Electronic Supplementary Material (ESI) for New Journal of Chemistry.

# Supplementary materials for “Machine learning and graph neural network for finding potential drugs related to multiple myeloma”

Haohuai He[1#], Guanxing Chen[1#], Calvin Yu-Chian Chen[1,2,3*]

[1] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, 510275, China

[2] Department of Medical Research, China Medical University Hospital, Taichung 40447, Taiwan

[3] Department of Bioinformatics and Medical Engineering, Asia University, Taichung 41354, Taiwan

[#] The authors contribute equally

* Corresponding Authors

Calvin Yu-Chian Chen, Ph.D.

School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, 510275, China

TEL: 15626413023

E-mail: chenyuchian@mail.sysu.edu.cn

# 20% proportional split dataset test

We use 8:2 to divide the training set and test set, which has little change compared with the original division.

Table S1. the $R^2$ score and Std of train and test set on 8:2 split rate

| Model | Train $R^2$score | Test $R^2$score |
|---|---|---|
| Adaboost | 0.788±0.009 | 0.761±0.031 |
| Random Forest | 0.850±0.008 | 0.822±0.035 |
| ExtraTrees | 0.831±0.007 | 0.830±0.030 |
| XGBoost | 0.841±0.007 | 0.811±0.039 |
| SVR | 0.655±0.006 | 0.509±0.045 |
| Elastic Net | 0.675±0.007 | 0.710±0.025 |
| KNN | 0.862±0.007 | 0.694±0.018 |
| MLP | 0.493±0.237 | 0.475±0.256 |

# Molecular descriptors test

We use PaDEL[1], QuBiLS-MAS[2], and QuBiLS-MIDAS[3,4] as molecular descriptors to build QSAR models. the results of the four molecular descriptors are shown in table X when the same feature screening pipeline and ExtraTree model are used.

We find that the 204 molecular descriptors of GFA have the best results, so we still use GFA molecular descriptors to build the QSAR model.

Table S2. the $R^2$ score and Std of train and test set on different molecular descriptors

| Description | Train $R^2$score | Test $R^2$score |
|---|---|---|
| PaDEL | 0.860±0.006 | 0.749±0.034 |
| QuBiLS-MAS | 0.848±0.005 | 0.731±0.029 |
| QuBiLS-MIDAS | 0.784±0.007 | 0.721±0.025 |
| GFA | 0.839±0.009 | 0.833±0.049 |

## Data availability

The data we use is from the ChEMBL dataset (https://www.ebi.ac.uk/chembl/) and TCM datasets (https://tcm.sysu.edu.cn/).

We also added the 881 EZH2 related small molecules data obtained from the ChEMBL data set to the supplementary materials (“Supplementary materials-dataset.xlsx”).

# References

[1]Yap, Chun Wei. "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints." *Journal of computational chemistry* 32.7 (2011): 1466-1474.

[2]Valdés-Martiní, José R., et al. "QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5 D) algebraic molecular descriptors computations." *Journal of cheminformatics* 9.1 (2017): 1-26.

[3]García-Jacas, César R., et al. "QuBiLS-MIDAS: A parallel free-software for molecular descriptors computation based on multilinear algebraic maps." (2014): 1395-1409.

[4]García-Jacas, César R., et al. "Distributed and multicore QuBiLS-MIDAS software v2. 0: Computing chiral, fuzzy, weighted and truncated geometrical molecular descriptors based on tensor algebra." *Journal of computational chemistry* 41.12 (2020): 1209-1227.