Molecular property prediction with quantum chemical and chemoinformatics approach

○Takayuki Serizawa, Kazufumi Okawa, Takaya Yamaguchi, Kenichiro Takaba
Laboratory for Medicinal Chemistry, Pharmaceuticals Research Canter, Asahi Kasei Pharma Corporation, 632-1 Mifuku, Izunokuni, Shizuoka 410-2321, Japan

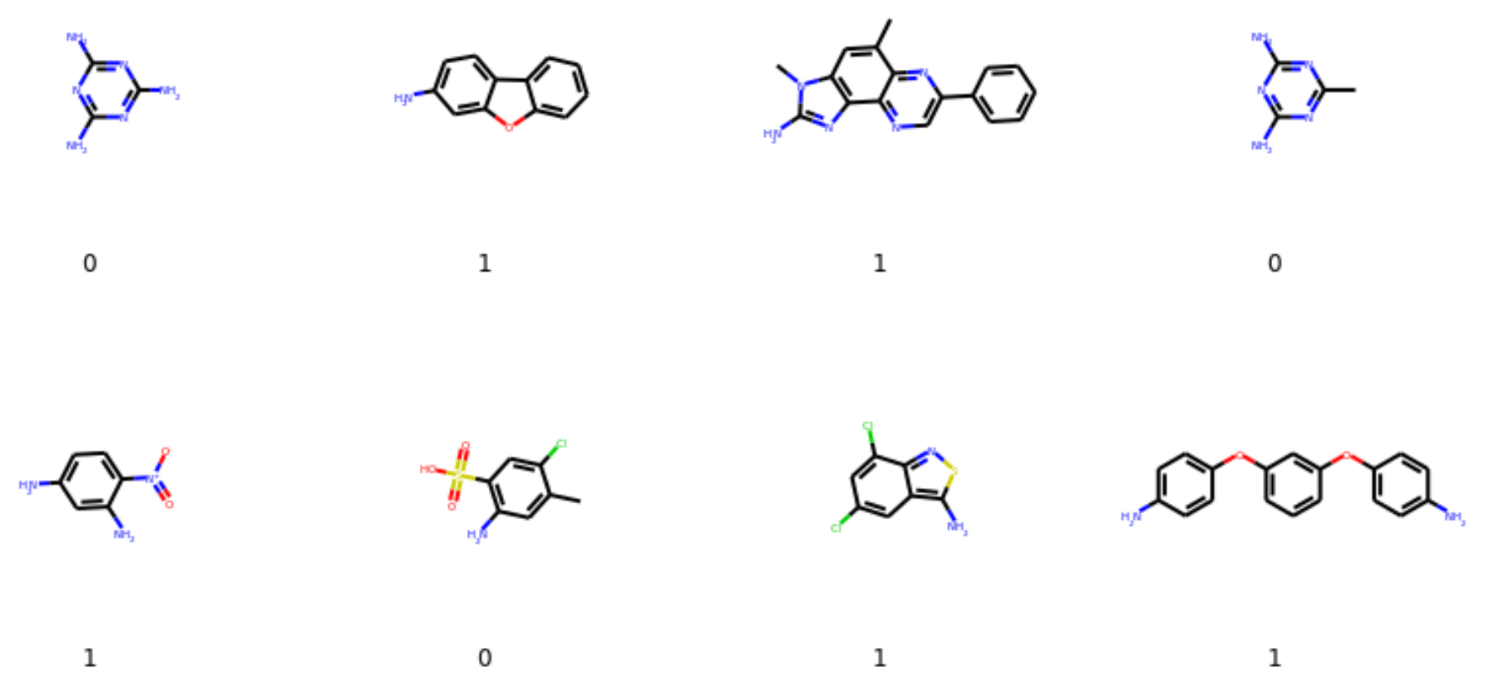
1. Introduction

We have applied machine learning to several internal projects. However, for such system to be truly accelerate drug discovery projects, it is necessary to improve the accuracy of predictive model.

Two approaches are investigated, one is quantum chemistry based which uses molecular energy. And the other is deep learning(GNN) with a new input feature from quantum chemistry. Here we present our effort of quantum chemistry and chemoinformatics approach for molecular property prediction.

2. Data Source[1]

Tested publicly available AMES[1] data set to evaluate model performance. Molecules are limited primary aromatic amine. All models are trained and tested as a binary classification model.
positive:456mols
negative:185 mols



Examples of molecules

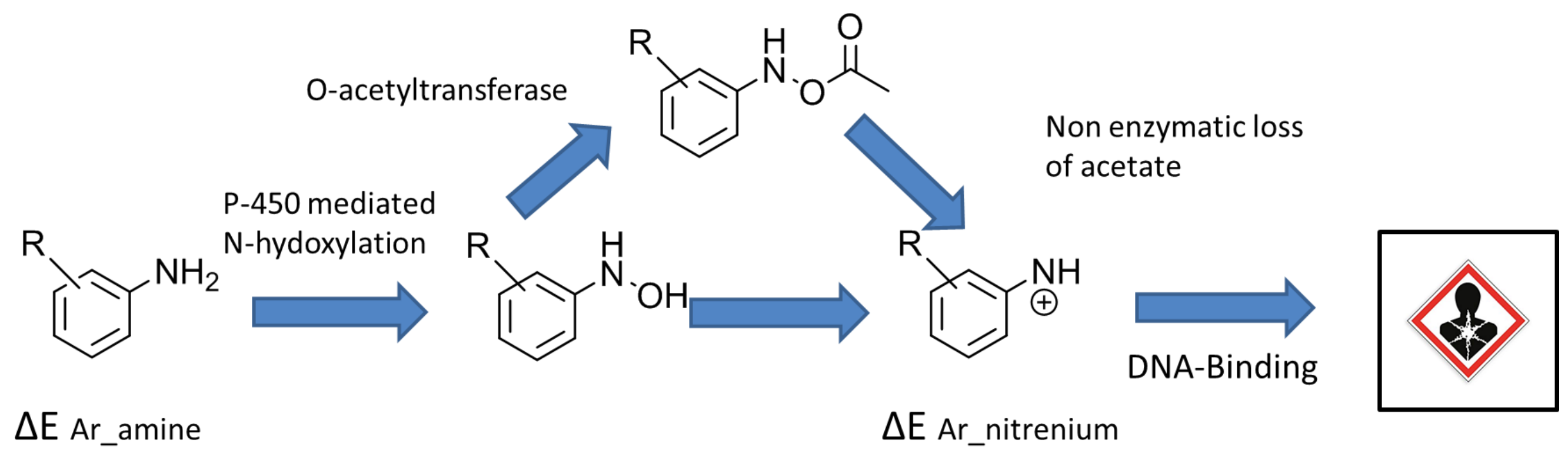
3. Quantum Chemical(QC) based approach

- Nitrenium ion stability correlates AMES toxicity.[2]
- Aryl amine and its nitrenium ion energy is calculated with Psi4 wrapper, psikit[3]. (Basis set SCF/6-31G**)

$$\Delta\Delta E = (\Delta E\mathit{Ar\_nitrenium} - \Delta E\mathit{Ar\_amine})$$
$$- (\Delta E\mathit{aniline\_nitrenium} - \Delta E\mathit{aniline})$$

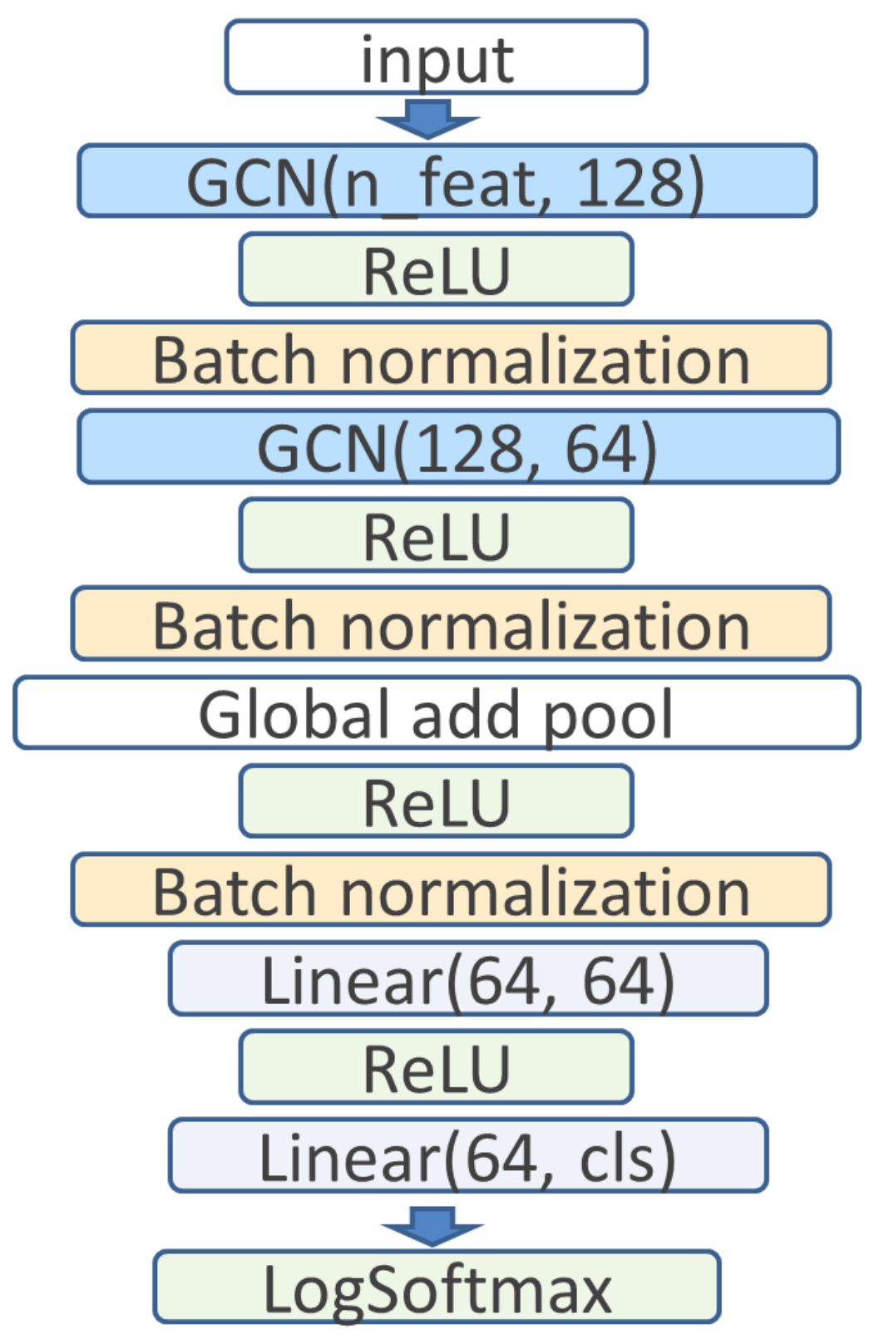
ΔΔE < 0 Ar amine => AMES negative

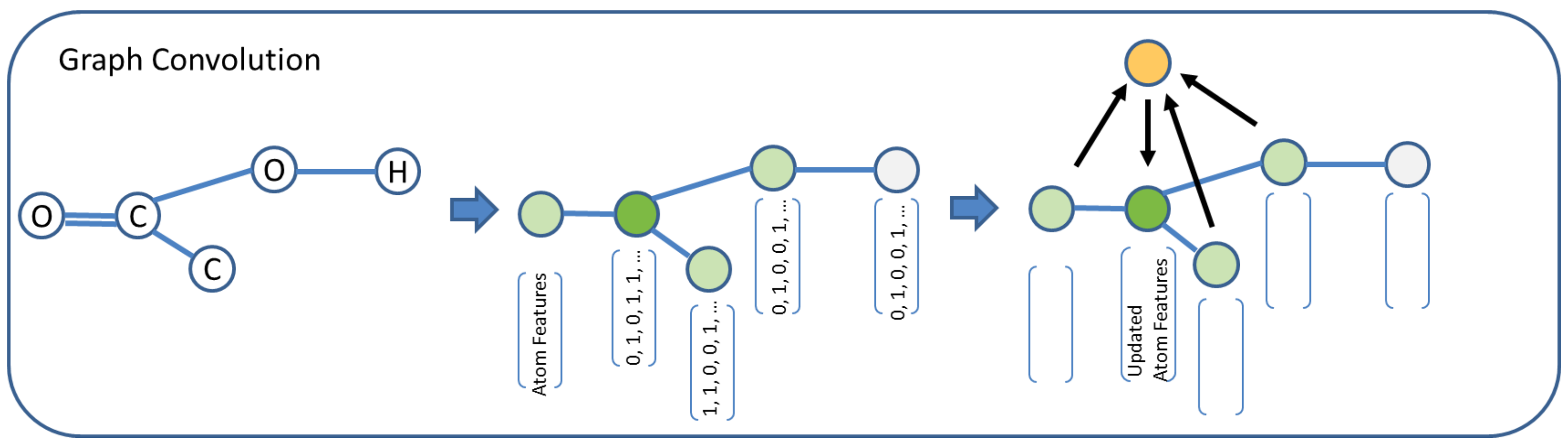
ΔΔE > 0 Ar amine => AMES positive



4. Graph Neural Network(GNN) based approach

- GNN models are easy to modify input features and outperform fingerprint based models.
- Build GNN model with 2 GCN layers.
- RESP charge is used as the new node feature because atomic partial charge used to predict reactivity. And investigated the effect of the input atom features.
- GNN network is implemented with pytorch_geometric[4].
- Support vector machine with ECFP4 is used for baseline.


input
GCN(n_feat, 128)
ReLU
Batch normalization
GCN(128, 64)
ReLU
Batch normalization
Global add pool
ReLU
Batch normalization
Linear(64, 64)
ReLU
Linear(64, cls)
LogSoftmax


Graph Convolution

Atom features used in this study

Feature	Description	Size
Atom type	Type of atom	44
Degree	Degree of atom	11
Valence	Valence of atom	7
Formal charge	Integer electronic charge of atom	1
Hybridization	sp, sp2, sp3, sp3d, sp3d2	5
Num of radical electron	Number of radial electrons	1
Aromatic	Aromatic atom or not aromatic	1
Num of Hs	Number of bonded hydrogen atoms	5
*RESP Charge	Partial charge of atoms	1

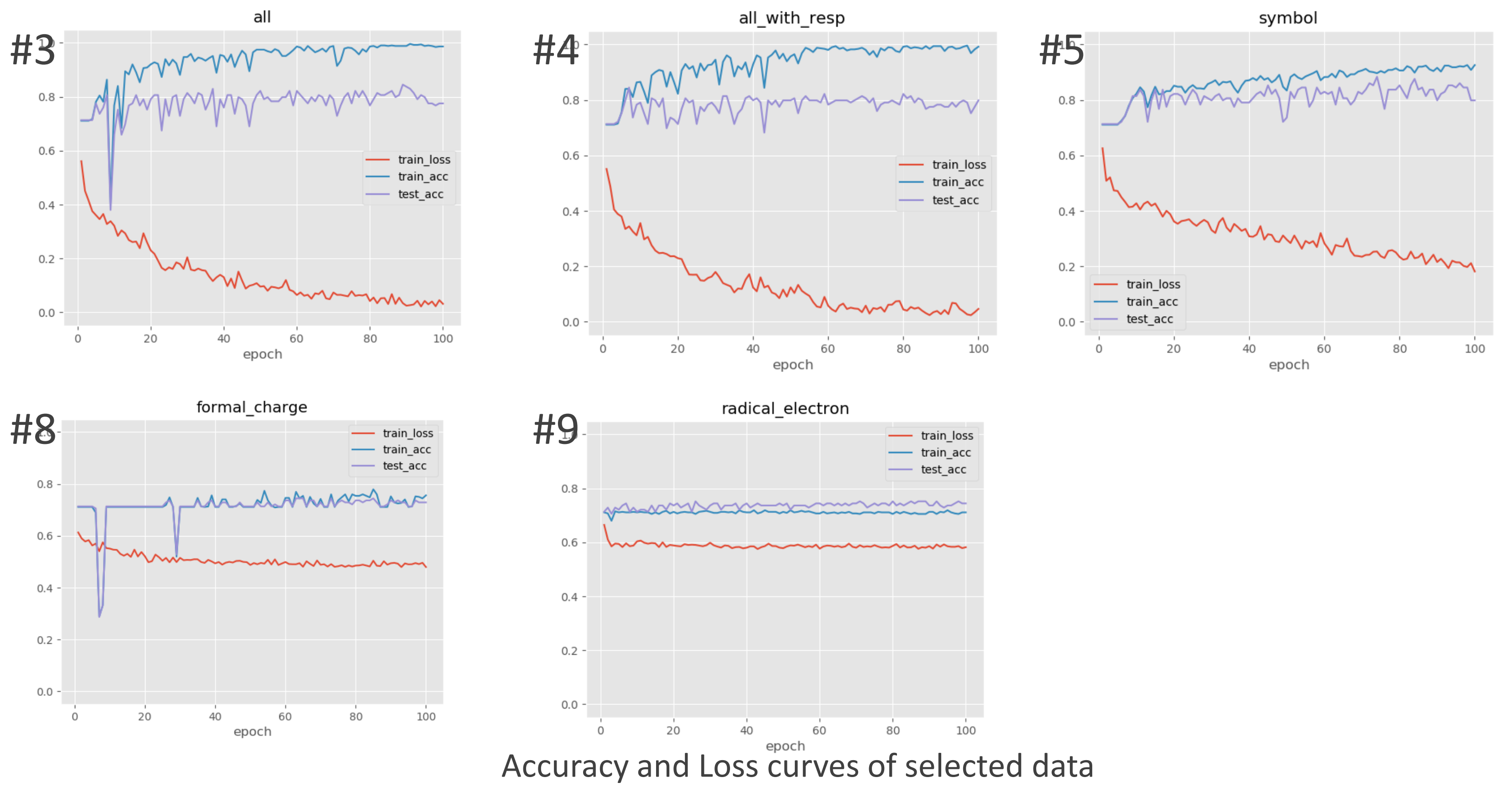
*RESP Charge was calculated with psikit, Basis set : SCF/6-31G**

5. Prediction Results of AMES dataset

#	Method	Used atom features	F1_SCORE	ROC_AUC
1	QC based	ΔΔE	0.74	-
2	SVC	ECFP4	0.85	-
3	GNN	All	0.84	0.75
4	GNN	All + RESP Charge	0.85	0.76
5	GNN	Atom Symbol	0.86	0.76
6	GNN	Atom Degree	0.80	0.66
7	GNN	Implicit Valence	0.82	0.71
8	GNN	Formal Charge	0.83	0.56
9	GNN	Num of Radical Electron	0.85	0.55
10	GNN	Hybridization	0.70	0.63
11	GNN	Aromatic atom or not	0.84	0.59
12	GNN	Num of bonded Hs	0.82	0.65

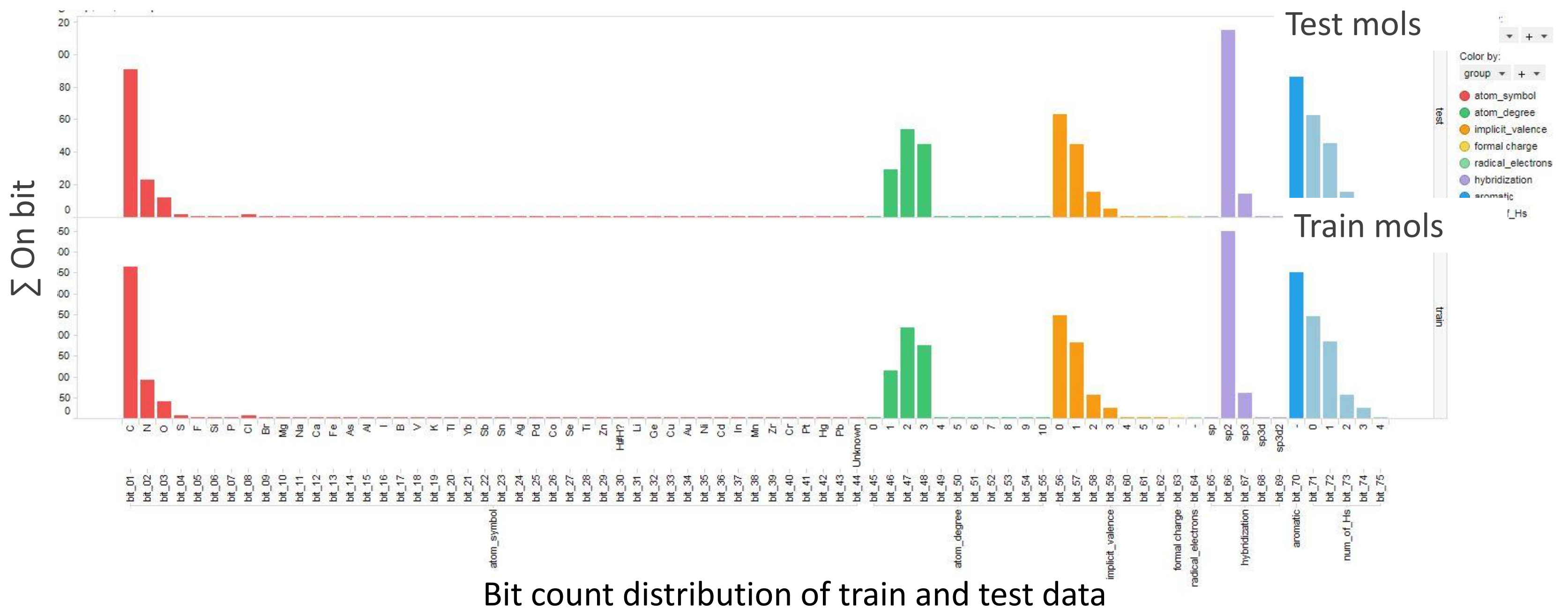
Addition of RESP Charge didn't improve the performance. Because current GCN doesn't consider the 3D conformation and convolves features with 2D adjacency matrix.

#8, 9, 11, 13 showed low ROC_AUC score. #3, #4 and #13 showed almost same performance. Selected accuracy and loss curves are shown below.



Accuracy and Loss curves of selected data

Count On-bit of all atom features in the dataset are shown. Most of atom symbol is 'Carbon'. But atom symbol was important features. It's indicate that topological information of the molecules is important feature. On the other hand, formal charge and num. radical electrons are zero for all molecules, so GNN could not learn useful information from these features.



Bit count distribution of train and test data

6. Summary

GNN and SVC showed almost same performance and outperformed QC approach in this study. Nitrenium hypothesis based QC approach is useful because it isn't required training data but limited available molecules (primary aromatic amine).

The effect of input feature for GNN is investigated. As new additional feature, RESP charge was used but it didn't improve the performance. On the other hand, the model trained with only atom symbol feature showed almost same performance to original model. These results indicate that current input features for GNN are redundant. Additional research is required for input features. We are challenging new input features development such as quantum chemistry based descriptors for GNN.

7. References

[1] *Katja H. et al., J. Chem. Inf. Model. 2009, 49, 2077–2081*

[2] *Jorg B. et al., J. Chem. Inf. Model. 2010, 50, 274–297*

[3] https://github.com/psi4/psi4, https://github.com/Mishima-syk/psikit

[4] https://pytorch-geometric.readthedocs.io/en/latest/