A comparative research between atomic-based and chemical calculation-based descriptor for cocrystal screening machine learning model

Yingquan Hao, Ying-Chieh Hung, Yusuke Shimoyama

Tokyo Institute of Technology, Tokyo, Japan;

yshimo@chemeng.titech.ac.jp

Introduction: In the past decades, the solubility of the pharmaceuticals in human body has been become lower. To solve this problem, the cocrystal have been considerate. By forming a new crystal structure with an additive which is called as coformer (CF), the dissolution property of the active pharmaceutical ingredient (API) can be modified. But by now, the screening of the cocrystal is mostly carried by experiments. Even there are some approaches attempt to screening the API and CF pair for cocrystal formation by machine learning, but most of them use semi-empirical descriptor or atomic-based molecular structure descriptor to make the chemistry understandable to the computer. However, this make it is very hard to get a machine learning model which can be generalized to the cases which is not learned before. So, in this research, 3D convolution neural network (3D-CNN) is employed with 3D charge distribution calculated by Universal ForceField (UFF) and Gasteiger partial charge method (GPC) to achieve a general model. Also, the performance of new build model is compared with the atomic-based methodology used before such as Graph Convolutional Network (GCN), neural network with extensive connectivity fingerprint (ECFP-NN).

Method & Result: The experimental datasets from the literature is used for this research. UFF and GPC are applied to get the initial position and charge of the API and CF, then charge is mapped to a new 3D coordinate system which set the longest edge of the Oriented Bounding Box (OBB) of the molecular as the x-axis, the middle edge as y-axis, the shortest one as z-axis, and the centre of the OBB as the (0,0,0). Finally, the new mapped charge of APIs and CFs is transformed to 3D arrays for inputs to 3D-CNN. The hyperparameter of 3D-CNN is determined by training datasets with Bayesian-optimization with 5-flod cross-validation. 3D-CNN shows a training accuracy 80% and accuracy 71% with the test datasets that contain none of the molecular in training datasets, while the ECFP-NN and GCN only give test accuracy lower than 65%. Because the 3D charge information is directly link to the cocrystal formation between API and CF.

Keywords: cocrystal, machine-learning, molecular informatics