**SYBA: Fragment based prediction of hard-to-synthesize structures**

Voršilák M.1,2, Svozil D.1,2

*1 CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic.*

*2 CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic.*

*e-mail: milan.vorsilak@vscht.cz, daniel.svozil@vscht.cz*

Machine learning methods are often used in cheminformatics to predict activity, to cluster similar structures or to classify structures into distinctive classes. While a biological activity or toxicity can be experimentally measured, another important molecular property, the synthetic feasibility, is a more abstract feature that can’t be easily assessed. Furthermore, synthetic feasibility is not only abstract, but even hard-to-synthesize (HS) structures are not readily available from any database, which is needed for classification. Thus, we developed HS structure generator called Nonpher to address this issue [1].

Classification based on Bayes probability relies on the sum of feature contributions that are calculated as differences between feature probabilities in training datasets. As a feature space, we utilized fragments generated with Morgan algorithm. Random samples of the ZINC database were analyzed to obtain natural (easy-to-synthesize, ES) fragment scores, conversely, Nonpher generated dataset to obtain HS fragment scores and the combination of individual fragment scores assemblies SYBA classifier. Our model promises easy interpretation because ES and HS fragments are known and on the test set also high accuracy.

Literature:

[1] M. Voršilák, D. Svozil, J. Cheminform. 2017, 9:20.