedding.

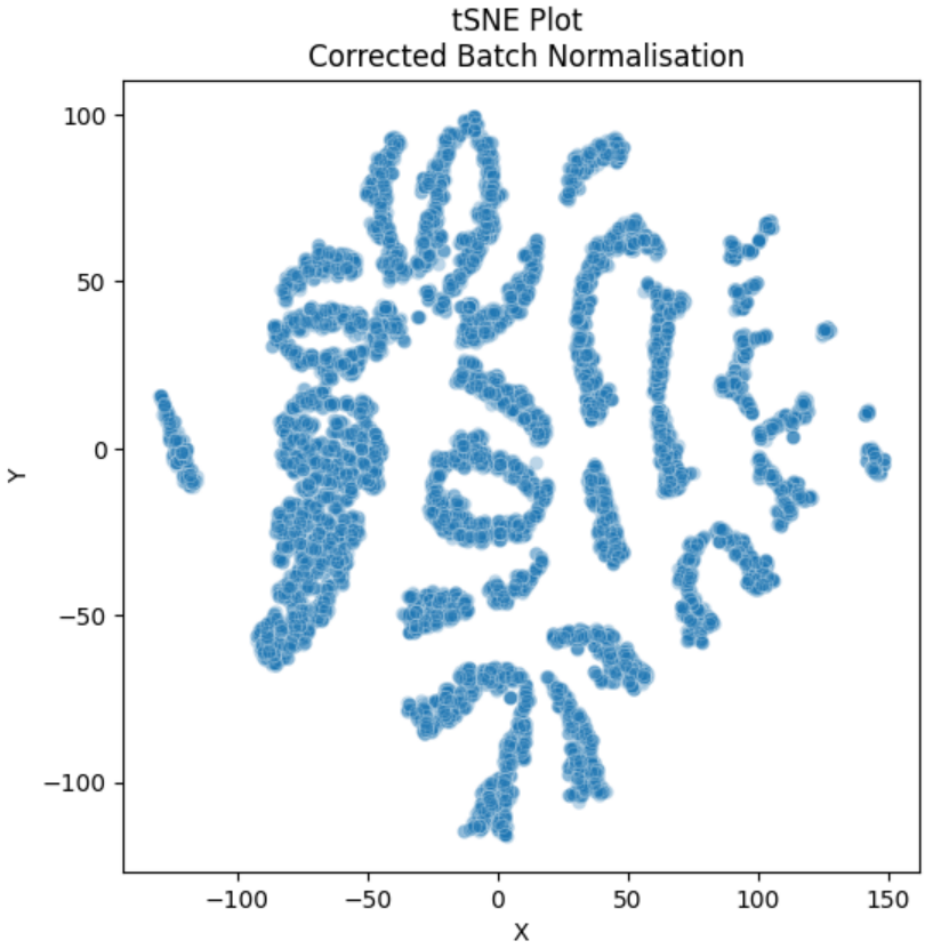


GNNs Embeddings

Improvment

Investigating the 2D tSNE projection of the embedded space, we realise that the batch normalisation layers are fixed for different populations on the original trained network (left Fig.). After updating the model for the task's molecules, we will see communities in the embedding space (right Fig.).

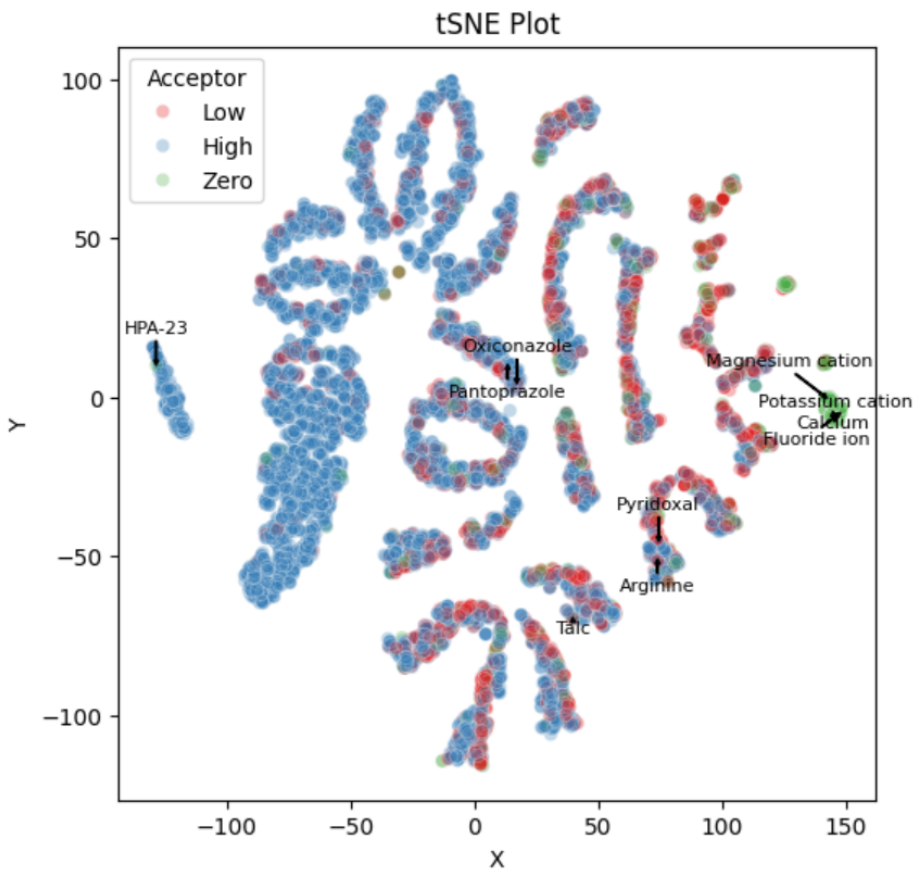


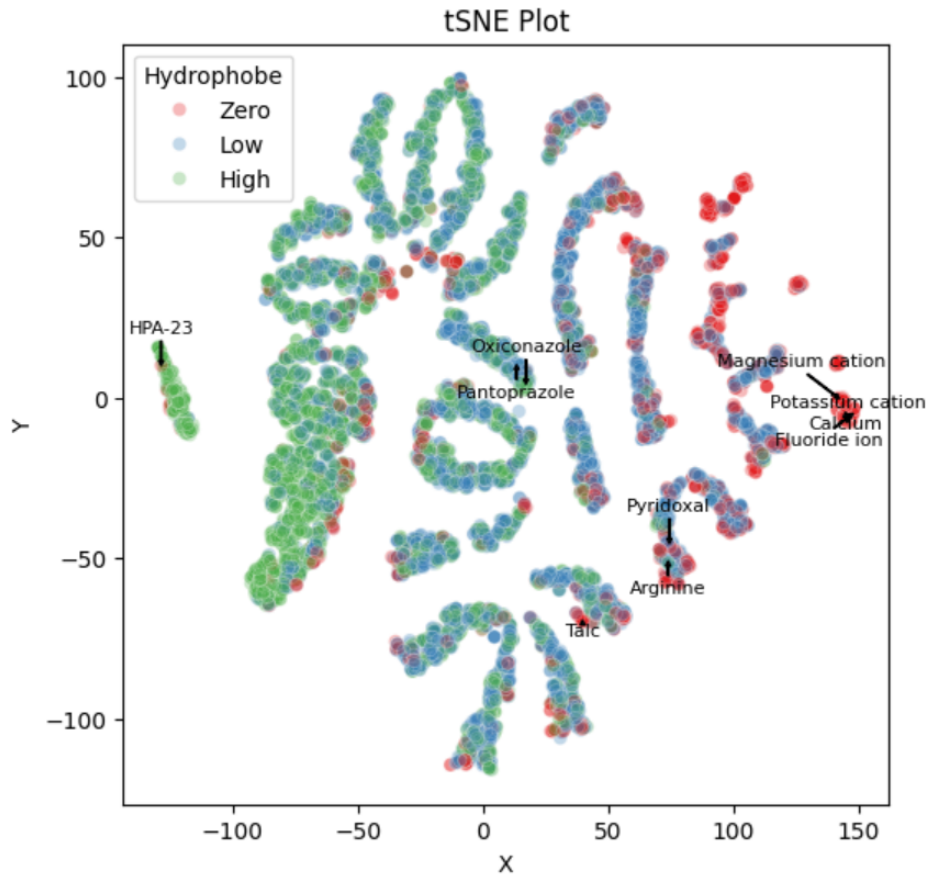


GNNs Embeddings

Clusters' Properties

Below, the embedded space is colour-coded based on the molecules' "acceptor" and "hydrophobe" features. We can see the clusters are more or less similar. Also, it is interesting to see some hand-selected ions embedded closely in this space on the middle right side of the plots.





GNNs Embeddings

Discussion and Future work

- Although the embedding space shows some merits regarding the similarity of close points, there is no reason to assume it can achieve a high score in the RAIDEN test since the original embedding was trained for a similar but not necessarily same downstream task.
- If we do not achieve an appropriate score, combining the current embedding architecture and the RAIDEN as the downstream task in the training phase can improve the embedding quality.
- A new idea for future work can be the introduction of molecular space similarity in an unsupervised fashion. For example, in [2], the author proposed a multi-layer network of different GNNs representing a tissue-specific protein-protein interaction. The tissue network "uses a hierarchy to model dependencies between protein graphs" [2].
- Domain knowledge about possible use cases for the downstream tasks (e.g., RAIDEN) can be incorporated by creating a hierarchy model based on the molecule features we think are relevant to the task.

References

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh.

Graphdta: predicting drug–target binding affinity with graph neural networks.

Bioinformatics, 37(8):1140–1147, 10 2020.

Marinka Zitnik and Jure Leskovec.

Predicting multicellular function through multi-layer tissue networks.

Bioinformatics, 33(14):i190–i198, July 2017.

arXiv:1707.04638 [cs].