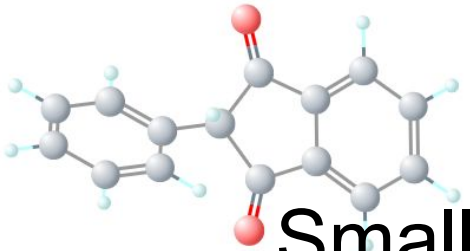
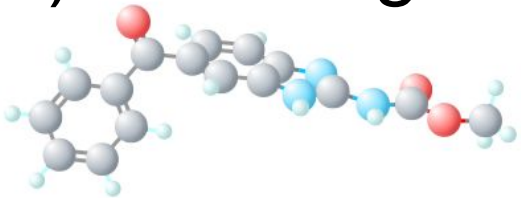


Small Molecule Property Prediction: Machine learning Tools For Quantitative Structure-Activity relation (QSAR) Modeling

Presented by *YA-TING CHANG*



INTRODUCTION

- *MicroRNAs(miRNA)* are small endogenously transcribed regulatory RNA which modulates gene expression
- Significant evidence has showed the fundamental role of miRNAs *in the development of many diseases*
- RNA-binding *small molecules* offer an attractive strategy for modulating microRNAs' function
- Utilize Machine Learning techniques for *predictive modeling of small molecules* with potential to inhibit specific miRNA

MOLECULE REPRESENTATION

Three types of molecule representation:

- Whole molecule (1D)
  Bulk molecular properties such as the number of stereo-centres, molecular weight…

- 2D representation ⇦ Focus in this project
  - Computed from a chemical structure diagram that is encoded as a connection table detailing all of the atoms and bonds in a molecule as a labelled graph
  - Most important type: descriptor and fingerprint

- 3D representation
  Calculated from the 3D coordinates of the atoms. Capture the 3D information regarding the molecule size, shape and atom distribution

FRAMEWORK

Algorithms

- Supervised: XGBoost
- Unsupervised: Gaussian Mixture Models

Software

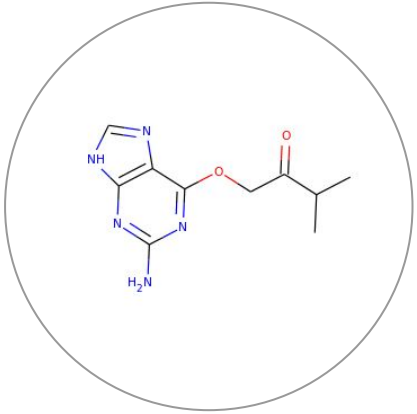
- Descriptor Generation: Mold2
- Fingerprint Generation: CDK(Java’s Chemistry Development Kit), R
- Modeling: Python scikit-learn
- Computing Platform: PSC Bridges

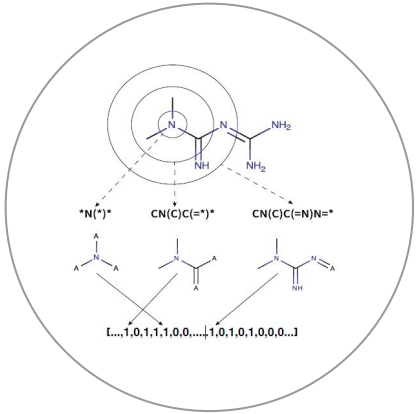
Dataset

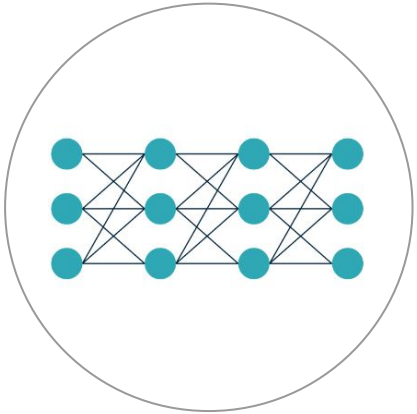
PubChem, ChEMBL, DrugBank

Supervised Machine Learning: *Molecule Prediction*

WORKFLOW







Compound Extraction

- Source: PubChem
- Target: miR21
- Each molecule is labeled as active or inactive
- Among 302,630 molecules, only 648 are active

Chemical Property Creation

- Generate 777 descriptors for each molecule

ML modeling

- Use xGboost to tackle imbalanced data set
- Put more weight on active cases

APPLICATION

Molecule from DrugBank

	Descriptor 1	Descriptor 2	⋯	**Label**
M 1				?
M 2				?
⋮				⋮

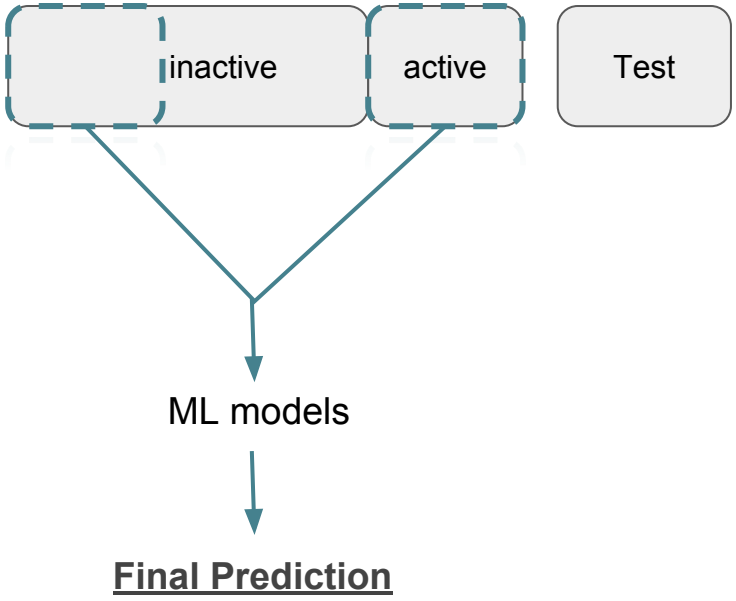
ML model

Prediction

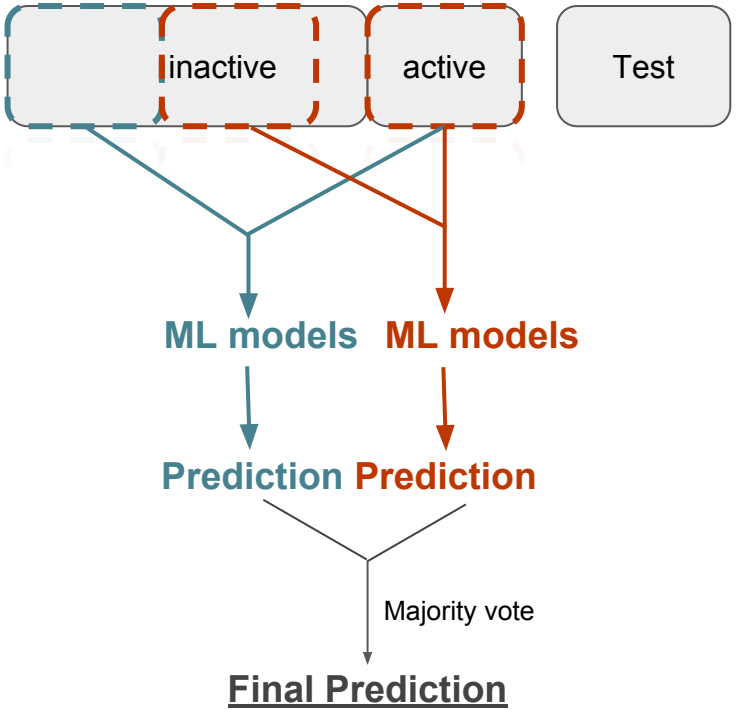
	Descriptor 1	Descriptor 2	⋯	**Label**
M 1				inactive
M 2				inactive
⋮				⋮

METHODS

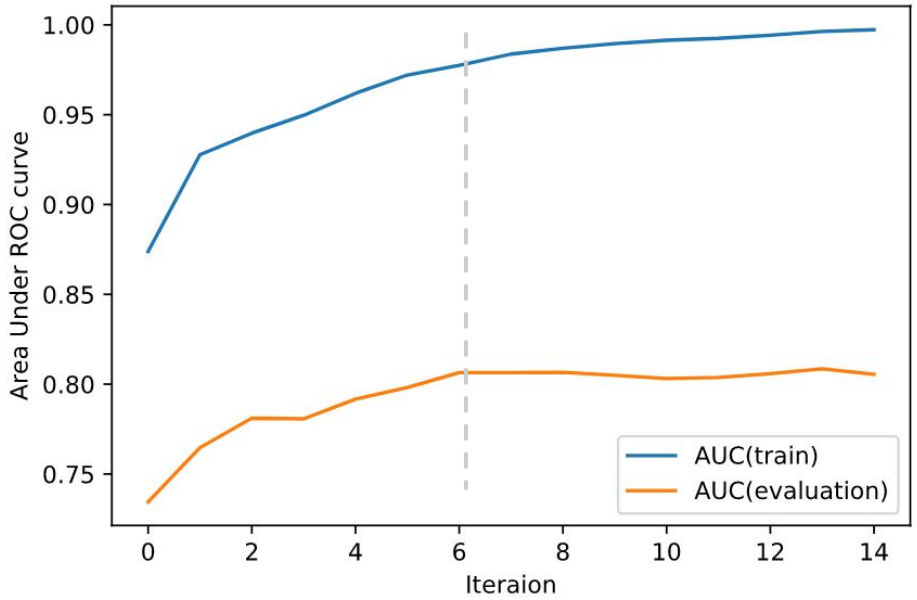
<u>without bagging + under Sampling</u>

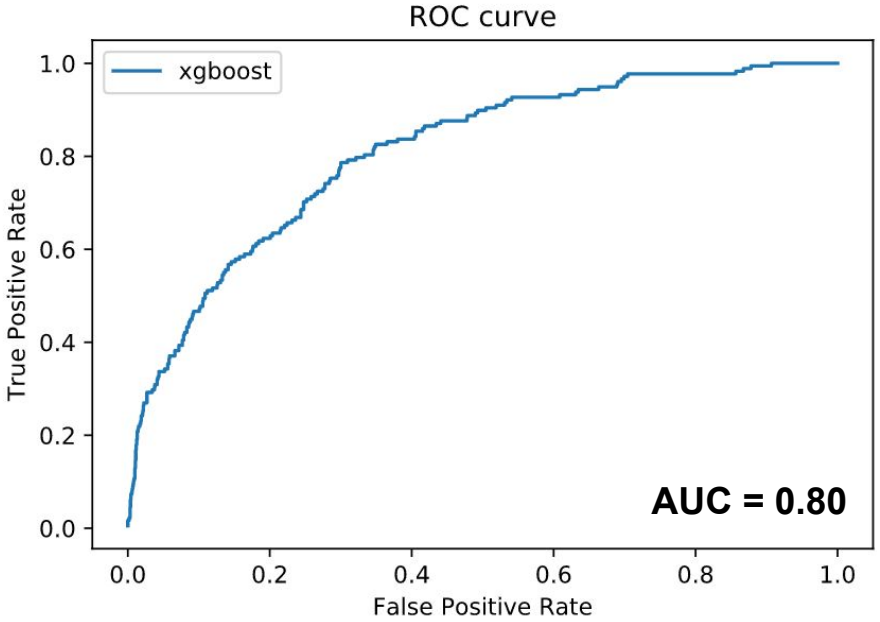


<u>bagging + under Sampling</u>



EVALUATION (without bagging)



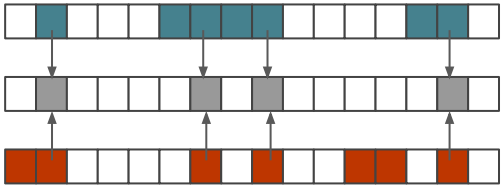


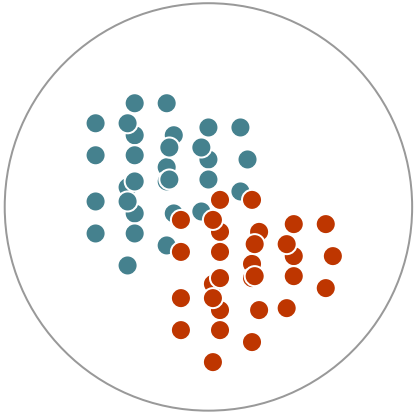
EVALUATION (without bagging)

Accuracy: 0.81
True positive rate: 0.58
False Negative rate: 0.41

Unsupervised Machine Learning: *Similarity Analysis*

WORKFLOW





Compound Extraction

- Source: ChEMBL
- Target: five proteins target

Chemical Property Creation

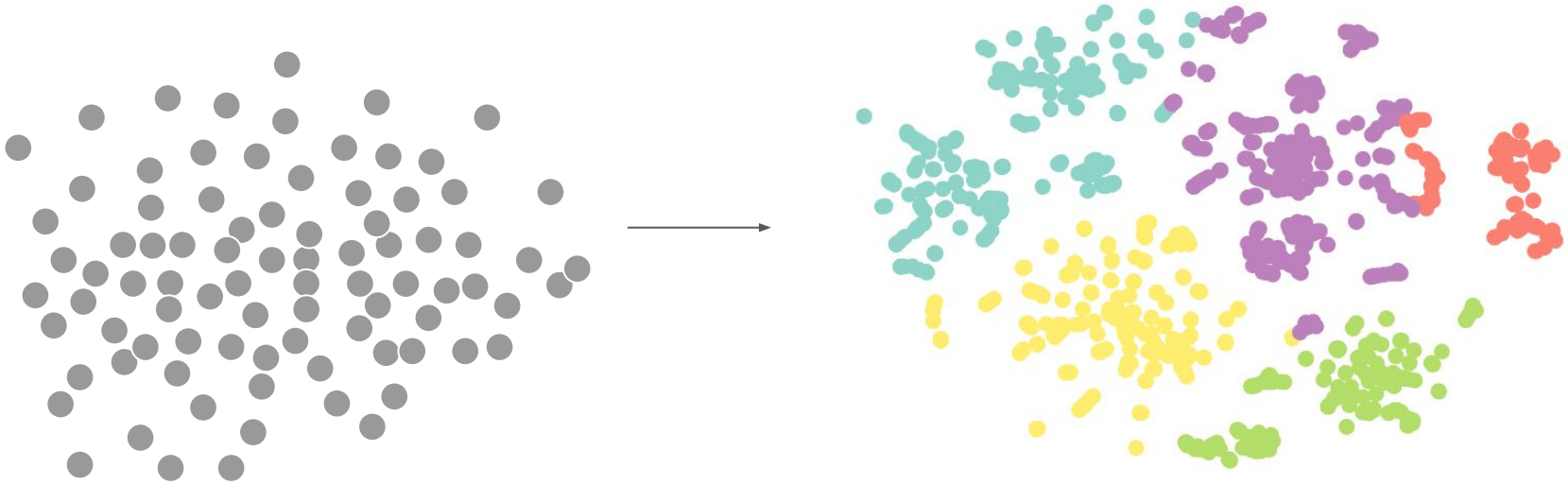
- Generate 1,024 fingerprints

Database Clustering

- Use Gaussian Mixture Models for clustering

RESULT

Similarity Analysis Assumption:
Molecules that are structurally similar are likely to have similar properties



FUTURE WORK

- Incorporate multiple data types and sources that aggregate structural, genetic and pharmacological data for modeling

- It’s likely that 3D characteristics of molecule will play an increasingly important role in chemoinformatics

- Recent evolution in deep learning networks has proven to be promising architecture for efficient learning from massive dataset for modern drug discovery

REFERENCES

[1] Karthikeyan, M., & Vyas, R. (2014). Machine Learning Methods in Chemoinformatics for Drug Discovery. Practical Chemoinformatics, 133-194. doi:10.1007/978-81-322-1780-0_3

[2] D., & Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: Paradigm for drug design. Drug Discovery Today, 21(8), 1291-1302. doi:10.1016/j.drudis.2016.06.013

[3] Heikamp, K., & Bajorath, J. (2011). Large-Scale Similarity Search Profiling of ChEMBL Compound Data Sets. Journal of Chemical Information and Modeling, 51(8), 1831-1839. doi:10.1021/ci200199u

[4] Willett, P. (2014). The Calculation of Molecular Structural Similarity: Principles and Practice. Molecular Informatics, 33(6-7), 403-413. doi:10.1002/minf.201400024

[5] ajorath, J. (2010). ChemInform Abstract: Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. ChemInform, 32(22). doi:10.1002/chin.200122290

THANK YOU!