Electronic supplementary materials *Mendeleev Commun.*, 2021, **31**, 769-780

Machine learning modelling of chemical reaction characteristics: yesterday, today, tomorrow

Timur I. Madzhidov, Assima Rakhimbekova, Valentina A. Afonina, Timur R. Gimadiev, Ravil N. Mukhametgaleev, Ramil I. Nugmanov, Igor I. Baskin and Alexandre Varnek

Annex A1. Benchmarking of reaction descriptors

Reaction-center (RC) free, RC-implicit, and RC-explicit descriptors were benchmarked on the four datasets of balanced reactions annotated with rate constants of S_N2 [1], E2 [2], Diels-Alder [3], or equilibrium constants of tautomerization reactions [4].

Representation of chemical reactions

RC-free. In the RC-free methodology, descriptors of reaction are molecular descriptors of reactants and products which considering molecular descriptors of reactants or products or their concatenation. Therefore, we considered descriptors of reaction as (i) ISIDA fragment descriptors set of reactants; (ii) ISIDA fragment descriptors set of products; (iii) concatenation of ISIDA fragment descriptors sets of reactants and products. ISIDA fragment descriptors calculation procedure is described below.

*RC-implicit***.** RC-implicit descriptors consider reaction descriptors as subtraction of products and reactants descriptors. The descriptors were (i) ISIDA fragment descriptors and (ii) RDkit fingerprints. ISIDA fragment descriptors and fingerprints calculation is described below.

RC-explicit descriptors. The chemical transformations in RC-explicit descriptors were encoded by Condensed Graph of Reaction (CGR). In CGR approach, a reaction is represented by a single 2D graph, some sort of pseudomolecule that contains both conventional chemical bonds and so-called dynamic bonds characterizing changed/broken/formed chemical bonds [6,7]. Thus, CGR represents the whole transformation, i.e., reactant-product pair, as a single molecular graph. CGRtools library was used to generate CGRs [8]. ISIDA fragment descriptors calculation is computed for CGR.

Descriptors

ISIDA fragment descriptors. ISIDA fragment descriptors were computed using the ISIDA Fragmentor [9] program. They represent the subgraphs of different topologies and sizes. Each subgraph is considered as a descriptor type whereas its occurrence in a molecule is the descriptor value. In this study, two types of subgraphs were considered: sequences of atoms and/or bonds and augmented atoms (atoms with first, second, etc. coordination spheres). The length of monitored fragments varied from 2 to 14 for sequences and from 2 to 6 for atom-centered fragments. The following options were also used: charges on atoms (Formal Charge), accounting for the terminal atoms of a fragment exclusively (Atom Pairs) and exploring all possible paths instead of shortest paths (DoAllWays).

An important option regulating the amount of the overall generated CGR fragments is the ‘dynamic bond’ inclusion. Toggled on, the option produces the fragments, that contains the bonds forming/breaking while chemical reaction (local fragments) and omits the ‘generic’ fragments, not assigned to the reaction centre. That could be used to generate fragments that describe the local environment of the reaction centre exclusively. Therefore, for each fragmentation type for CGR, the descriptors vector included either all generated fragments or only fragments containing dynamic bonds or atoms.

Difference reaction fingerprint. A reaction fingerprint is a difference between count-based fingerprints of products and reactants. In our study, we used three types of reaction fingerprints developed by Schneider et al. [15] and implemented in RDKit software [33]: (i) atom pairs representing two particular atoms with the specified number of non-hydrogen neighbor atoms separated by up to three bonds [34], (ii) Morgan fingerprints identical to extended-connectivity fingerprints with radius 2 or 3 and vector sizes were 512, 1024, 2048 [35] and (iii) topological torsions representing four consecutively linked non-hydrogen atoms with the specified number of π-electrons and the number of non-hydrogen neighbor atoms [36].

Descriptors of the reaction conditions. Descriptor vectors were complemented with experimental conditions (solvent parameters and temperature) as described in reference [10]. Each solvent was described by 15 descriptors that represent polarity, polarizability, H-acidity, and basicity: Catalan SPP [11], SA [12], and SB constants [11], Camlet–Taft constants α [13], β [14], and π^* [15], four functions depending on the dielectric constant, three functions depending on the refractive index as shown in paper [16]. The latter 7 descriptors reflect the polarity and polarizability of the bulk of the solvent. The inverse absolute temperature, 1/T (in Kelvin degrees) was also used as a descriptor of temperature influence. Since some of the solvents were a water-organic mixture, the molar ratio of organic solvent was used as a descriptor as well (100% for pure solvent).

Here, we compare models obtained on all descriptor sets generated using mentioned possible options of descriptor calculation, and results are provided as boxplots. In practice, usually optimal fragmentation scheme is selected based on cross-validation (which corresponds to the best model in boxplot).

Model building and validation

The models were built using Random Forest approach implemented in the scikit-learn library [17]. The number of trees was equal to 500 in all cases, the optimized hyperparameter was the values of features selected upon tree branching (option max_features). The other parameters were set to their default values. The predictive performance of the best models was estimated using the nested cross-validation technique. In this approach, an outer (external validation) loop split the initial data set into an external test set (used for assessment of the model performance) and modelling set (used for the model building including internal cross-validation).

Here, two strategies of such split have been tested: conventional random 5-fold cross-validation (denoted as ‘reaction-out’ CV) and ‘transformation-out’ CV [18]. ‘Reaction-out’ CV approach was simply a regular five times repeated five-fold CV. “Transformation-out” CV estimates the ability to predict characteristics of chemical reactions with a novel reactant-product pair.

A hyperparameter (the number of features to consider when looking for the best split) of the Random Forest was optimized on modelling set in the conventional 5-fold cross-validation using grid search. Its best value found in internal cross-validation was used to build a model on the whole modelling set, followed by the application of resulting models to the external test set. Predictions for the objects in the external test set folds were merged and then performance metrics were calculated. Notice, that performance on ‘reaction-out’ CV and ‘transformation-out’ CV reflects predictive ability on the external test set.

Predictive performance was evaluated using determination coefficient (R^2):

$$R^2 = 1 - \frac{\sum_i (y_{i,pred} - y_{i,obs})^2}{\sum_i (y_{i,pred} - \bar{\bar{y}}_{obs})^2}$$

here $y_{i,obs}$ and $y_{i,pred}$ are, respectively, experimental and predicted values of the rate constant log$k$, or equilibrium constants of tautomerization reactions logk_T, N is the number of data points.

References

1. T. Gimadiev, T. Madzhidov, I. Tetko, R. Nugmanov, Iu. Casciuc, O. Klimchuk, A. Bodrov, P. Polishchuk, I. Antipin and A. Varnek, *Mol. Inf.*, 2018, **38**, 1800104.

2. T. I. Madzhidov, A. V. Bodrov, T. R. Gimadiev, R. I. Nugmanov, I. S. Antipin and A. A. Varnek, *J. Struct. Chem.*, 2015, **56**, 1227.
3. T. I. Madzhidov, T. R. Gimadiev, D. A. Malakhova, R. I. Nugmanov, I. I. Baskin, I. S. Antipin and A. A. Varnek, *J. Struct. Chem.*, 2017, **58**, 650.
4. T. R. Gimadiev, T. I. Madzhidov, R. I. Nugmanov, I. I. Baskin, I. S. Antipin and A. Varnek, *J. Comput.-Aided Mol. Des.*, 2018, **32**, 401.
5. T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, A. V. Bodrov, A. I. Lin, I. I. Baskin, A. A. Varnek and I. S. Antipin, *Russ. J. Org. Chem.*, 2014, **50**, 459 (*Zh. Org. Khim.*, 2014, **50**, 473).
6. A. Varnek, D. Fourches, F. Hoonakker and V. P. Solov'ev, *J. Comput.-Aided Mol. Des.*, 2005, **19**, 693.
7. F. Hoonakker, N. Lachiche, A. Varnek and A. Wagner, *Int. J. Artif. Intell. Tools*, 2011, **20**, 253.
8. R. I. Nugmanov, R. I. Nugmanov, R. N. Mukhametgaleev, T. Akhmetshin, T. R. Gimadiev, V. A. Afonina, T. I. Madzhidov and A. Varnek, *J. Chem. Inf. Model.*, 2019, **59**, 2516.
9. A. Varnek, D. Fourches, D. Horvath and O. Klimchuk, *Curr. Comput.-Aided Drug Des.*, 2008, **4**, 191.
10. R. I. Nugmanov, T. I. Madzhidov, I. S. Antipin and A. A. Varnek, *J. Struct. Chem.*, 2014, **55**, 1026.
11. J. Catalán, V. López, P. Pérez, R. Martin-Villamil, J.-G. Rodríguez, *Liebigs Ann.*, 1995, 241.
12. J. Catalán and C. Díaz, *Liebigs Ann.*, 1997, 1941.
13. R. W. Taft and M. J. Kamlet, *J. Am. Chem. Soc.*, 1976, **98**, 2886.
14. M. J. Kamlet and R.W. Taft, *J. Am. Chem. Soc.*, 1976, **98**, 377.
15. M. J. Kamlet, J. L. Abboud and R. W. Taft, *J. Am. Chem. Soc.*, 1977, **99**, 6027.
16. T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, A. V. Bodrov, A. I. Lin, I. I. Baskin, A. A. Varnek and I. S. Antipin, *Russ. J. Org. Chem.*, 2014, **50**, 459 (*Zh. Org. Khim.*, 2014, **50**, 473).
17. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825.
18. A. Rakhimbekova, T. N. Akhmetshin, G. I. Minibaeva, R. I. Nugmanov, T. R. Gimadiev, T. I. Madzhidov, I. I. Baskin and A. Varnek, *SAR QSAR Environ. Res.*, 2021, **32**, 207.