**Investigation of chemical structure recognition by encoder-decoder models in learning progress**

Shumpei Nemoto,a Tadahaya Mizuno a† and Hiroyuki Kusuhara a

*a Graduate school of Pharmaceutical Sciences, Bunkyo-ku, 113-0033, Tokyo, Japan*

*† tadahaya@mol.f.u-tokyo.ac.jp*

****

**Supplementary Figure 1**. AUC and MCC of 113 assays prediction compared between perfect accuracy of Encoder-Decoder models. Bar height and error bar indicate mean and standard deviation, respectively.

**Supplementary Figure 2**. Coefficient of determination of Lipophilicity and FleeSolv prediction compared between perfect accuracy of Encoder-Decoder models. Lipophilicity and FleeSolv data were obtained from MoleculeNet (<https://moleculenet.org/>). XGBoost was used as machine learning algorithm for prediction. Hyperparameters listed in **Supplementary Table 1** were optimized using Optuna for each dataset prediction with optimization index of RMSE and n\_trials of 50.

**Supplementary Table 1**. Optimized hyperparameters of XGBoost by Optuna. Each hyperparameter takes values from min to max.

**Supplementary Table 2**. 113 HTS assays predicted by XGBoost from ToxCast. All assays have a sample size of at least 7000 (listed in Supplementary\_Table.xlsx).

**Supplementary Table 3**. List of tanimoto coefficients with similar compound groups and representative compounds (listed in Supplementary\_Table.xlsx).