Structure–reactivity modeling using mixture-based representation of chemical reactions

Pavel Polishchuk[1,2,3] · Timur Madzhidov[3] · Timur Gimadiev[3,5] · Andrey Bodrov[3,4] · Ramil Nugmanov[3] · Alexandre Varnek[3,5]

Received: 19 September 2016 / Accepted: 23 July 2017 / Published online: 27 July 2017

Abstract We describe a novel approach of reaction representation as a combination of two mixtures: a mixture of reactants and a mixture of products. In turn, each mixture can be encoded using an earlier reported approach involving simplex descriptors (SiRMS). The feature vector representing these two mixtures results from either concatenated product and reactant descriptors or the difference between descriptors of products and reactants. This reaction representation doesn't need an explicit labeling of a reaction center. The rigorous "product-out" cross-validation (CV) strategy has been suggested. Unlike the naïve "reaction-out" CV approach based on a random selection of items, the proposed one provides with more realistic estimation of prediction accuracy for reactions resulting in novel products. The new methodology has been applied to model rate constants of E2 reactions. It has been demonstrated that the use of the fragment control domain applicability approach significantly increases prediction accuracy of the models. The models obtained with new "mixture" approach performed better than those required either explicit (Condensed Graph of Reaction) or implicit (reaction fingerprints) reaction center labeling.

Keywords Chemical reactions · Simplex representation of molecular structure · Condensed graph of reaction · Reaction fingerprints · Rate constant prediction · Mixtures

Introduction

Structure–property modeling of chemical reactions represents a difficult task because of the complexity issue: any chemical reaction involves several molecular species of two types—reactants and products. The major question concerns the preparation of a descriptor vector encoding a chemical reaction which can serve as an input to a modeling software. Earlier, two different methodologies have been used for this purpose. The first one is based on the explicit consideration of a reaction center identified either manually or automatically using atom-to-atom mapping procedure [1]. This approach has been used in most of reported QSPR studies of reactions. Thus, Gasteiger et al. used some physicochemical parameters (charges, polarizabilities, steric accessibilities, parameters for inductive and resonance effects) for selected atoms and bonds to prepare the models for pK_a for aliphatic carboxylic acids [2] and for kinetics of amide hydrolysis [3]. ISIDA fragment descriptors [4, 5] issued from Condensed Graph of

Electronic supplementary material The online version of this article (doi:10.1007/s10822-017-0044-3) contains supplementary material, which is available to authorized users.

✉ Pavel Polishchuk
pavlo.polishchuk@upol.cz

✉ Timur Madzhidov
timur.madzhidov@kpfu.ru

✉ Alexandre Varnek
varnek@unistra.fr

[1] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czech Republic

[2] A.V. Bogatsky Physico-Chemical Institute of National Academy of Sciences of Ukraine, Odessa, Ukraine

[3] A.M. Butlerov Institute of Chemistry, Kazan Federal University, Kazan, Russia

[4] Department of General and Organic Chemistry, Kazan State Medical University, Kazan, Russia

[5] Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France

Reaction [6, 7] have been used for the reaction data analysis [8] and for the modeling of the rate of S_N2 [7, 9, 10] and E2 [11] reactions and optimal conditions for Michael reaction [12].

Another approach is based on the implicit representation of a reaction center, in which the feature vector for the reaction is calculated as the difference between descriptors of products and reactants [13–16] or by using only combined descriptors of substrates [17]. This methodology has been successfully applied in different reaction classification tasks [15, 18] and in building the regression model for prediction of optimal conditions of Michael reaction [12], S_N2 rate constant prediction [17] and and S_N1/S_N2 reactions classification [19]. Both approaches—with and without reaction center detection—have their own drawbacks. Unless detected manually for small congeneric data set, the reaction center detection needs atom-to-atom mapping procedure which is error-prone and time-consuming [20]. Calculation of reaction vectors [21] or reaction fingerprints [15, 18] requires perfectly balanced reactions; otherwise the resulting feature vector would contain chemically meaningless terms. Since most of raw reaction data in the widely used databases like CAS REACT or Reaxys are not balanced, the data curation step is needed before using modeling methods. However, application of e-notebooks for new chemical reaction registration in synthetic laboratories might potentially be helpful to feed the databases with perfectly balanced reactions.

In this article we describe an approach which doesn't need explicit encoding of a reaction center. A reaction is considered as an ensemble of two mixtures—a mixture of reactants and a mixture of products. Each mixture can be represented by special descriptors. Two different reaction representations were investigated: (*i*) concatenated feature vectors of reactants and products mixtures and (*ii*) a difference between these two vectors.

Earlier, we described an approach to prepare feature vectors for binary mixtures involving SiRMS descriptors [22]. Here, we extended this technique to mixtures having an arbitrary number of components. This new "mixture-like" methodology has been applied to model the rate constant of E2 reactions. For the comparison purpose, the models have also been built using either reaction fingerprints [15] issued from the implicit encoding of a reaction center or fragment ISIDA descriptors [4, 5] generated from the condensed graphs of reactions [6, 7] which explicitly label a reaction center. A rigorous cross-validation strategy has been suggested in order to provide with a realistic assessment of the models' performance.

Computational procedure

Dataset

A dataset of 313 E2 bimolecular elimination reactions carried out in pure solvents at different temperatures has been collected from the literature [23]. An E2 reaction proceeds in a single step with a single transition state. It results in a formation of a π-bond due to synchronous *trans*-elimination of a leaving group (**L**) in the presence of a base (**B**$^-$) needed to tie in the hydrogen atom (Fig. 1).

The dataset involves 90 distinct substrates and 60 distinct products, the most representative of them are listed in Table 1. The most representative substrates are ((1,1′-biphenyl)-4-yl)(1-bromocyclohexyl)methanone, (2-chloroethanesulfonyl)benzene and 2-fluorohexane, whereas the products are ethenesulfonylbenzene,

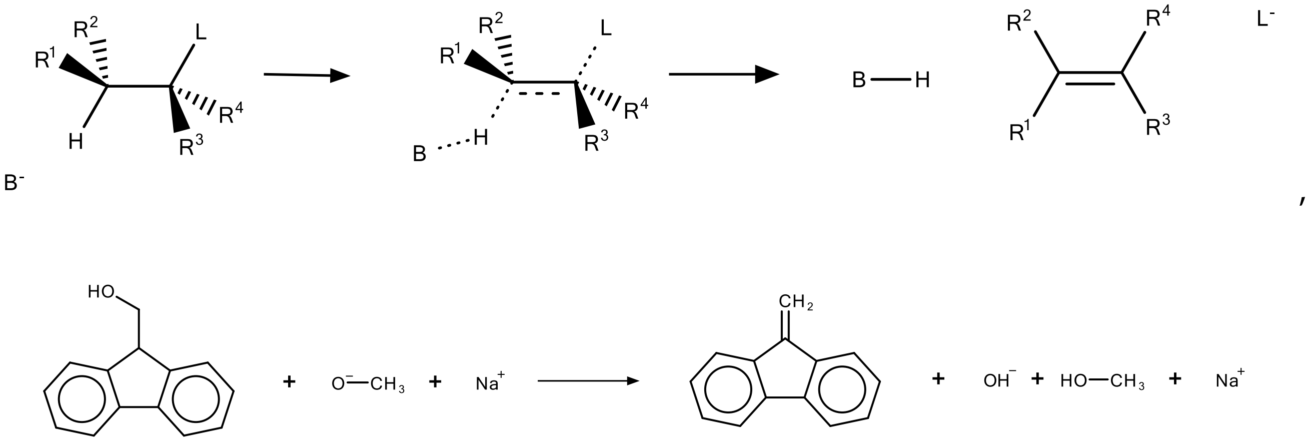
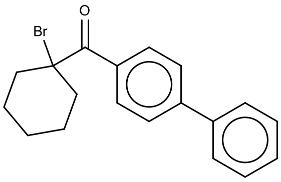
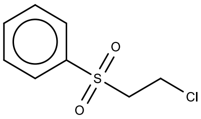
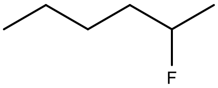
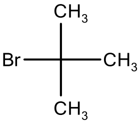
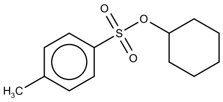
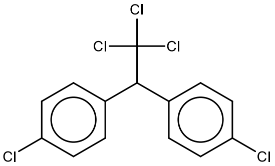
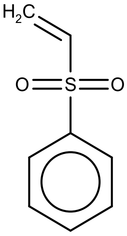
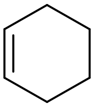
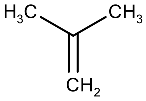
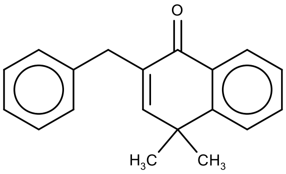
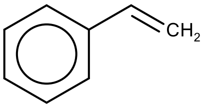
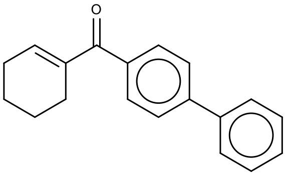
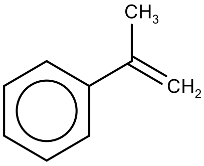
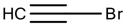
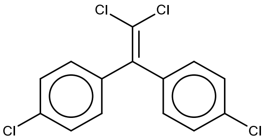


Fig. 1 A bimolecular elimination reaction. (*top*) Schematic representation of the E2 reaction mechanism, where B$^-$ is a base and L is a leaving group. (*bottom*) An example of an E2 transformation of (9*H*-fluoren-9-yl)methanol into 9-methylene-9*H*-fluoren, where CH_3O^- (from sodium methylate) is a base and hydroxide ion is a leaving group

Table 1 The most frequently occurred substrates and products

Reactants		
 13	 13	 12
 11	 10	 10
Products		
 27	 25	 25
 15	 14	 13
 11	 10	 10

cyclohexene and *iso*-butylene. Among the most representative leaving groups one can mention bromide and chloride anions occurred in 101 and 93 reactions, respectively, as well as *p*-tosylate and trimethylamine which occurred in 35 reactions each. The other seven leaving groups are occurred in very few reactions. Overall, 23 bases were detected, the most representative of them were methoxide occurred in 59 reactions, ethoxide (in 38 reactions), *tert*-bytoxide (30), thiophenyl (30), triethylamine (24), bromide (20), chloride (14) and hydroxide (14) ions and piperidine (10).

Representation of chemical reactions

The structures of reactants and products were encoded in reaction feature vectors using three different approaches: (*i*) the extended SiRMS mixture representation approach, (*ii*) ISIDA fragments calculated from condensed graphs of reactions and (*iii*) reaction fingerprints. Dipole moment, refraction, dielectric permittivity, Catalan acidity [24], basicity [25] and polarity/polarizability [26], Kamlet-Taft alpha [27], beta [28] and π constants [29] used as solvent parameters and reaction temperature were concatenated with all reaction feature vectors.

SiRMS-based mixture representation of chemical reactions

In the framework of the SiRMS methodology, a single compound can be represented as a set of tetraatomic fragments (simplexes) of fixed composition and topology (Fig. 2). The counts of identical simplexes are used as descriptor values.

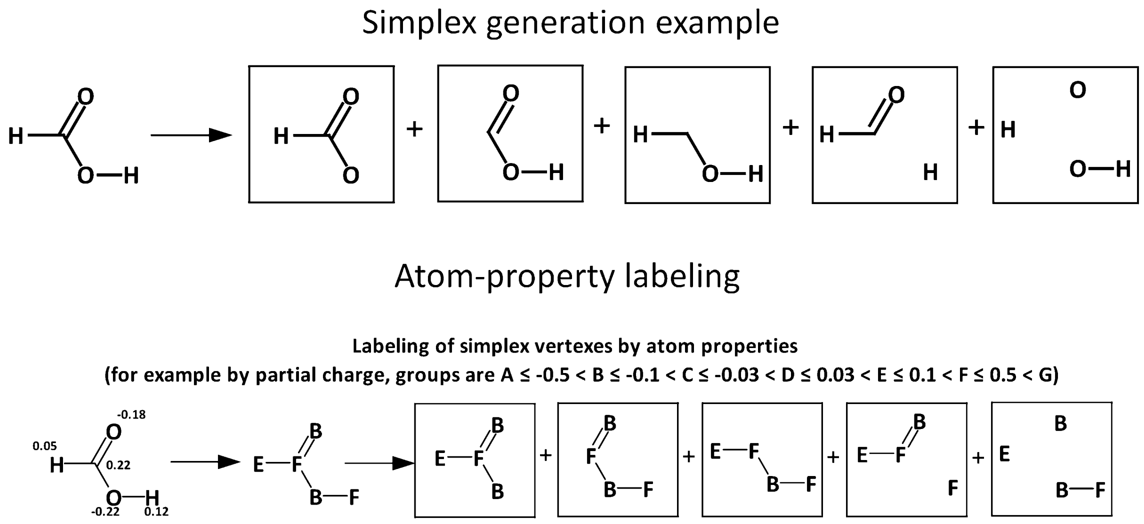

Simplex generation example
H
O
O—H
+
Atom-property labeling
Labeling of simplex vertexes by atom properties
(for example by partial charge, groups are A ≤ -0.5 < B ≤ -0.1 < C ≤ -0.03 < D ≤ 0.03 < E ≤ 0.1 < F ≤ 0.5 < G)
-0.18
0.05
0.22
-0.22
0.12
E
F
B

Fig. 2 An example of simplex descriptor generation for individual compounds

Generated simplexes can also be labeled according to different atomic properties (partial atomic charges, lipophilicity, H-bond donor/acceptor, etc). Partial atomic charges seem to be a relevant parameter for the reactivity modeling. Therefore, Gasteiger charges on atoms were calculated by *cxcalc* tool [30]. Then, the whole range of charge values was split onto seven bins labeled from A to G: $A \leq -0.5 < B \leq -0.1 < C \leq -0.03 < D \leq 0.03 < E \leq 0.1 < F \leq 0.5 < G$. In such a way, each atom received the corresponding label further used for simplex encoding (see Fig. 2). In order to avoid a combinatorial explosion, we enumerated either fully connected fragments (similar to the first three simplexes in Fig. 2) or fragments containing two disconnected parts (similar to the 4th simplex in Fig. 2). For more details about the SiRMS approach see our earlier studies [31, 32]. Notice that in this work we considered simplexes for which the numbers of atoms in fragments varied from 2 to 6.

The preparation of mixture descriptors for the mixture of three equally occurred components (here, reactants of E2 reactions) is illustrated in Fig. 3. It proceeds in three steps:

I. simplex descriptors representing connected or disconnected molecular subgraphs of N atoms (in this study $N=2$–6) are generated. For the mixture of three components A, B and C, the program generates simplexes of individual species including atoms of only A and B, as well as mixture simplexes including atoms of two (AB, BC, AC) or three (ABC) components. For molecular species containing less than 2 atoms (e.g., component C), individual simplexes are not generated. Each type of fragments is considered as an individual descriptor and its count weighted by the corresponding component occurrences is the descriptor value. In this study occurrences of all components were 1.

II. the feature vectors of individual simplexes are summed up which results in vector $D_S = A + B + C$. Similarly, superposition of the vectors of mixture simplexes AB, BC, AC and ABC results in D_M vector.

III. concatenation of D_S and D_M results in *SiRMS-mix*—the feature vector of the whole mixture.

Since a chemical reaction can be represented as an ensemble of two mixtures: a mixture of starting materials (reactants) and a mixture of products, the reaction feature vector can be computed as their combination. Two different ways of combining mixture feature vectors into reaction feature vector have been investigated: (*i*) their concatenation and (*ii*) by calculation of the difference between product and reactant mixture descriptors (Fig. 4).

In this study, simplexes included from 2 to 6 atoms; only pair-wise and triple-wise combinations of components were used for mixture simplex generation. The atoms were labeled either by symbols of chemical elements or by bin labels corresponding to partial atomic charges (see above).

Condensed graph of reaction

A Condensed Graph of Reaction (CGR) results from merging molecular graphs of reactants and products into one single connected or disconnected molecular graph described by conventional bonds (single, double, aromatic, etc) and dynamic bonds characterizing chemical transformations (single-to-double, double-to-single, etc) [6], see example in Fig. 5. In CGR, the changes of atomic charges in a course of a reaction can be accounted by introducing dynamic atoms (Fig. 5). A CGR can be prepared by superposing identically numbered atoms of reactants and products which needs to perform atom-to-atom mapping as a preliminary step.

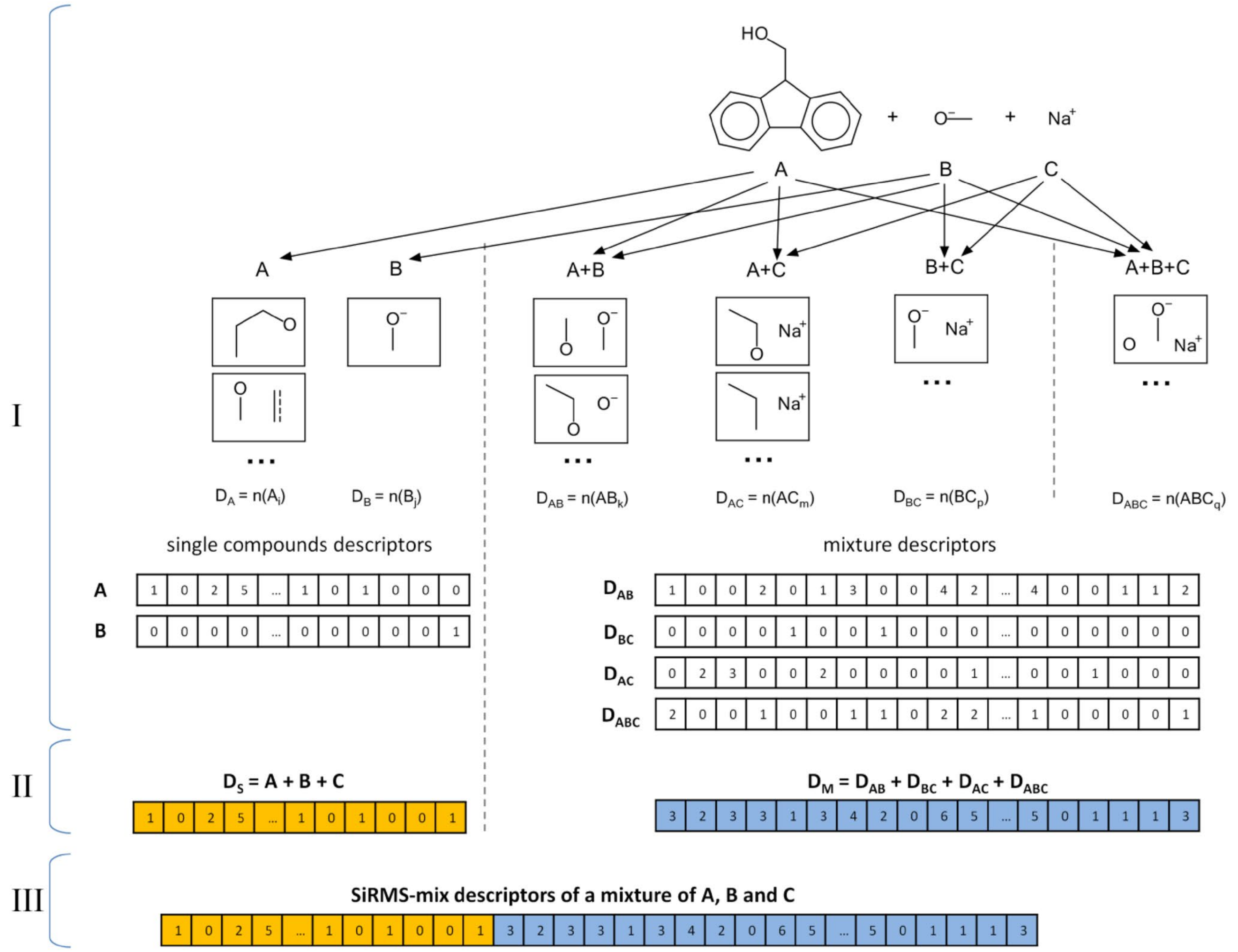

I
HO
+ O⁻ + Na⁺
A
B
C
A
B
A+B
A+C
B+C
A+B+C
O
O⁻
Na⁺
...
D_A = n(A_i)
D_B = n(B_j)
D_AB = n(AB_k)
D_AC = n(AC_m)
D_BC = n(BC_p)
D_ABC = n(ABC_q)
single compounds descriptors
mixture descriptors
A 1 0 2 5 ... 1 0 1 0 0 0
B 0 0 0 0 ... 0 0 0 0 0 1
D_AB 1 0 0 2 0 1 3 0 0 4 2 ... 4 0 0 1 1 2
D_BC 0 0 0 0 1 0 0 1 0 0 0 ... 0 0 0 0 0 0
D_AC 0 2 3 0 0 2 0 0 0 0 1 ... 0 0 1 0 0 0
D_ABC 2 0 0 1 0 0 1 1 0 2 2 ... 1 0 0 0 0 1
II
D_S = A + B + C
1 0 2 5 ... 1 0 1 0 0 1
D_M = D_AB + D_BC + D_AC + D_ABC
3 2 3 3 1 3 4 2 0 6 5 ... 5 0 1 1 1 3
III
SiRMS-mix descriptors of a mixture of A, B and C
1 0 2 5 ... 1 0 1 0 0 1 3 2 3 3 1 3 4 2 0 6 5 ... 5 0 1 1 1 3

Fig. 3 Generation of simplex descriptors for a mixture of three components

Since a CGR represents some sort of pseudomolecule, it can be encoded by fragment descriptors.

Here, two different types of ISIDA fragment descriptors—augmented atoms and sequences with length varying from 1 to 8 atoms—were calculated using ISIDA Fragmenter tool [6]. In order to reduce the number of generated fragments, the hydrogen suppressed graphs were used. Dynamic bond and atom labels were added to the specifications of the fragments.

Reaction fingerprints

A reaction fingerprint is the 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 [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].

Models building and validation

The models were built by the Random Forest approach using the randomForest R package [37]. The optimal number of variables used to select the best split of trees nodes was estimated by a grid search using *caret* package [38]. Number of trees was equal to 500 in all cases. All other parameters were set to their default values provided by randomForest R package. Since Random Forest proved to be able to handle many descriptors with complex relationships, no variable selection has been performed.

Two model validation strategies were applied. The first one is a “reaction-out” approach which is consisted in ten

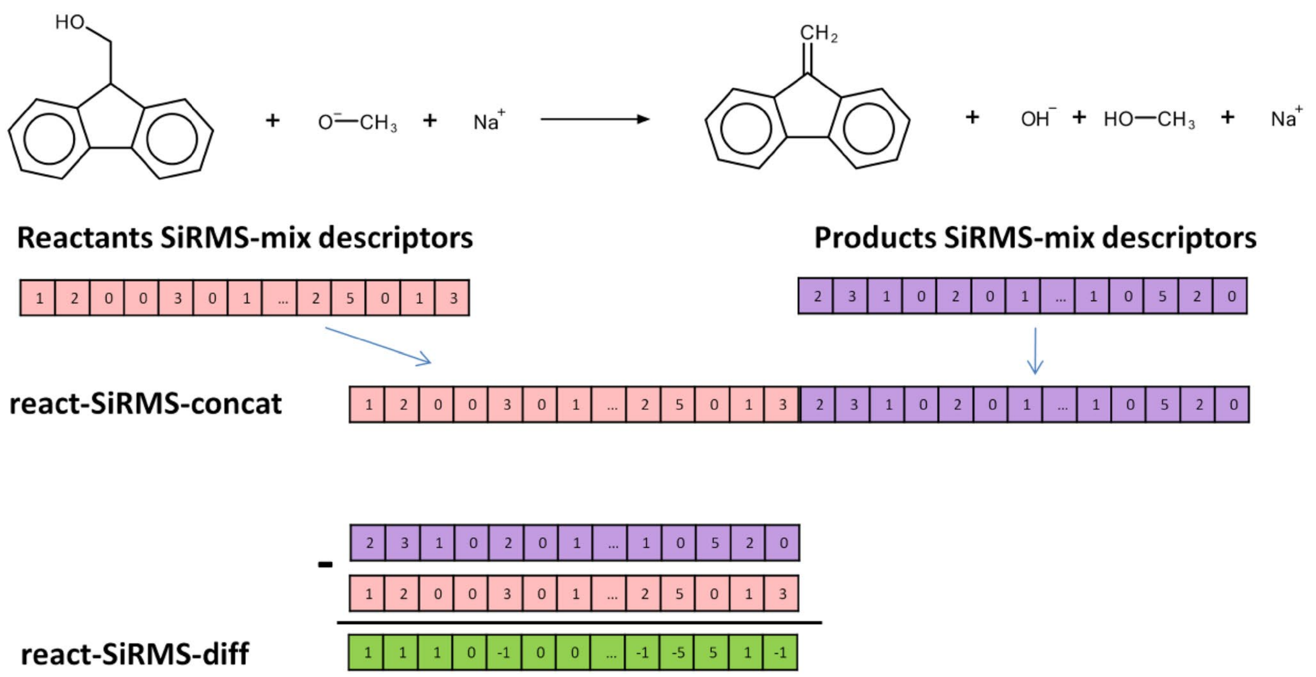
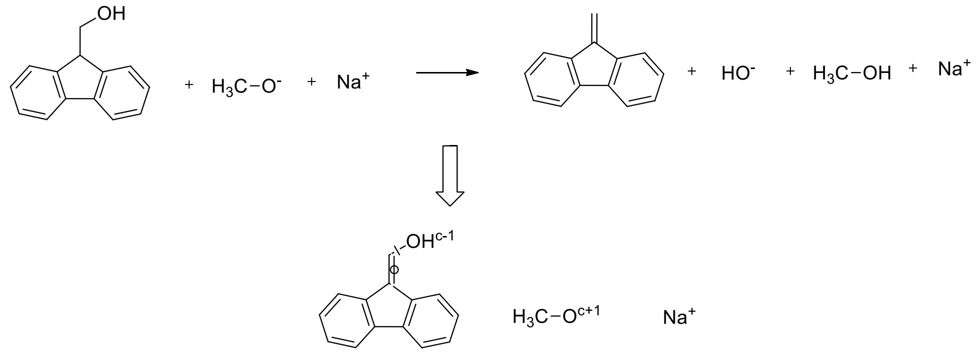


Fig. 4 Reaction descriptor vectors based on the concatenated product and reactant mixture descriptors (*react-SiRMS-concat*) and on their difference (r*eact-SiRMS-diff)*

Fig. 5 An example of encoding of an E2 reaction into a Condensed Graph of Reaction. The broken and formed bonds are labeled by a crossing and a circle, respectively. Oxygen atoms changing their formal charges are denoted by symbols “c + 1” (negative-to-neutral) and “c − 1” (neutral-to-negative)



times repeated fivefold cross-validation where folds were randomly generated. However, this conventional validation procedure overestimates the model performance because the same reaction may proceed under different conditions and, hence, it might become simultaneously a part of both training and test sets. Therefore, a more rigorous “product-out” strategy has been suggested. It assumes that in a particular fold, all reactions with the same main product are placed in the test set. Since the number of reactions with the same product significantly varies (from 1 to 27 reactions) the randomly created “product-out” folds may contain substantially different number of objects. More balanced folds were prepared using Monte-Carlo optimization of the variance of reaction counts across folds and ten the most diverse sets of folds were selected. Functions (*create_folds_mc, groupwise_tanimoto* and *select_folds*) used to generate the balanced folds are available in *pfpp* R package (https://github.com/DrrDom/pfpp).

The prediction performance of models was measured by Q^2 and root mean square error (RMSE).

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

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i (y_{i,pred} - y_{i,obs})^2}{N - 1}}$$

Since the cross-validation procedure was repeated 10 times, we were able to estimate statistical significance of difference between averaged performances of the best models using paired t-test.

Applicability domain of models

In order to discard reactions dissimilar to those in the training set, the "Fragment Control" applicability domain (AD) approach has been used [4]. The "Fragment Control" AD discards any test set reaction containing fragments which don't occur in the training set reactions. An AD was applied to the test set reactions at each fold followed by assembling the results for all folds. In such a way, statistical parameters were calculated for the entire set. Data coverage was assessed as a ratio of the number of reactions accepted by AD to the total number of reactions.

Results and discussion

Generally, the *SiRMS-mix* descriptors vary as a function of several parameters defining their size, complexity and labeling. The size of any simplex is defined by the minimal and maximal number of constituting atoms. Each atom was labeled either by element symbol or by partial charge category. Different mixture simplexes—pair-wise and triple-wise, etc.,—could take part of mixture descriptors. Since a huge number of *SiRMS-mix* descriptors corresponding to different combinations of the above parameters could be considered, we decided first to select their optimal values leading to the most performant models. These calculations were performed on concatenated reaction descriptors *react-SiRMS-concat*. Then, selected parameters were used in the modeling with difference reaction descriptors *react-SiRMS-diff*.

Selection of optimal parameters of *SiRMS* descriptors

Predictive performances of the models built on the concatenated reaction descriptors (*react-SiRMS-concat*) as a function of size, complexity and labeling of simplex descriptors is given in Fig. 6. One may see that more complex descriptors including both pair-wise and triple-wise mixture simplexes (*mult* = *3*, Fig. 6) perform similarly to descriptors including pair-wise mixture simplexes only (*mult* = *2*).

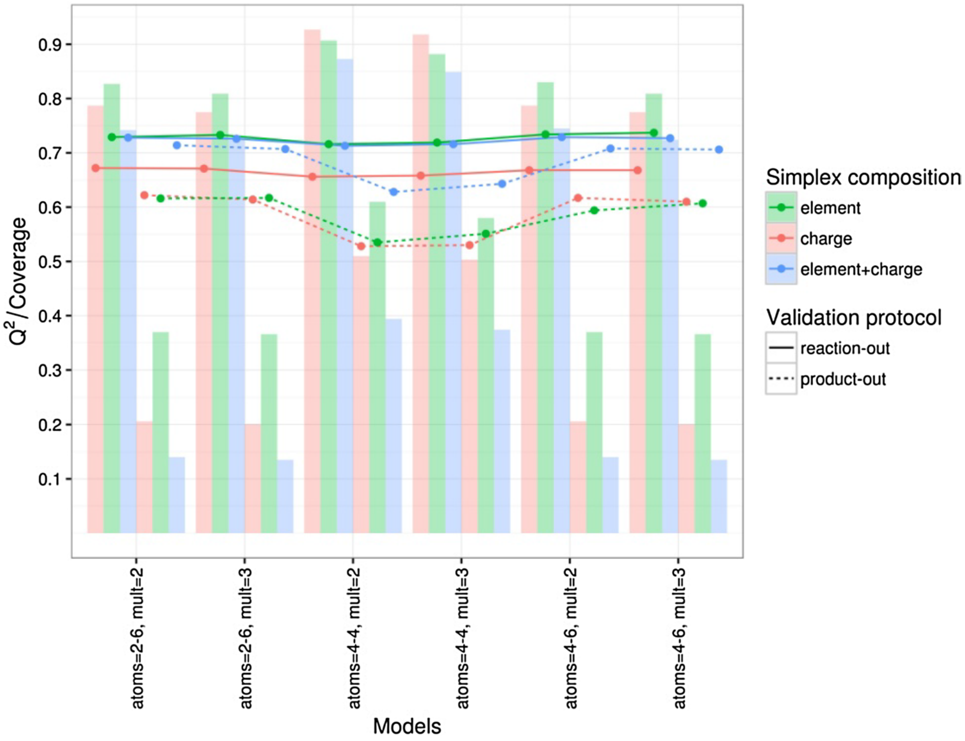


Fig. 6 The cross-validation determination coefficient Q^2 and the data coverage of the *react-SiRMS-concat* models as a function of size, complexity and composition of simplex descriptors. The model applicability domain was taken into account. *Solid and dashed lines* represent the Q^2 values, respectively, for "reaction-out" and "product-out" validation strategies, whereas corresponding *bars* show the data coverage. The color code reflects the composition of simplex descriptors including subgraphs encoding by elements (*green*), by charges (*red*), and, by both elements and charges (*blue*). The labels at the *horizontal axis* specify the minimal and maximal number of atoms in simplexes (e.g., "*atoms* = *2–6*") and complexity of mixture simplexes used: *mult* = *2* for pair-wise only and *mult* = *3* for pair-wise and triple-wise combinations

Variation of the number of atoms in simplexes doesn't impact the models performance for "reaction-out" CV (solid line in Fig. 6). However, Q^2 values for "product-out" CV significantly vary as a function of maximal number of atoms (N_{max}): the models with $N_{max}=6$ perform better than those with $N_{max}=4$. This could be explained by the fact that larger fragments better characterize substrates specificity. On the other hand, the occurrence of fragments in the training set decreases with their size. This explains significant reduction of data coverage due to application of "fragment control" applicability domain. Notice that models performance doesn't significantly vary as a function of minimal number of atoms (N_{min}). Indeed, at $N_{max}=6$, within the given validation strategy and atoms labeling, the models with $N_{min}=2$ and 4 perform very similarly (see Fig. 6).

Comparison of different schemes of atoms labeling in simplexes shows that consideration of atomic charges together with element types (blue lines on Fig. 6) increases the models' performance. This suggests particular importance of charge encoding for the reactivity modeling.

Benchmarking calculations

The results of benchmarking calculations comparing performances of the models based on *SiRMS-mix*, ISIDA/CGR descriptors as well as on different types of fingerprints are summarized in Table 2. One can see that two strategies of preparation of the reaction feature vector—either products and reactants vectors concatenation (*react-SiRMS-concat*) or their subtraction (*react-SiRMS-diff*)—lead to models of similar performances. Reasonable statistical parameters were obtained in "reaction-out" CV (Q^2 = 0.62–0.69, RMSE = 0.78–0.90), whereas "product-out" CV led to much worse statistical parameters (Q^2 = 0.37–0.47, RMSE = 1.03–1.15). The use of model AD significantly improved the model performance, especially in "product-out" CV (Q^2 = 0.59–0.74, RMSE = 0.75–0.86) which was close to the "reaction-out" CV performance (Q^2 = 0.67–0.74, RMSE = 0.74–0.90).

Table 2 Statistical parameters of the best QSAR models based on *SiRMS-mix* descriptors, different types of reaction fingerprints and ISIDA/CGR descriptors

No	Descriptors	Labeling scheme[a]	Models validation[b]	mtry[c]	Q^2	RMSE	Q$^2_{DA}$	RMSE$_{DA}$	Coverage
1	React-SiRMS-diff[d]	chg	R-OUT	3243	0.62	0.87	0.68	0.83	0.80
2		elm/chg		4066	0.68	0.81	0.74	0.74	0.76
3		elm		1371	0.69	0.78	0.74	0.74	0.85
4		chg	P-OUT	1622	0.36	1.14	0.64	0.86	0.22
5		elm/chg		2033	0.42	1.08	0.74	0.75	0.15
6		elm		411	0.47	1.03	0.64	0.90	0.38
7	React-SiRMS-concat[d]	chg	R-OUT	4053	0.63	0.86	0.67	0.84	0.79
8		elm/chg		4029	0.67	0.81	0.73	0.76	0.75
9		elm		2648	0.69	0.79	0.73	0.75	0.83
10		chg	P-OUT	2026	0.35	1.15	0.62	0.89	0.21
11		elm/chg		2821	0.39	1.11	0.71	0.80	0.14
12		elm		794	0.43	1.07	0.59	0.90	0.37
13	Atom pairs fingerprints		R-OUT	100	0.61	0.89	0.62	0.87	0.97
14			P-OUT	10	0.35	1.14	0.41	1.07	0.64
15	Morgan fingerprints		R-OUT	250	0.67	0.82	0.70	0.79	0.92
16			P-OUT	50	0.40	1.10	0.67	0.81	0.33
17	Topological torsion fingerprints		R-OUT	75	0.60	0.90	0.62	0.88	0.94
18			P-OUT	10	0.34	1.15	0.51	1.03	0.45
19	ISIDA/CGR[e]		R-OUT	519	0.69	0.79	0.74	0.74	0.88
20			P-OUT	156	0.41	1.09	0.61	0.90	0.16

[a] Atom labeling for SiRMS-mix descriptors: *chg* partial atomic charge, *elm* elements, *elm/chg* both schemes

[b] R-OUR and P-OUT correspond to "reaction-out" and "product-out" validation strategies, correspondingly

[c] The number of variable selected as candidates at each node split of RF model

[d] React-SiRMS-diff and react-SiRMS-concat are SiRMS-mix descriptor generated by concatenation or difference methods considering only pairwise component combinations with overall number of atoms in fragments from 4 to 6

[e] Augmented atoms descriptors with distance from 1 to 8

The observed performance improvement is linked to decrease of the data coverage which varies from 75 to 83% in the “reaction-out” CV and from 14 to 38% in “product-out” CV.

The comparison of the best *SiRMS* model (No. 5, Table 2) with the models involving different types of fingerprints and ISIDA/CGR descriptors is given on Fig. 7. One can see that all models in the “reaction-out” CV protocol perform similarly. However, this is not a case for the “product-out” cross-validation where the statistical parameters of the models built on Atom Pairs fingerprints and Topological Torsion fingerprints are very little predictive ($Q^2_{DA} < 0.5$). The best *SiRMS* model performs better than the models based on ISIDA/CGR descriptors (model No. 20, Table 2; p-value = 0.0002) and Morgan fingerprints (model No. 16, Table 2; p-value = 0.0080).

Although in “reaction-out” cross validations all sets of descriptors perform reasonably well this doesn’t reflect the predictive ability of models with respect to reactions leading to new products which can be assessed in “product-out” cross-validation. The Q^2 and RMSE values obtained in “product-out” CV are relatively low. Fragment control applicability domain significantly improves the model performance discarding up to 85% of reactions. Such big lost in the data coverage can be explained by high structural diversity and relatively small size of the data set due to which the test set objects often contain the fragments absent in the training set.

Conclusion

The suggested here mixture-based simplex representation of chemical reactions has been applied to the modeling of rate constants of E2 reactions. This approach doesn’t need any explicit information about reaction center and, therefore, atom-to-atom mapping is not required. The latter represents a significant advantage compared to methods based on explicit consideration of the reaction center because AAM procedure is time consuming and may lead to erroneous results [20]. However, as any other method of implicit encoding of a reaction center, our approach requires complete reaction representation (all products and all reactants). The *SiRMS-mix* models perform better than the models built on ISIDA/CGR descriptors and Morgan reaction fingerprint and much better than those involving reaction fingerprints encoding atom pairs or topological torsions. However, *SiRMS-mix* models have the lowest coverage according to the chosen Fragment control applicability domain approach.

A clear advantage of SiRMS approach is a possibility to vary the size and composition of considered molecular subgraphs (simplexes) and, in such a way, to select the descriptors set which fits modeled property. Thus, addition of simplexes labeled by partial atomic charge improves predictive performance of the models, which might be explained by significant role of electrostatic interactions in the E2 reaction mechanism. The SiRMS approach explicitly encodes different combinations of fragments occurred in reactants

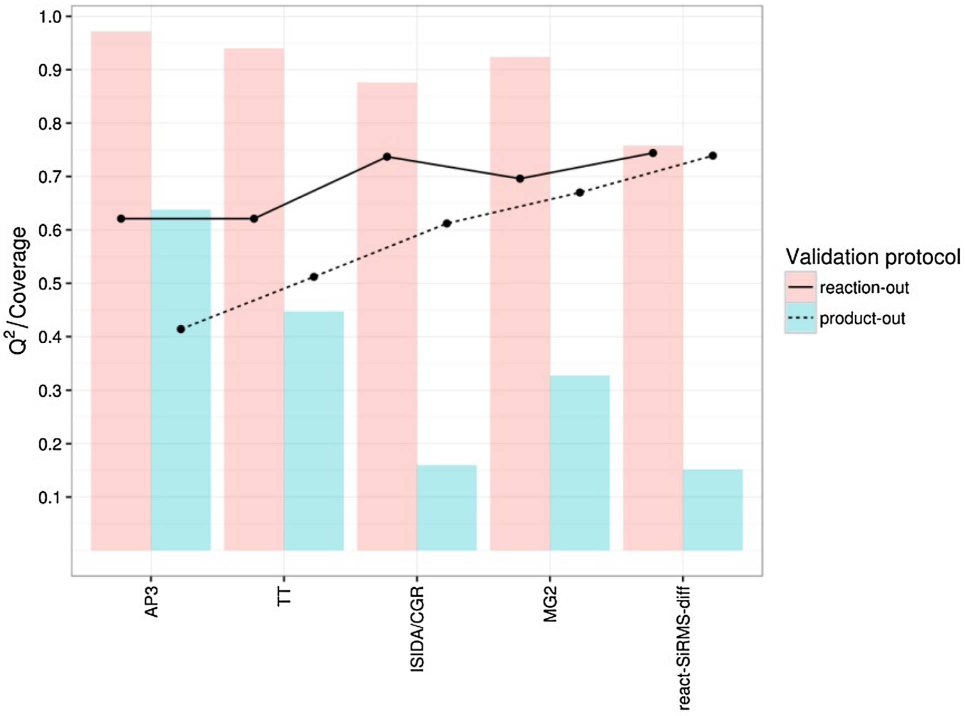


Fig. 7 Benchmarking of the models for E2 reaction rate constants involving different descriptors. The *dots* connected by *solid and dashed lines* represent Q^2_{DA} values calculated considering applicability domain for “reaction-out” and “product-out” validation strategies correspondingly. The *bars* represent coverage of the corresponding models. The labels on the *x axis* mean *AP3* atom pairs fingerprints, *TT* topological torsion fingerprints, *MG2* Morgan fingerprints

and products. Therefore, compared to reaction fingerprints from RDKit, SiRMS includes information not only about chemical transformations but also about all chemical functions present in reactants and products.

It has been demonstrated that the Fragment control AD could significantly improve the model performance. However, at the same time this leads to the reduction of the data coverage which is explained by small size and high diversity of the studied data set.

In parallel with the classical "reaction-out" cross-validation strategy we suggested to apply the more aggressive "product-out" cross-validation protocol which reliably assesses the accuracy of predictions for the reactions leading to new products.

Software implementation

The described reaction SiRMS descriptors were implemented in the open-source software written on Python 3 which is available in the Github repository https://github.com/DrrDom/sirms/releases/tag/v1.0.1.

Acknowledgements This work was supported by Russian Science Foundation, Grant No. 14-43-00024.

References

1. Chen WL, Chen DZ, Taylor KT (2013) Automatic reaction mapping and reaction center detection. Wiley Interdiscip Rev Comput Mol Sci 3(6):560–593. doi:10.1002/wcms.1140
2. Zhang J, Kleinöder T, Gasteiger J (2006) Prediction of pKa values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. J Chem Inf Model 46(6):2256–2266. doi:10.1021/ci060129d
3. Gasteiger J, Hondelmann U, Rose P, Witzenbichler W (1995) Computer-assisted prediction of the degradation of chemicals: hydrolysis of amides and benzoylphenylureas. J Chem Soc Perkin Trans 2(2):193–204. doi:10.1039/p29950000193
4. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko IV, Marcou G (2008) ISIDA—platform for virtual screening based on fragment and pharmacophoric descriptors. Curr Comput Aided Drug Des 4(3):191–198. doi:10.2174/157340908785747465
5. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) ISIDA property-labelled fragment descriptors. Mol Inform 29(12):855–868. doi:10.1002/minf.201000099
6. Varnek A, Fourches D, Hoonakker F, Solov'ev VP (2005) Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. J Comput Aided Mol Des 19(9):693–703. doi:10.1007/s10822-005-9008-0
7. Hoonakker F, Lachiche N, Varnek A, Wagner A (2011) A representation to apply usual data mining techniques to chemical reactions—illustration on the rate constant of SN2 reactions in water. Int J Artif Intell Tools 20(02):253–270. doi:10.1142/S0218213011000140
8. de Luca A, Horvath D, Marcou G, Solov'ev V, Varnek A (2012) Mining chemical reactions using neighborhood behavior and condensed graphs of reactions approaches. J Chem Inf Model 52(9):2325–2338. doi:10.1021/ci300149n
9. Madzhidov TI, Polishchuk PG, Nugmanov RI, Bodrov AV, Lin AI, Baskin II, Varnek AA, Antipin IS (2014) Structure-reactivity relationships in terms of the condensed graphs of reactions. Russ J Org Chem 50(4):459–463. doi:10.1134/S1070428014040010
10. Nugmanov RI, Madzhidov TI, Haliullina GR, Baskin II, Antipin IS, Varnek A (2014) Development of "structure-reactivity" models for nucleophilic substitution reactions with participation of azides. J Struct Chem 55(6):1080–1087
11. Madzhidov T, Bodrov A, Gimadiev T, Nugmanov R, Antipin I, Varnek A (2015) Obtaining structure-reactivity relationships for bimolecular elimination reactions with Condensed Reaction Graph approach. J Struct Chem 56(7):1227–1234
12. Marcou G, Aires de Sousa J, Latino DARS, de Luca A, Horvath D, Rietsch V, Varnek A (2015) Expert system for predicting reaction conditions: the michael reaction case. J Chem Inf Model 55(2):239–250. doi:10.1021/ci500698a
13. Faulon J-L, Visco DP, Pophale RS (2003) The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. J Chem Inf Comput Sci 43(3):707–720. doi:10.1021/ci020345w
14. Ridder L, Wagener M (2008) SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. ChemMedChem 3(5):821–832. doi:10.1002/cmdc.200700312
15. Schneider N, Lowe DM, Sayle RA, Landrum GA (2015) Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. J Chem Inf Model 55(1):39–53. doi:10.1021/ci5006614
16. Zhang Q-Y, Aires-de-Sousa J (2005) Structure-based classification of chemical reactions without assignment of reaction centers. J Chem Inf Model 45(6):1775–1783. doi:10.1021/ci0502707
17. Kravtsov AA, Karpov PV, Baskin II, Palyulin VA, Zefirov NS (2011) Prediction of rate constants of S_N2 reactions by the multicomponent QSPR method. Dokl Chem 440 (2):299–301. doi:10.1134/s0012500811100107
18. Faulon J-L, Misra M, Martin S, Sale K, Sapra R (2008) Genome scale enzyme—metabolite and drug—target interaction predictions using the signature molecular descriptor. Bioinformatics 24(2):225–233. doi:10.1093/bioinformatics/btm580
19. Kravtsov AA, Karpov PV, Baskin II, Palyulin VA, Zefirov NS (2011) Prediction of the preferable mechanism of nucleophilic substitution at saturated carbon atom and prognosis of S N 1 rate constants by means of QSPR. Dokl Chem 441 (1):314–317. doi:10.1134/s0012500811110048
20. Muller C, Marcou G, Horvath D, Aires-de-Sousa J, Varnek A (2012) Models for identification of erroneous atom-to-atom mapping of reactions performed by automated algorithms. J Chem Inf Model 52(12):3116–3122. doi:10.1021/ci300418q
21. Patel H, Bodkin MJ, Chen B, Gillet VJ (2009) Knowledge-based approach to de novo design using reaction vectors. J Chem Inf Model 49(5):1163–1184. doi:10.1021/ci800413m
22. Oprisiu I, Varlamova E, Muratov E, Artemenko A, Marcou G, Polishchuk P, Kuz'min V, Varnek A (2012) QSPR approach to predict nonadditive properties of mixtures. Application to bubble point temperatures of binary mixtures of liquids. Mol Inform 31(6–7):491–502. doi:10.1002/minf.201200006
23. Palm VA (1974–1978) Tables of rate and equilibrium constants of heterolytic organic reactions, vol 1–5. Moscow
24. Catalán J, Díaz C (1997) A generalized solvent acidity scale: the solvatochromism of o-tert-butylstilbazolium betaine dye and its homomorph o, o′-di-tert-butylstilbazolium betaine dye. Liebigs Ann 1997 (9):1941–1949. doi:10.1002/jlac.199719970921
25. Catalán J, Díaz C, López V, Pérez P, De Paz J-LG, Rodríguez JG (1996) A generalized solvent basicity scale: the solvatochromism of 5-nitroindoline and its homomorph

1-methyl-5-nitroindoline. Liebigs Ann 1996 (11):1785–1794. doi:10.1002/jlac.199619961112

26. Catalán J, López V, Pérez P, Martin-Villamil R, Rodríguez J-G (1995) Progress towards a generalized solvent polarity scale: The solvatochromism of 2-(dimethylamino)-7-nitrofluorene and its homomorph 2-fluoro-7-nitrofluorene. Liebigs Ann 1995 (2):241–252. doi:10.1002/jlac.199519950234
27. Taft RW, Kamlet MJ (1976) The solvatochromic comparison method. 2. The .alpha.-scale of solvent hydrogen-bond donor (HBD) acidities. J Am Chem Soc 98(10):2886–2894. doi:10.1021/ja00426a036
28. Kamlet MJ, Taft RW (1976) The solvatochromic comparison method. I. The .beta.-scale of solvent hydrogen-bond acceptor (HBA) basicities. J Am Chem Soc 98(2):377–383. doi:10.1021/ja00418a009
29. Kamlet MJ, Abboud JL, Taft RW (1977) The solvatochromic comparison method. 6. The .pi.* scale of solvent polarities. J Am Chem Soc 99(18):6027–6038. doi:10.1021/ja00460a031
30. cxcalc. 5.4 edn. Chemaxon, Budapest, Hungary
31. Kuz'min VE, Artemenko AG, Muratov EN (2008) Hierarchical QSAR technology based on the Simplex representation of molecular structure. J Comput Aided Mol Des 22(6–7):403–421. doi:10.1007/s10822-008-9179-6
32. Kuz'min VE, Artemenko AG, Polischuk PG, Muratov EN, Khromov AI, Liahovskiy AV, Andronati SA, Makan SY (2005) Hierarchic system of QSAR models (1D-4D) on the base of simplex representation of molecular structure. J Mol Model 11:457–467. doi:10.1007/s00894-005-0237-x
33. RDKit: Open-Source Cheminformatics. http://www.rdkit.org
34. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25(2):64–73. doi:10.1021/ci00046a002
35. Rogers D, Hahn M (2010) Extended-Connectivity Fingerprints. J Chem Inf Model 50(5):742–754. doi:10.1021/ci100050t
36. Nilakantan R, Bauman N, Dixon JS, Venkataraghavan R (1987) Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. J Chem Inf Comput Sci 27(2):82–85. doi:10.1021/ci00054a008
37. Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22
38. Max Kuhn. Contributions from Jed Wing and Steve Weston and Andre Williams and Chris Keefer and Allan Engelhardt and Tony Cooper and Zachary Mayer and the R Core Team caret: Classification and Regression Training (2014). R package version 6.0–30 edn.