Efficacy Comparison of Descriptor modeling and Graph modeling for Ligand Based Virtual Screening

Using the MUV dataset

Hugo Hakem – Meng Bioengineering

04/25/2024

Under the supervision of:

- Professor. Teresa Head-Gordon
- GSI. Yingze (Eric) Wang

Chem / BioE 242

Efficacy Comparison of Descriptor modeling and Graph modeling for Ligand Based Virtual Screening

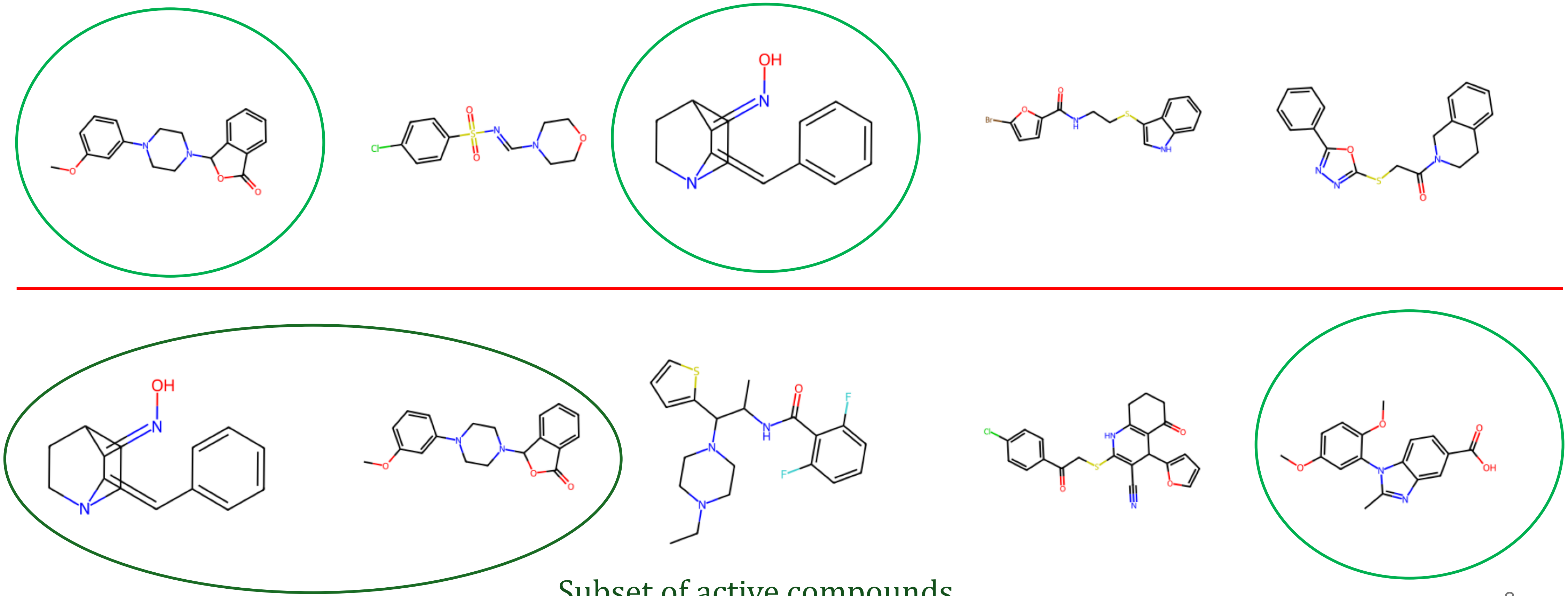
1.

Using the MUV dataset

1. Ligand Based Virtual Screening (LBVS)

- From a **large virtual databases** of available compounds, extract **small focused subsets** with an **enriched fraction of active compounds** (in regard of a target) in order to speed up biological testing. [1][2][3]

Subset of unactive compounds



Subset of active compounds

Efficacy Comparison of Descriptor modeling and Graph modeling for Ligand Based Virtual Screening

1.

Using the MUV dataset

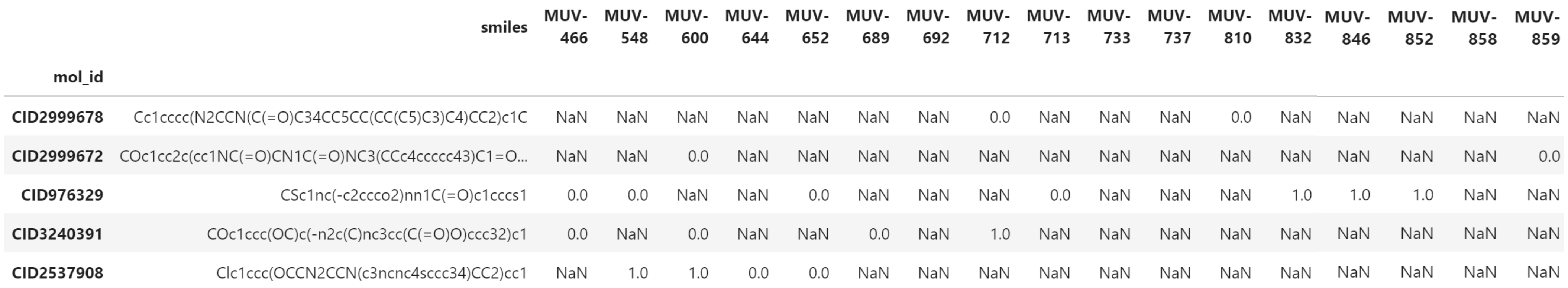
2.

2. MUV Dataset

- MUV dataset is created from bioactivity data from PubChem BioAssay. It can be downloaded on MoleculeNet [4].

17 Targets

93087 mols



	smiles	MUV-466	MUV-548	MUV-600	MUV-644	MUV-652	MUV-689	MUV-692	MUV-712	MUV-713	MUV-733	MUV-737	MUV-810	MUV-832	MUV-846	MUV-852	MUV-858	MUV-859
mol_id																		
CID2999678	Cc1cccc(N2CCN(C(=O)C34CC5CC(CC(C5)C3)C4)CC2)c1C	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
CID2999672	COc1cc2c(cc1NC(=O)CN1C(=O)NC3(CCc4ccccc43)C1=O...	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0
CID976329	CSc1nc(-c2ccco2)nn1C(=O)c1cccs1	0.0	0.0	NaN	NaN	0.0	NaN	NaN	NaN	0.0	NaN	NaN	NaN	1.0	1.0	1.0	NaN	NaN
CID3240391	COc1ccc(OC)c(-n2c(C)nc3cc(C(=O)O)ccc32)c1	0.0	NaN	0.0	NaN	NaN	0.0	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CID2537908	Clc1ccc(OCCN2CCN(c3ncnc4sccc34)CC2)cc1	NaN	1.0	1.0	0.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Statistics per Targets:
~14700 mols
~ 29 mols active
~ 0.2% Positive rate

Statistics on active compounds:
471 mols active 1 times
16 mols active 2 times
2 mols active 3 times

2. MUV Dataset

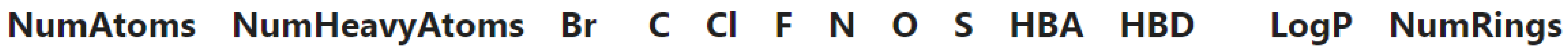
- Why this dataset ?

  → **Benchmarking Dataset for LBVS. It stands for Maximum Unbiased Validation Dataset** [2][3].

  - Adress Artificial Enrichment Bias:
    - **Unactive** compound **too different from Active** compound
  - Adress Analogue Bias:
    - **Active** compound **too similar** with each others

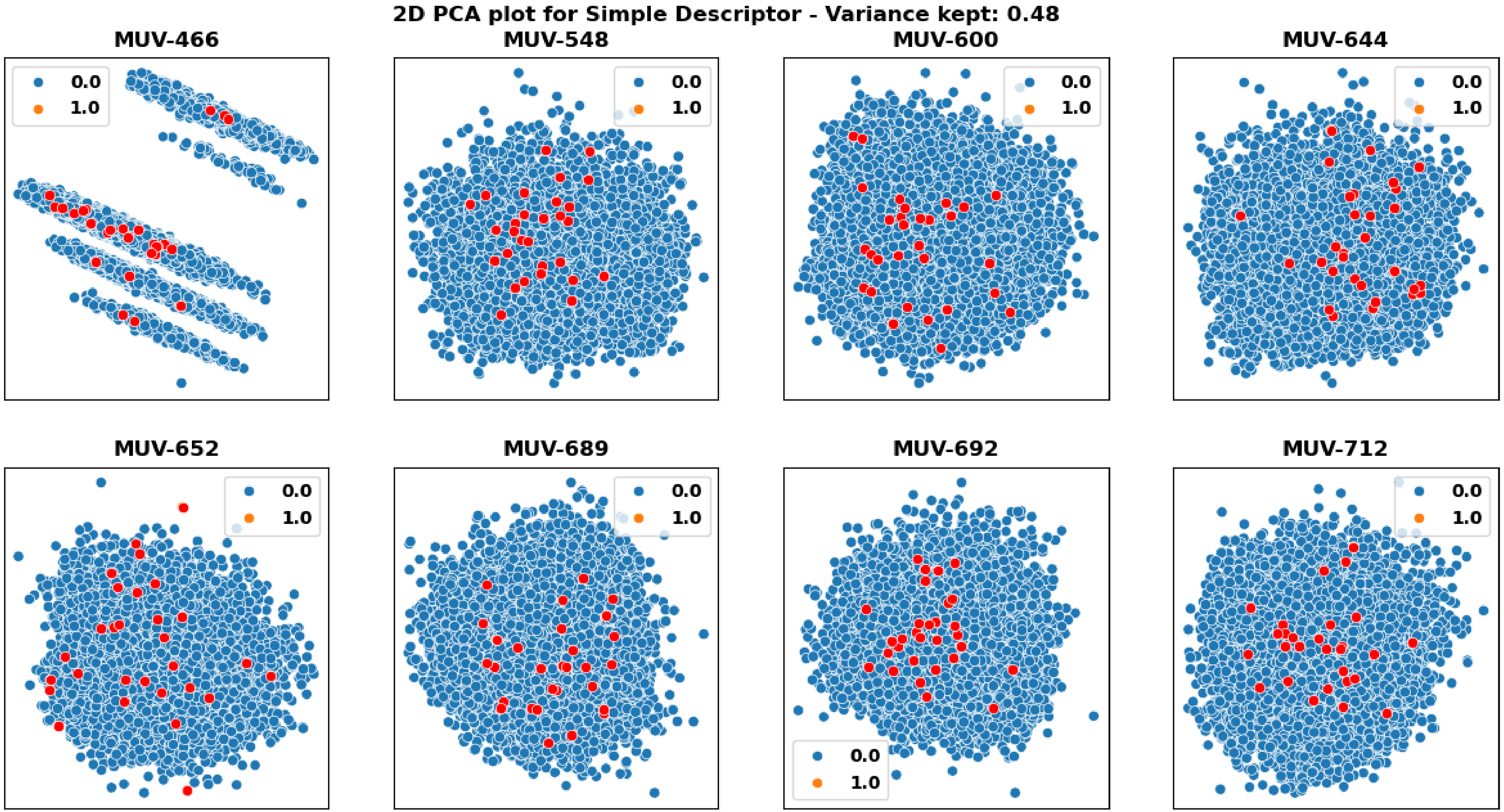
- How the chemical similarity has been computed?

  → In term of simple descriptor:



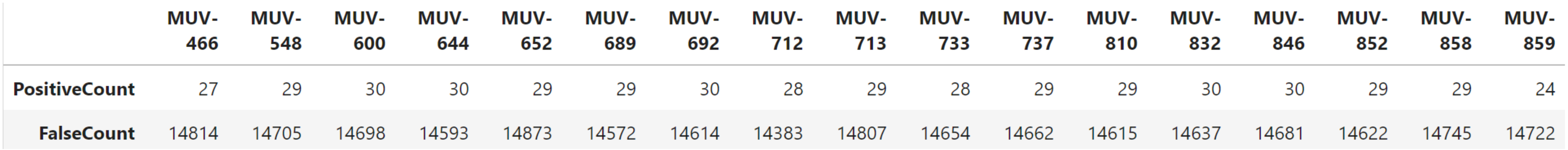
NumAtoms	NumHeavyAtoms	Br	C	Cl	F	N	O	S	HBA	HBD	LogP	NumRings

2. MUV Dataset



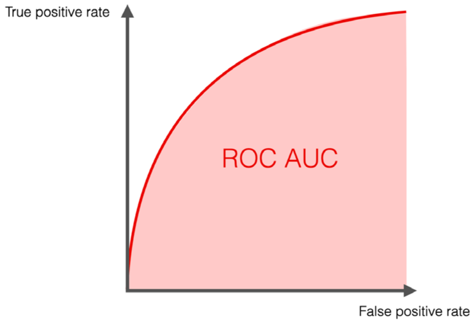
2. MUV Dataset

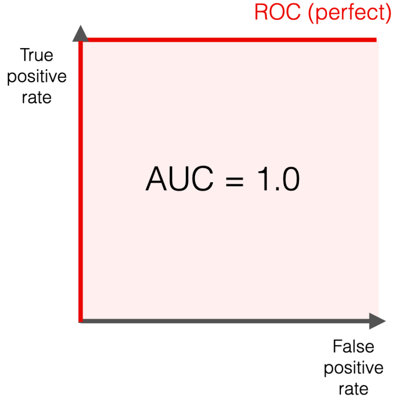
- **What is the goal with this Dataset?**
  → Perform all 17 Classification Tasks

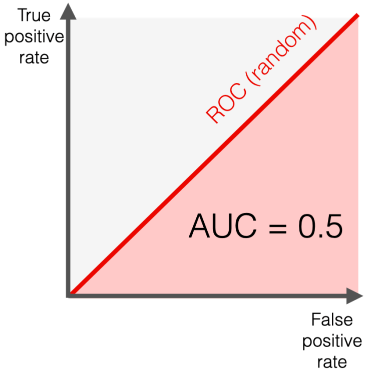


	MUV-466	MUV-548	MUV-600	MUV-644	MUV-652	MUV-689	MUV-692	MUV-712	MUV-713	MUV-733	MUV-737	MUV-810	MUV-832	MUV-846	MUV-852	MUV-858	MUV-859
PositiveCount	27	29	30	30	29	29	30	28	29	28	29	29	30	30	29	29	24
FalseCount	14814	14705	14698	14593	14873	14572	14614	14383	14807	14654	14662	14615	14637	14681	14622	14745	14722

- **Which metric of success?**
  → Very unalanced dataset: **ROC_AUC** / focused on positive classification [5]







3.

Efficacy Comparison of Descriptor modeling and Graph modeling for Ligand Based Virtual Screening

1.

Using the MUV dataset

2.

3. Descriptor Modeling

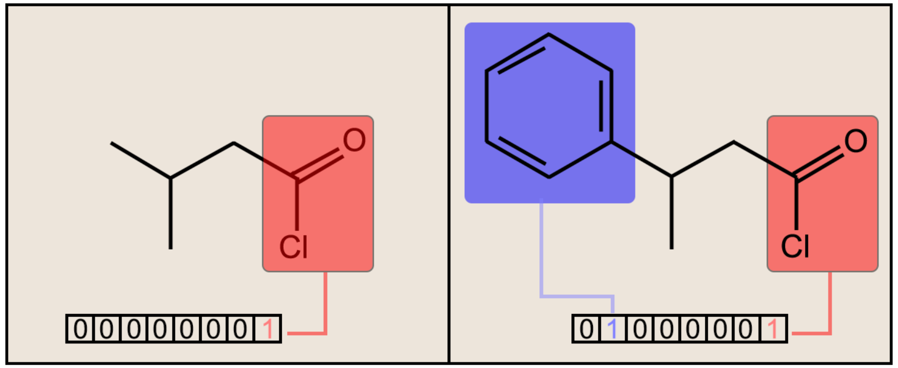
Which set of Descriptor ? [6]

- 13 Simple Descriptor, 1D

NumAtoms NumHeavyAtoms Br C Cl F N O S HBA HBD LogP NumRings

- 166 MACCS Keys Descriptor, 2D → After Processing 149

`rdkit.Chem.rdMolDescriptors.`**`GetMACCSKeysFingerprint`**`(`*`(Mol)mol`*`)`



Merge Descriptor 307

- 210 Complex Descriptor, mix of 1D and 2D → After Processing 203

`rdkit.Chem.Descriptors.`**`CalcMolDescriptors`**`(`*`(Mol)mol`*`)`

3. Descriptor Modeling

Which classification model?

Random Forest [✓ XGBoost]

Fine tuning:

- Loss Function [7]

$$Focal\ Loss(p) = -(y(1-p)^{\gamma}\log p + (1-y)p^{\gamma}\log(1-p))$$

`pip install imbalance-xgboost` [8]

- Other hyper-parameters:

n_estimator	max_depth	eta (learning rate)	focal_gamma
500	8	0.1	1.4

Was Grid Search possible?

- 20s training / target with CPU→ **5min40s** for 17 target and **22min40s** for each descriptors

3. Descriptor Modeling

Which Data Splitter?

- **Random** Train/Test: 0.8 / 0.2
- **Scaffold [9]** Train/Test: 0.8/ 0.2

  → Split made on core structure of molecules, to make the model learn on a Train split with very different molecule than the Test split.

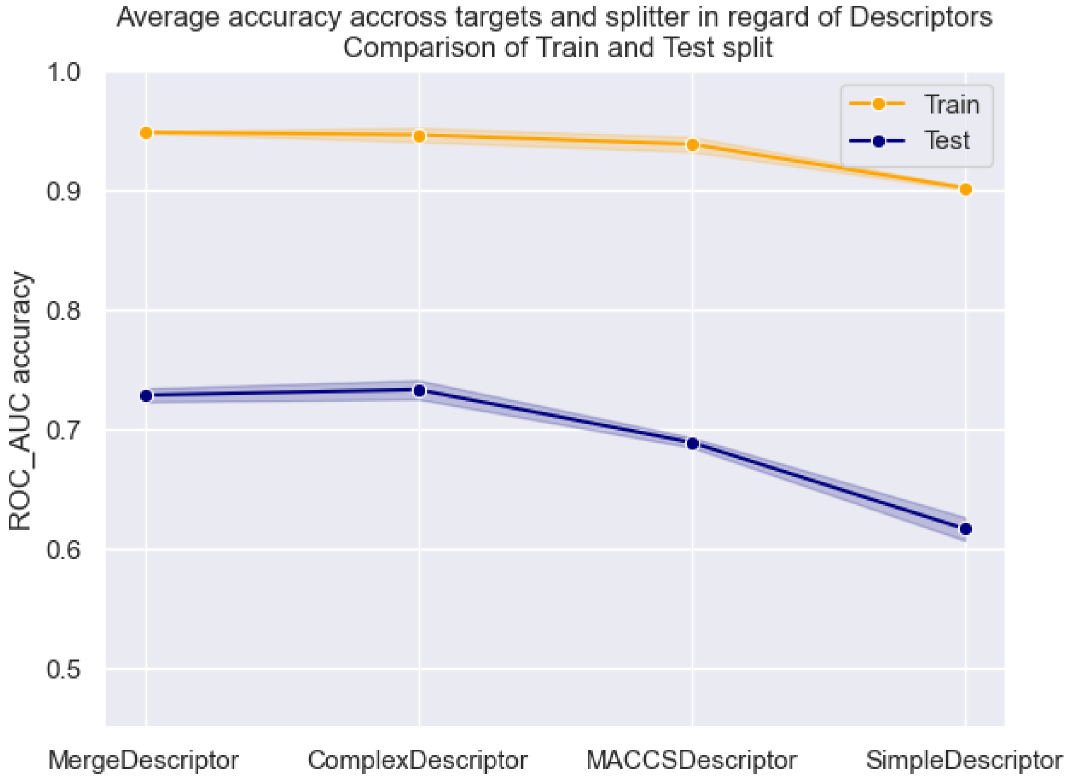
    → It challenges the generalization of the model.

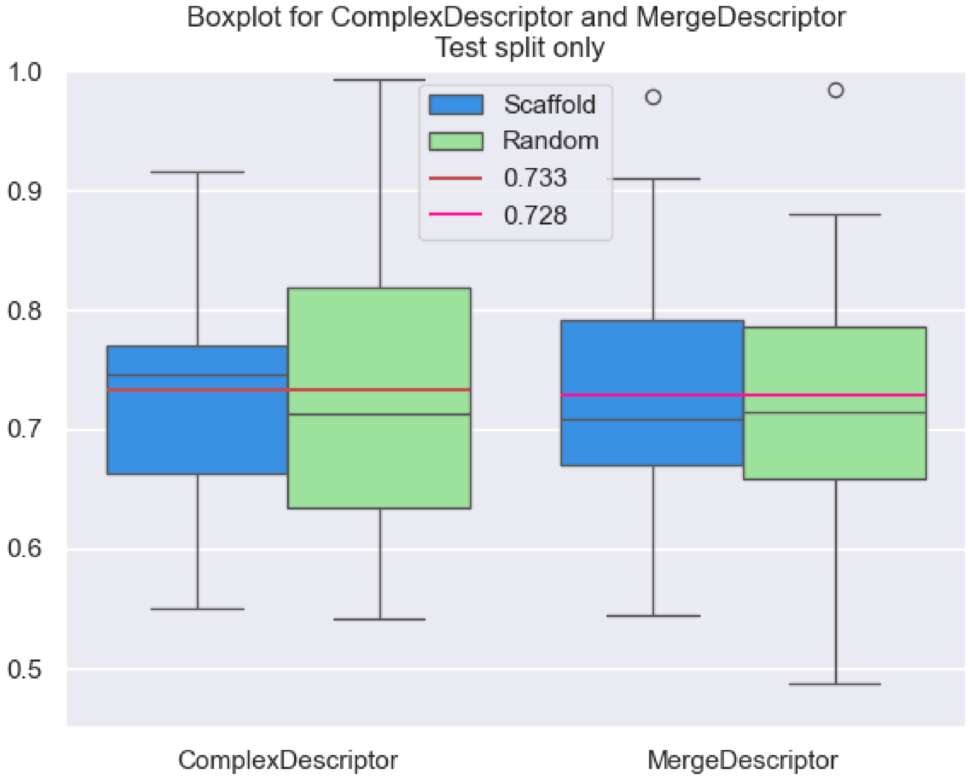
Since data are **unbalanced**:

- Need to make sure that 0.8/0.2 of positive in Train/Test

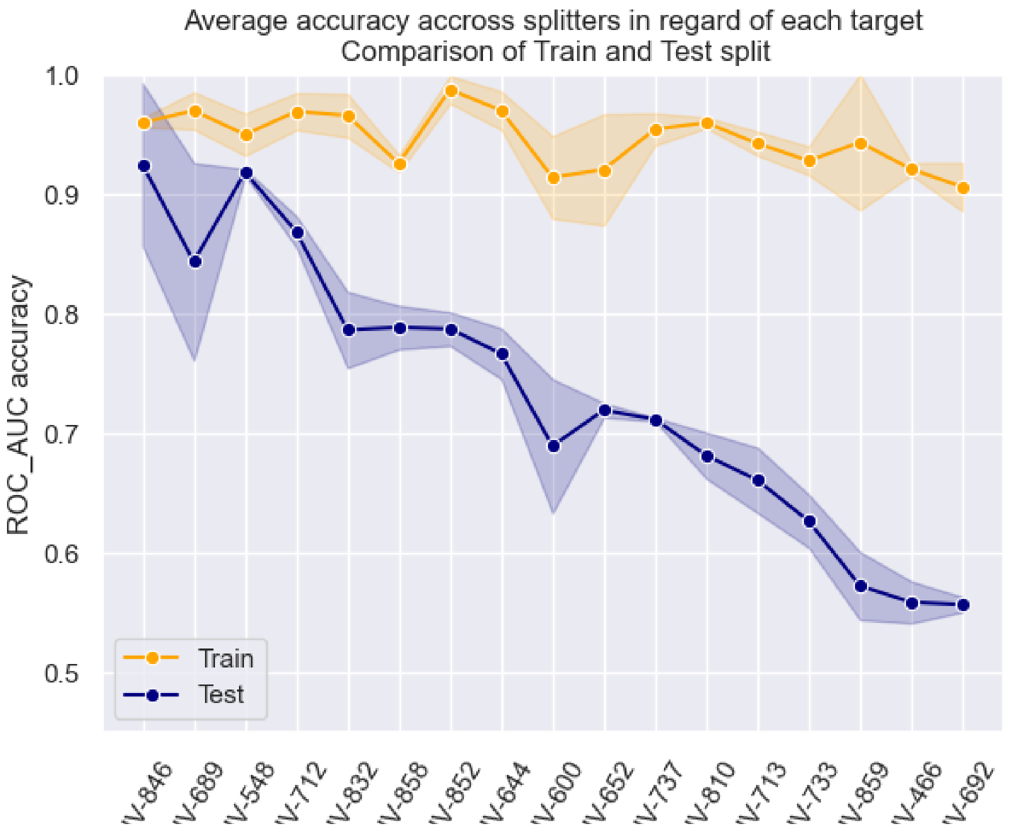
To not miss-estimate the performance of Random Split, **3 generation of Random split is tested and averaged**

3. Descriptor Modeling

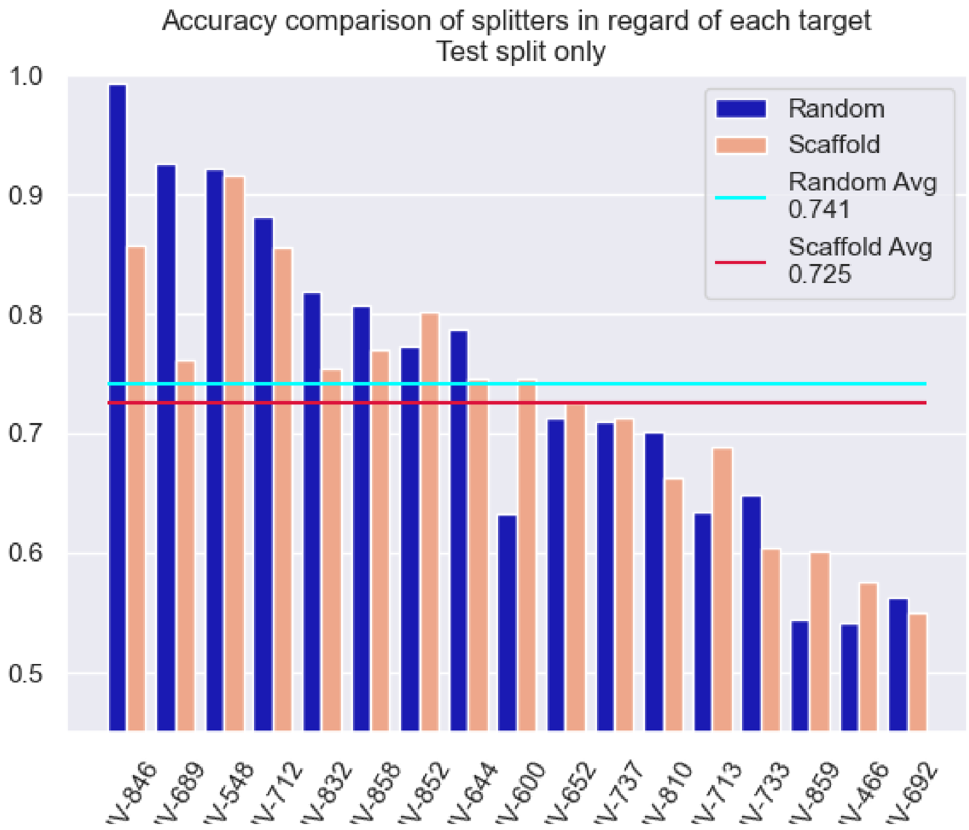




3. Descriptor Modeling



Average accuracy accross splitters in regard of each target
Comparison of Train and Test split



Accuracy comparison of splitters in regard of each target
Test split only

3.

Efficacy Comparison of Descriptor modeling and Graph modeling for Ligand Based Virtual Screening

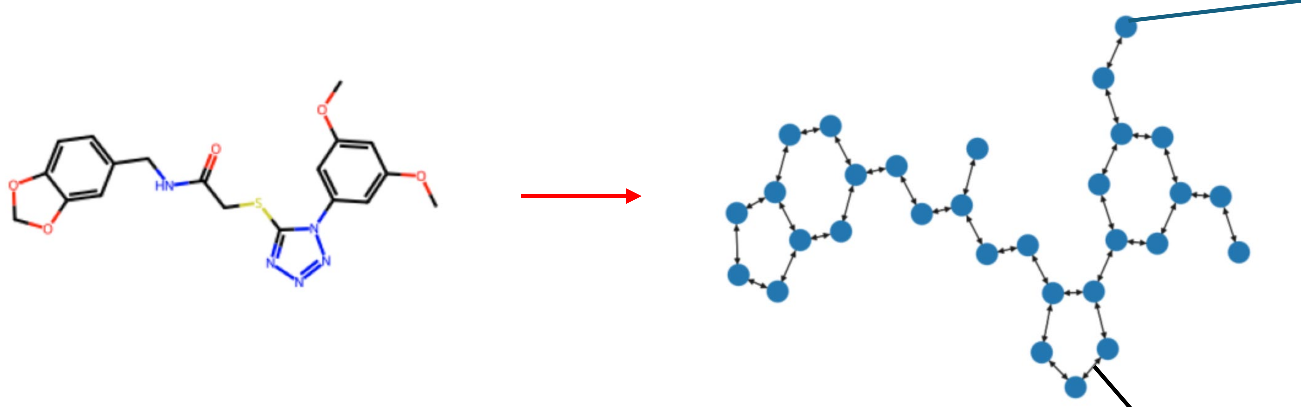
4. 1.

Using the MUV dataset

2.

4. Graph Modeling

Molecule Graph Representation, downloadable using torchgeometric, based on feature from OGB database [10]



Node Features 9
bond type: [6, 7, 8, 9, 16, 17, 35]
~~chirality: [0]~~
degree: [1, 2, 3, 4]
formal electric charge: [4, 5, 6]
number of hydrogen atoms connected: [0, 1, 2, 3]
~~number of radical electrons: [0]~~
hybridization state: [2, 3, 4]
part of ring: [0, 1]
aromaticity: [0, 1]

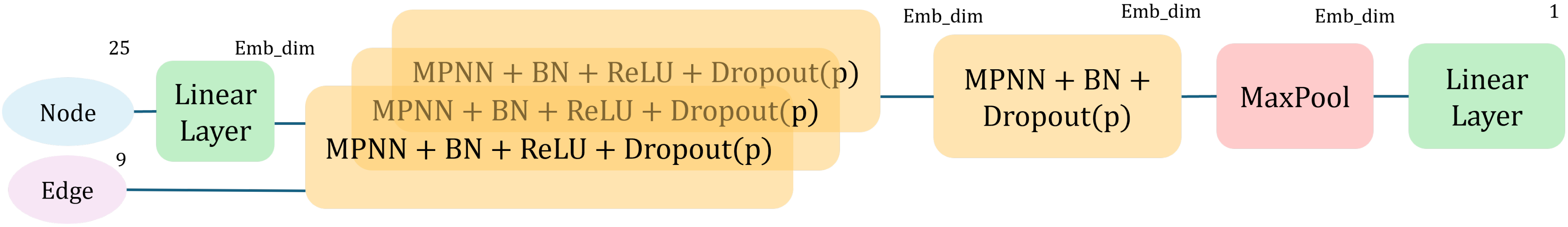
Bond Features 3
bond type: [1, 2, 3, 12]
stereoisomery info: [0, 2, 3]
conjugation: [0, 1]

One-Hot-Encoding of those categorical features:

- Node features: 25
- Bond features: 9

4. Graph Modeling

Graph Neural Network architecture inspired from OGB gitHub [10]:



Model hyper-parameters

Num_MPNN	Drop_ratio(p)	Emb_dim
4	0.1	100

Trainer hyper-parameters

Batch_size	Epoch	Learning_rate	L2	Gamma
128	30	5.10^{-4}	10^{-5}	1.4

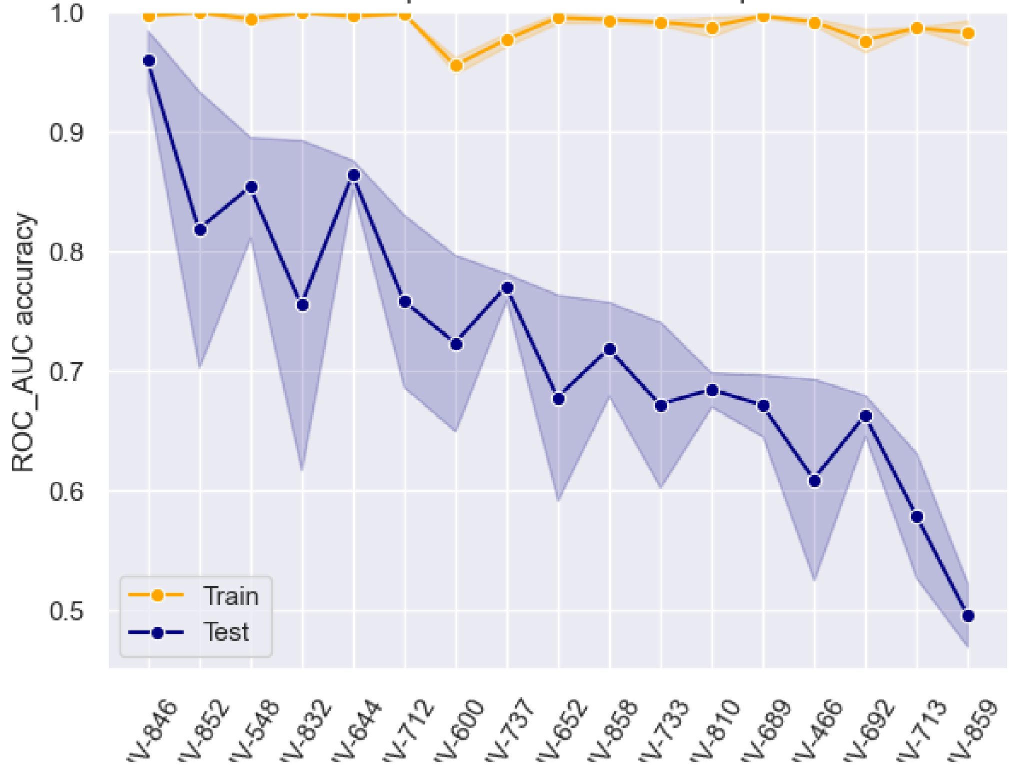
Was Grid Search possible?

- 3min training / target with CUDA → **51min** for 17 target
- Subset of **1 Target known to be challenging thanks to XGBoost**

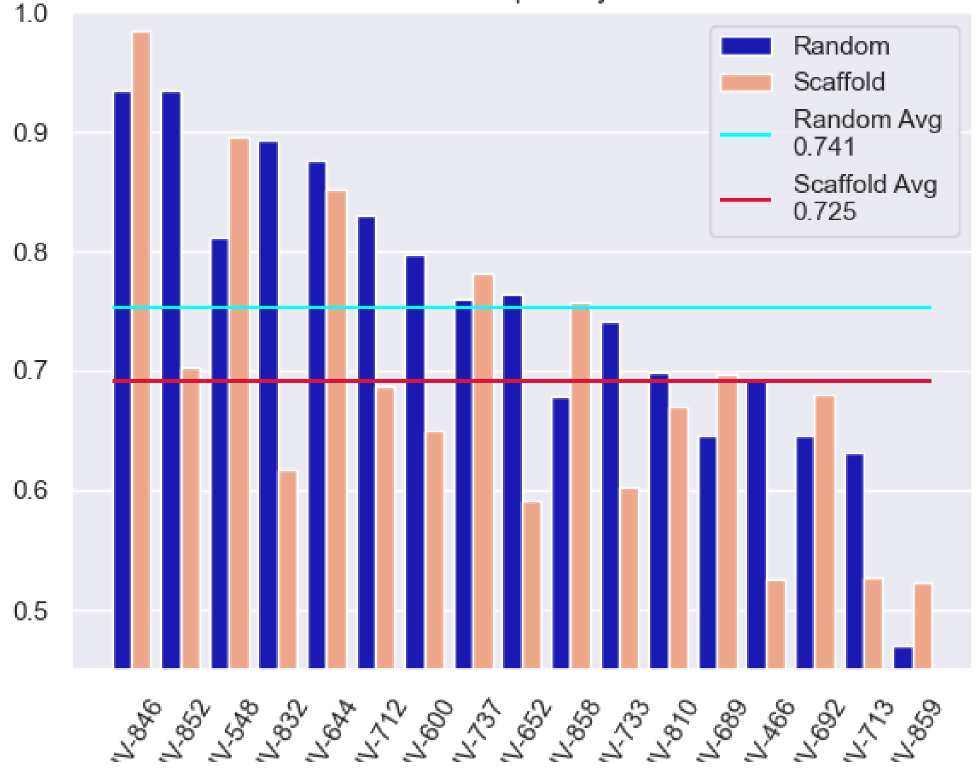
4. Graph Modeling

Accuracy Result

Average accuracy accross splitters in regard of each target
Comparison of Train and Test split

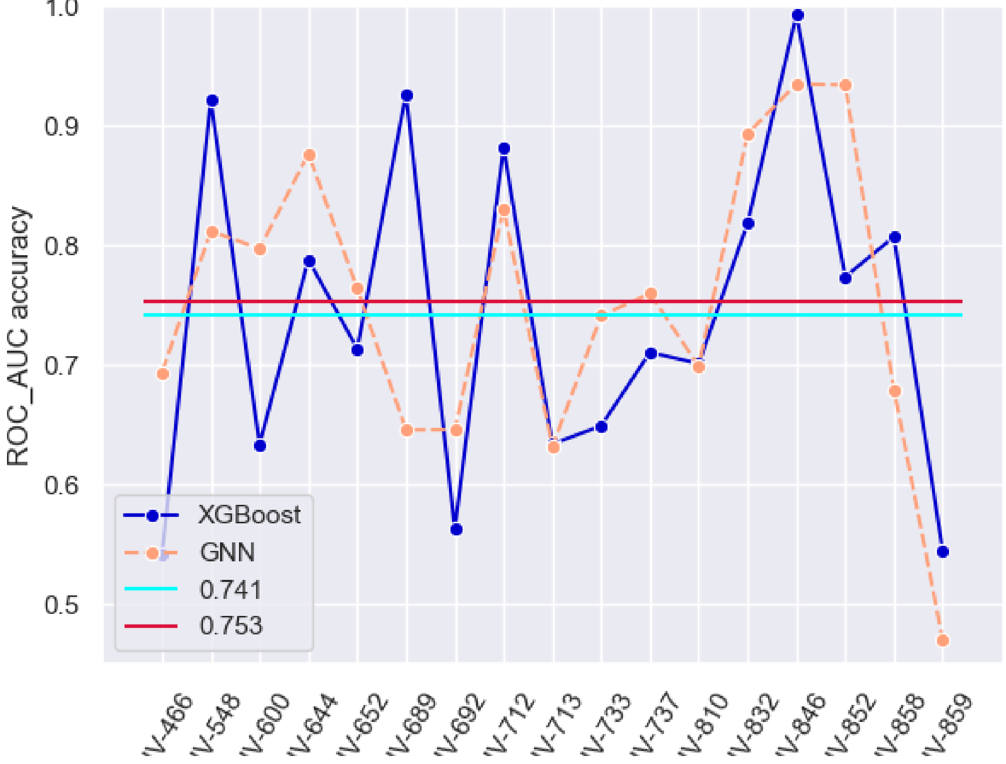


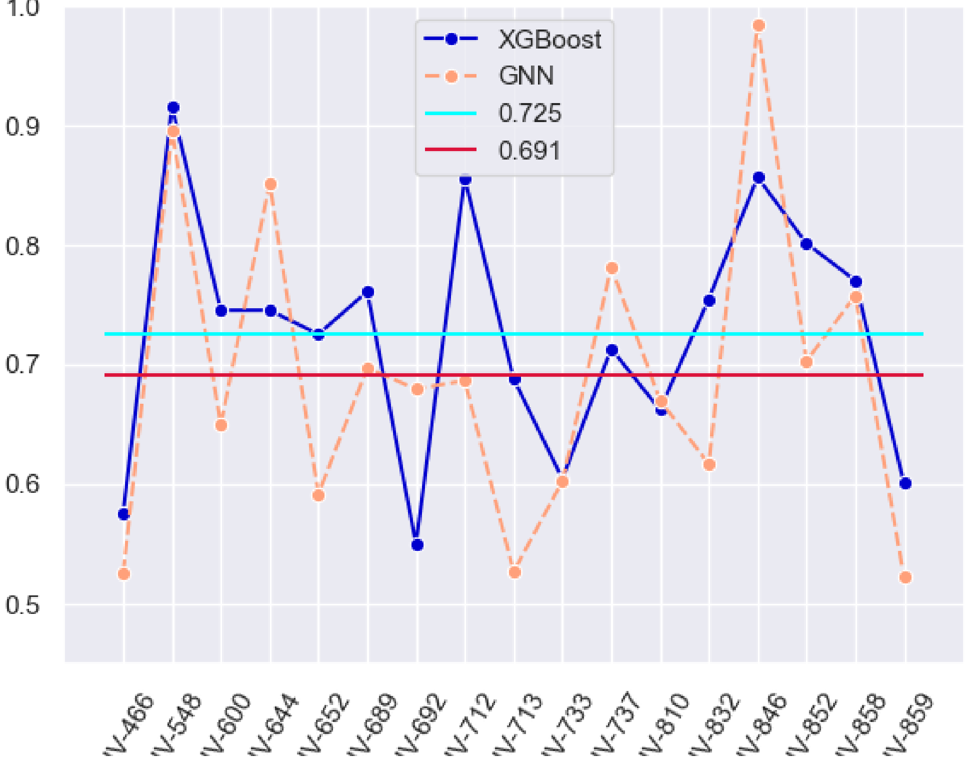
Accuracy comparison of splitters in regard of each target
Test split only



4. Graph Modeling

Test Accuracy comparison with XGBoost


Accuracy comparison of GNN and XGBoost for Random Split
Test split only


Accuracy comparison of GNN and XGBoost for Scaffold Split
Test split only

Conclusion

Test Accuracy comparison with litterature

XGBoost / split	Scaffold Split	Random Split
MoleculeNet [4]	//	0.720
Mine	**0.725**	**0.741**

GNN / split	Scaffold Split	Random Split
MoleculeNet [4]	//	0.775
GNN + Dummy super node [7]	**0.789**	0.823 (by averaging their valid/test)
TrimNet [11]	//	**0.851** (by averaging their valid/test)
Mine	0.691	0.753

Conclusion

Pros and Cons for each model *(Efficacy Comparison of Descriptor modeling and Graph modeling for Ligand Based Virtual Screening)*

Feature / Model	XGBoost	GNN
Ease of tunning	✓	X
Speed	✓ (20s)	X (3min)
Freedom for tunning	X	✓
Potential to achieve greater accuracy	X (0.741)	✓ (0.851)

Conclusion

Success:

1. **Created a more efficient XGBoost than in the litterature**

   → Showed the importance of chosing the right Descriptor set
2. **Created a GNN performance with comparable performance to litterature**
3. **Addressed the shortcoming of the litterature** in showing the **performance difference for each targets**

Critics of my pathway:

1. **XGBoost tunning with Complex Descriptor instead of Merge Descriptor.**

   → It may bias the result saying that Complex, better than Merge
2. **GNN tunning on 1 target.**

   → May bias the fine tunning resulting in a not so great overall accuracy.
3. **GNN, One-Hot-Encoding instead of Embedding layer**

Thank You

Under the supervision of
Professor. Teresa Head-Gordon
GSI. Yingze (Eric) Wang

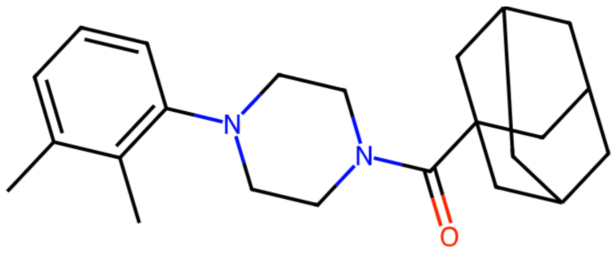
Special Credit to my friend for challenging me
Adrien Bourgain, Meng BioE

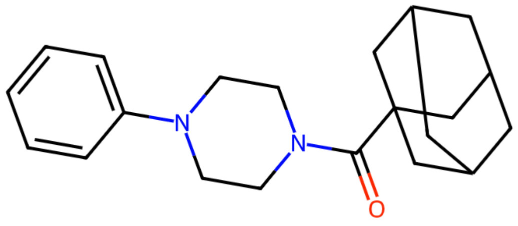
References

[1] Gimeno, Aleix, et al. « The Light and Dark Sides of Virtual Screening: What Is There to Know? » International Journal of Molecular Sciences, vol. 20, no 6, march 2019, p. 1375. PubMed Central, https://doi.org/10.3390/ijms20061375.
[2] Rohrer, Sebastian G., et Knut Baumann. « Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data ». Journal of Chemical Information and Modeling, vol. 49, no 2, february 2009, p. 169-84. DOI.org (Crossref), https://doi.org/10.1021/ci8002649.
[3] Xia, Jie, et al. « Benchmarking methods and data sets for ligand enrichment assessment in virtual screening ». Methods, vol. 71, january 2015, p. 146-57. ScienceDirect, https://doi.org/10.1016/j.ymeth.2014.11.015
[4] Datasets. https://moleculenet.org/datasets-1. Consulted 4 april 2024.
[5] « How to explain the ROC AUC score and ROC curve? » Consulté le: 25 avril 2024. [web]. Available: https://www.evidentlyai.com/classification-metrics/explain-roc-curve
[6] « 6.1: Molecular Descriptors ». Chemistry LibreTexts, 26 october 2019, https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/06%3A_Molecular_Similarity/6.01%3A_Molecular_Descriptors.
[7] J. Li, D. Cai, et X. He, « Learning Graph-Level Representation for Drug Discovery ». arXiv, 15 septembre 2017. Consulted: 25 april 2024. [Web] Available: http://arxiv.org/abs/1709.03741
[8] C. Wang, C. Deng, et S. Wang, « Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost ». 2019.
[9] « Introduction to Scaffold Splitting - Oloren AI ». Consulted: 25 april 2024. [Web]. Available: https://www.oloren.ai/blog/scaff-split
[10] W. Hu et al., « Open Graph Benchmark: Datasets for Machine Learning on Graphs », arXiv preprint arXiv:2005.00687, 2020.
[11] P. Li et al., « TrimNet: learning molecular representation from triplet messages for biomedicine », Briefings in Bioinformatics, nov. 2020, doi: 10.1093/bib/bbaa266.

Appendix

Scaffold Splitting [7]



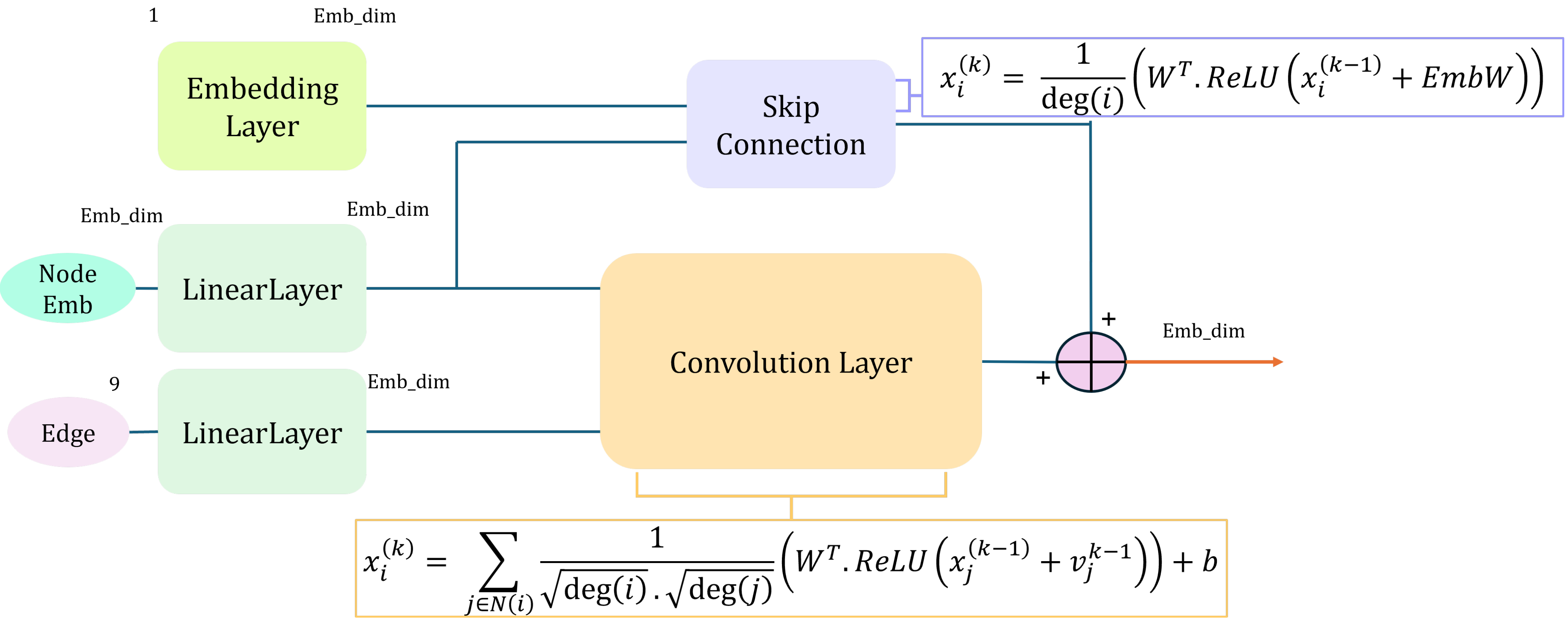


- **Group molecules by scaffold.**
- **Sort** by the number of molecules by scaffold.
- Molecules withing a **scaffold group with a large occurrence** will be put in priority in the **Training Split** until reaching 80%.
- Put the rest into the test split.

- Actually, a **difference is made between positive and negative** as the data set is unbalanced.

4. Graph Modeling

MPNN architecture detailed:

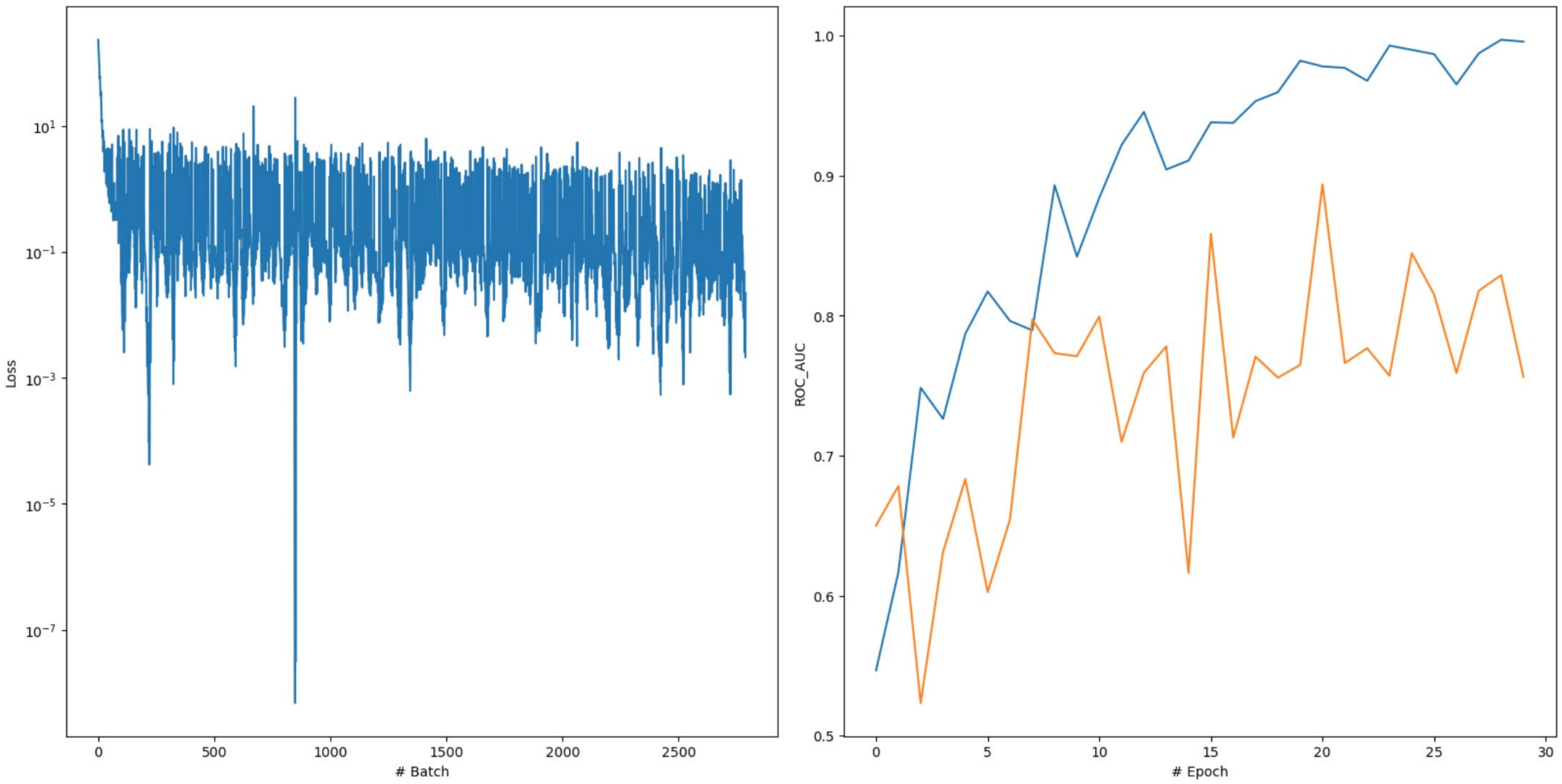


$$x_i^{(k)} = \frac{1}{\deg(i)}\left(W^T . ReLU\left(x_i^{(k-1)} + EmbW\right)\right)$$

$$x_i^{(k)} = \sum_{j\in N(i)} \frac{1}{\sqrt{\deg(i)}\,.\sqrt{\deg(j)}}\left(W^T . ReLU\left(x_j^{(k-1)} + v_j^{k-1}\right)\right) + b$$

4. Graph Modeling

Example of training for one arbitrary target


Loss
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
# Batch
ROC_AUC
# Epoch