[P8] Modeling of complex reaction data: the case of tautomeric equilibria

Glavatskikh Marta [1,2], Timur Madzhidov [2], Dragos Horvatz [1] Ramil Nugmanov [2], Timur Gimadiev [1,2], Igor Baskin [3], Gilles Marcou [2], Alexandre Varnek [2]

[1]Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France.
[2]Laboratory of Chemoinformatics and Molecular Modeling, Butlerov Institute of Chemistry, Kazan Federal University, Kremlevskaya str. 18, Kazan, Russia
[3]Lomonosov Moscow State University, Leninskiye Gory str. 1, Moscow, Russia

The existing tools for the prediction of ratio of tautomers are predominantly based on the calculation of pKa values of related tautomer. This may significantly affect the accuracy, especially, if the errors of the pKa predictions are comparable with the difference of tautomers' pKa values. Moreover, this calculation is usually restricted by aqueous solution and hence not applicable for other media. Here the prediction of tautomeric equilibria is performed directly for the equilibrium constant ($\log K_T$) for the reactions proceeding in aqueous and organic solutions or their mixtures.

The models were built on a data set of 695 reactions of 10 tautomeric classes, for which $\log K_T$ values were measured in different solvents and at different temperatures[1]. Support Vector Machine[2] (SVM) and Generative Topographic Mapping[3] (GTM) were used as machine learning methods. The structure of tautomers has been encoded by ISIDA fragments[4] whereas conditions were accounted for physico-chemical parameters of solvent and inverse temperature. Both SVM and GTM models perform well in cross-validation (RMSE (5CV)=0.63-0.67, R^2 (5CV)=0.82-0.84). Validation of these models on two external test sets, either included the transformations under new reaction conditions (test 1) or new structures (test 2), lead to reasonable statistical parameters (RMSE=0.59 and 1.96, R^2=0.62 and 0.65). Large RMSE value for test 2 is explained by the fact that more than half of the compounds were out of the model's applicability domain. The consensus SVM model is publicly available on our web-server: https://cimm.kpfu.ru/development/predictor.
As it is illustrated on Figure 1, a GTM map well separates both different tautomeric classes and the same equilibria proceeding in different solvents.

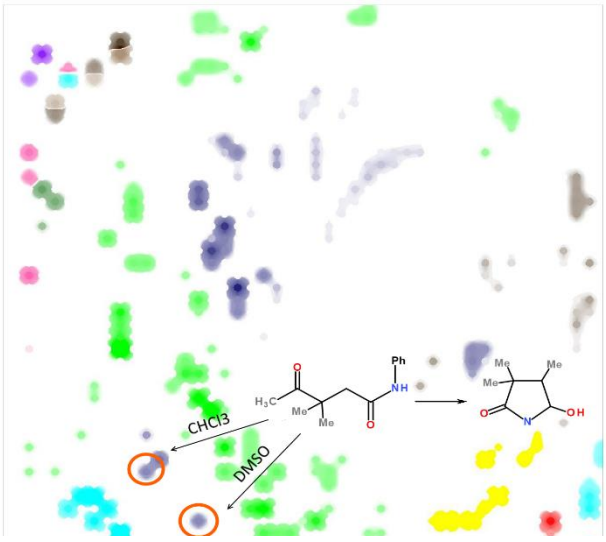

Figure 1. GTM map built for 697 tautomeric equilibria. The color code corresponds to 10 tautomeric classes. Selected data points correspond to the same equilibrium studied in CHCl3 ($\log K_T$ = -0.49) and DMSO ($\log K_T$ = 0.62).

Bibliography:

[1] Palm, V. A. ; VINITI: Moscow (1978).
[2] Chang, C.-C.; Lin, C.-J. Intell. Syst. Technol. (2011) 2 (3), 1-27.
[3] Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. J. Chem Inf. Model. (2013) 53 (12), 3318-3325.
[4] Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. Curr. Comput.-Aided Drug Des. (2008) 4 (3), 191-198.