[P2] Ranking of Reaction Condition Applicability Based of Artificial Networks

Valentina Afonina[1], Anastasiya Delova[1], Ramil Nugmanov[1], Timur Madzhidov[1], Olga Klimchuk[2], Alexandre Varnek[1,2]

[1] Chemoinformatics and Molecular Modeling Laboratory, Alexander Butlerov Institute of Chemistry, Kazan Federal University, Kremlyovskaya Str., 18, Kazan, Russia;
[2] Laboratory of Chemoinformatics, University of Strasbourg, Blaise Pascal Str., 1, Strasbourg, France.

The search for optimal reaction conditions is the most important task that arises during the development of the synthesis of the chemical compound of interest. At the same time, up to now there have not been developed tools that could be used to evaluate the conditions for conducting reactions on the basis of its structure. In this study we have tested the approach for predicting the optimal reaction conditions. This system takes into account that the reaction can be carried out in a variety of different conditions. Our approach is based on the use of machine learning methods, in particular, artificial neural networks.

Data

A set of catalytic hydrogenation reactions was taken from the Reaxys database. The initial dataset was standardized and information on temperature, pressure and catalysts were automatically curated. Reactions without information about temperature, pressure or catalyst were discarded. Thus, from the initial set of reactions containing 233905 reactions (corresponding to 116465 transformations), a standardized set of one step catalytic hydrogenation reactions was obtained, including all information on temperature, pressure, additives, and catalysts, comprising 38,739 reactions (corresponding to 30,986 transformations).

Modeling

The reactions of the obtained standardized data set were encoded as condensed graph of reaction (CGR) [1]. Then, binary fragment descriptors were generated using the ISIDA Fragmentor 2017 program [1]. Initial descriptor space containing more than 144 thousands descriptors was shrunk to 400 principal components (with explained variance 91,48%). Numerical values for temperature and pressure was binned into 3 bins (low, medium, high), catalysts and additives were encoded as indicator variables. For every reported condition of reaction a binary vector representing set of conditions (temperature, pressure, additive, catalyst) was formed in such a way. The total number of “permissible” sets of conditions is 4608.

The classification model based on the multilayer perceptron with one hidden layer that consists of 2000 neurons was built using shrunk fragment descriptors of transformation as input and vector of condition as output variables. Application of the model to a new reaction gives a vector of probabilities. On the basis of it “permissible” set of conditions was ranked. Top *k* values from this list were returned as predicted conditions.

Discussion

Cross-validation have shown that in 48.26% of cases the first ranked combination of conditions coincided with the really used reaction conditions, in 72.35% the real reaction conditions were found among the top 5 predictions. For the external set, the results were 50% and 74.06%, respectively. The zero model, in which the combinations of reaction conditions were ranked according to their frequency in the training set, gives much worse results – 32.03% and 47,36% for the training set and 30,88% and 46,82% for the test set, respectively.

Bibliography :

[1] Varnek A. et al. J. Comput. Aided. Mol. Des. 19 (2005) 693-703

The research was supported by Russian Science Foundation, grant 14-43-00024.