Deep conditional variational autoencoder for reaction conditions prediction

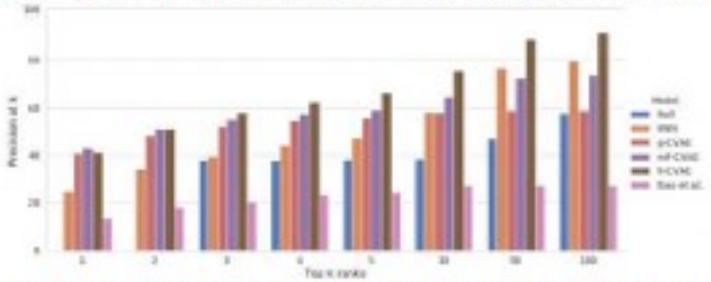
Timur I. Madzhidov[1], Daniyar A. Mazitov[1], Valentina A. Afonina[1], Tagir N. Akhmetshin[1], Juliya D. Skibina[1], Arkadii I. Lin[2], Ramil I. Nugmanov[1], Alexandre Varnek[2]

[1] Kazan Federal University
18 Kremlyovskaya Str., 420008 Kazan, Russia
[2] University of Strasbourg
1, rue Blaise Pascal, 67000 Strasbourg, France
E-mail: Timur.Madzhidov@kpfu.ru, Skype: tmadzhidov,
Twitter: tmadzhidov, Telegram: @tmadzhidov

Synthesis planning domain has a second birthday after seminal paper of Segler et al.[1] who proposed AI-powered technology for retrosynthetic route prediction. However, selection of optimal conditions for every reaction in synthetic route is almost as challenging and as important as prediction of the route itself. The main complication in prediction of optimal conditions are absence of negative examples, very large condition space, voids in reaction-condition matrix (all possible conditions for given reaction are unavailable). For some reactions, conditions are well-studied and documented, but for the other selection of conditions could be tricky and requires good knowledge and experience of chemists.

Here, we propose an approach based on deep neural networks that predicts combination of catalyst-reagent-temperature-pressure best suited for a particular reaction. Unlike existing approaches[2,3], here we do not rank possible conditions based on some technique but directly sample them using conditional variational autoencoder. Training and test sets comprised of hydrogenation reactions from Reaxys database, two datasets are collected: "small" consisting from 38K with limited set of conditions and "big" one including 196K reactions with vast variety of conditions. Three different latent distributions were compared, namely Gaussian (g-CVAE), Riemannian Normalizing Flow (mf-CVAE) and Hyperspherical Uniform (h-CVAE) distributions.

Proposed approaches were shown superior performance to null model, which predicts conditions based on their popularity in training data, as well as over nearest-neighbor approach (KNN). The latter ranks possible conditions based on the similarity of corresponding reactions to a given one. For hydrogenation reactions, previously proposed approach of Gao et al.[2] have shown mediocre performance (see figure).



The proposed network architecture is the first one that can generate possible set of applicable conditions for a particular reaction. It was shown that such sampling generates set of possible conditions even for very wide reaction condition space (see figure).

Research was partially supported by the Ministry of Education of Youth and Sports of the Czech Republic, agreement MSMT-5727/2018-2, as well as the Ministry of Higher Education and Science of the Russian Federation, agreement 14.587.21.0049 (unique project identifier RFMEFI58718X0049).

1. Segler, M. H. S. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604 (2018).
2. Gao, H. *et al.* Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* 4, 1465–1476 (2018).
3. Lin, A. I. *et al.* Automatized Assessment of Protective Group Reactivity: A Step Toward Big Reaction Data Analysis. *J. Chem. Inf. Model.* 56, 2140–2148 (2016).