

Bimolecular nucleophilic substitution reactions: predictive models for rate constants and Molecular Reaction Pairs analysis

Timur Gimadiev^[a,b], Timur Madzhidov^[a], Igor Tetko^[c], Ramil Nugmanov^[a], Iury Casciuc^[b], Olga Klimchuk^[b], Andrey Bodrov^[d], Pavel Polischuk^[d], Igor Antipin^[a], and Alexandre Varnek^{[b]*}

^[a] Laboratory of Chemoinformatics and Molecular Modeling, Butlerov Institute of Chemistry, Kazan Federal University, Kremlyovskaya str. 18, Kazan, Russia

^[b] Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France;

^[c] Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, b. 60w, D-85764 Neuherberg, Germany

^[d] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Křížkovského str. 511/8, Olomouc, Czech Republic.

Abstract

Here, we report the data visualization, analysis and modeling for a large set of 4830 S_N2 reactions the rate constant of which ($\log k$) was measured at different experimental conditions (solvent, temperature). The reactions were encoded by one single molecular graph - Condensed Graph of Reactions, which allowed us to use conventional chemoinformatics techniques developed for individual molecules. Thus, Matched Reaction Pairs approach was suggested and used for the analyses of substituents effects on the substrates and nucleophiles reactivity. The data were visualized with the help of the Generative Topographic Mapping approach. Consensus Support Vector Regression (SVR) model for the rate constant was prepared. Unbiased estimation of the model's performance was made in cross-validation on reactions measured on unique structural transformations. The model's performance in cross-validation (RMSE=0.61 $\log k$ units) and on the external test set (RMSE=0.80) is close to the noise in data. Performances of the local models obtained for selected subsets of reactions proceeding in particular solvents or with particular type of nucleophiles were similar to that of the model built on the entire set. Finally, four different definitions of model's applicability domains for reactions were examined.

Keywords: bimolecular nucleophilic substitution reactions, Condensed Graph of Reaction, Matched Reaction Pairs, Support Vector Regression, Generative Topographic Mapping, models applicability domain

Introduction

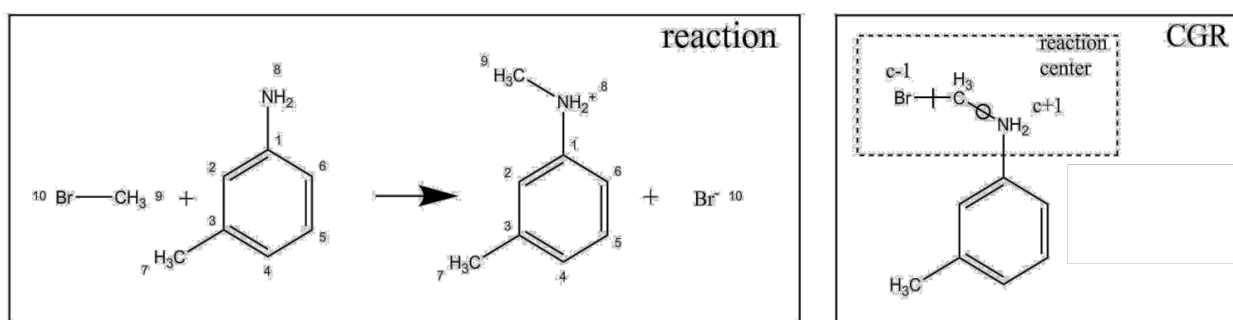
Compared to individual molecules, chemical reaction is a complex object because it involves several molecular species of two types (reactants and products) and its yield depends on experimental conditions (solvent, catalyst, temperature). This prevents applying to chemical reactions most of conventional methods designed for the analysis and modeling of individual compounds. This complexity could be reduced using the Condensed Graph of Reaction (CGR) approach [1] representing reaction by a single 2D graph, some sort of pseudomolecule, characterized by both conventional chemical bonds and such called dynamic bonds characterizing chemical transformations. Fragment descriptors generated for CGR can be successfully applied in any cheminformatics application used a descriptors vector as an input. This approach has successfully been used for similarity searching in reaction space [2], for data analysis using Generative Topographic Mapping [3] and for QSPR modeling of various kinetic and thermodynamic properties of reactions [4–6] or optimal reaction conditions [7]. On the other hand, some methods of data analysis considering chemical species as molecular graphs were never used for reactions analysis so far. One of these methods, Matched Molecular Pairs (MRP), is widely used in medicinal chemistry [8] for the analysis of the effects of replacement of one chemical group with another one. In this paper, we'll demonstrate how this approach can be extended to chemical reactions represented by their CGR.

Another goal of this paper is the development of predictive models for logarithm of rate constant ($\log k$) of bimolecular nucleophilic substitution reaction (S_N2). Nucleophilic substitution (S_N) is a fundamental class of reactions in which an electron rich molecule called nucleophile attacks the positive or partially positive charged atom of substrate molecule to replace a leaving group [9], called also nucleofuge (example shown on Figure 1). Bimolecular nucleophilic substitution S_N2 is referred to subclass of S_N reactions where the bond with leaving group is broken and the bond with nucleophile is formed synchronously. Notice that nucleophilic substitution reactions proceeding through formation of carbocation followed by ion recombination are denoted as S_N1 . Nucleophile could be either neutral (usually amine or alcohol) or negatively charged species (alcoholates, thiolates, halogen or other inorganic salt anions, deprotonated amines). Usually, only reactions with aliphatic carbon in reaction center are called as S_N2 reactions. Reactions involving substitution at aromatic or unsaturated carbon of substrate atom are usually asynchronous and follow either addition-elimination (S_NAr) or elimination-addition (S_N1) mechanisms.

Prediction models for S_N2 reaction rate were reported in some earlier publications [10–12]. Thus, Baskin et al [10] reported neural networks models for $\log k$ on a set of 1732 reactions

proceeding in various pure solvents at different temperatures. The feature vector combined descriptors encoding, on one hand, chemical structure of substrate and product molecules, and, on the other hand, solvent and temperature. Hoonakker et al [11] reported predictive models for $\log k$ built on a dataset of 1014 S_N2 reactions proceeding in water at different temperatures. They used three different machine-learning techniques (SVR, M5P, MLR) and ISIDA fragment descriptors derived from CGR in combination with inverse temperature. Madzhidov et al [12] reported Random Forest models built on a dataset of 1041 S_N2 reaction proceeding in different solvents and water-organic solvent mixtures, and at different temperatures but involving only neutral nucleophiles. The models were built on ISIDA or SiRMS [13] descriptors encoding structure of reactants completed by 14 solvent descriptors and inverse temperature. The study by Nugmanov et al [14] was focused on particular case of S_N2 reactions involving azide ion as nucleophile

Unlike previous studies, here the SVM consensus model for $\log k$ were built on the set of 4830 reactions involving both neutral and anionic nucleophiles and proceeding at different temperatures in 43 different solvents and water-organic mixtures. This is the biggest and the most diverse data set of S_N2 reactions considered so far. The data set was analyzed using the the GTM approach. Matched Reaction Pairs approach was described and used for the analyses of substituents effects on the substrates and nucleophiles reactivity. Unbiased estimation of the model's performance was performed in cross-validation on reactions measured on unique structural transformations. Performances of the global model built on the entire data set and local models obtained for the subsets of reactions proceeded in particular solvent were compared. Finally, different applicability domains for reactions were compared. The developed consensus model is available for the user at our server *cimm.kpfu.ru*.



Scheme 1. Example of S_N2 reaction and of related Condensed Graphs of Reaction. The CGR has two dynamic bonds and two dynamic atoms. The former describe one formed bond (denoted by circle) and one broken bond (denoted by a crossed line). The latter describe the atoms which change their charges from 0 to -1 (labeled by "c-1") and from 0 to +1 (labeled by "c+1").

2. Computational procedure

2.1 Data preparation

More than 8000 reactions were manually collected from the handbook by Palm and registered in the database created under IJChem (ChemAxon) environment [15]. Each database record included the following information: structure of reactants and products, rate constants, temperature, solvent, molar percentage of organic solvent in water-organic mixture (100% for pure solvent, including pure water), literature source. Reactions with, at least, one of mandatory field missed or those that proceed in undesired condition (binary mixture of organic solvents, tertiary mixture, or high-pressure reactions) were excluded.

Chemical structures were standardized with ChemAxon Standardizer [16] including functional group normalization, aromatization and atom-to-atom mapping (AAM). Explicitly specified hydrogen atoms were removed. Special attention was paid to duplicates filtering. The latter were defined as the same chemical transformations proceeding in the same conditions (solvent, temperature and composition of organic solvent-water mixture). Stereospecific reactions were also considered as duplicates. For each series of duplicated reactions, an average $\log k$ value and related standard deviation (SD) were calculated. Distribution of these SD values for the entire set shows that in most of cases $SD < 0.5 \log k$ units (Figure 1) which could be used as a reasonable estimation of experimental inter-laboratory error.

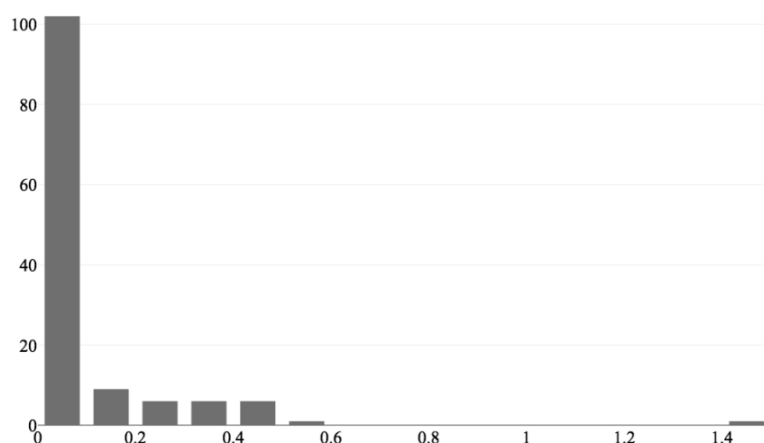
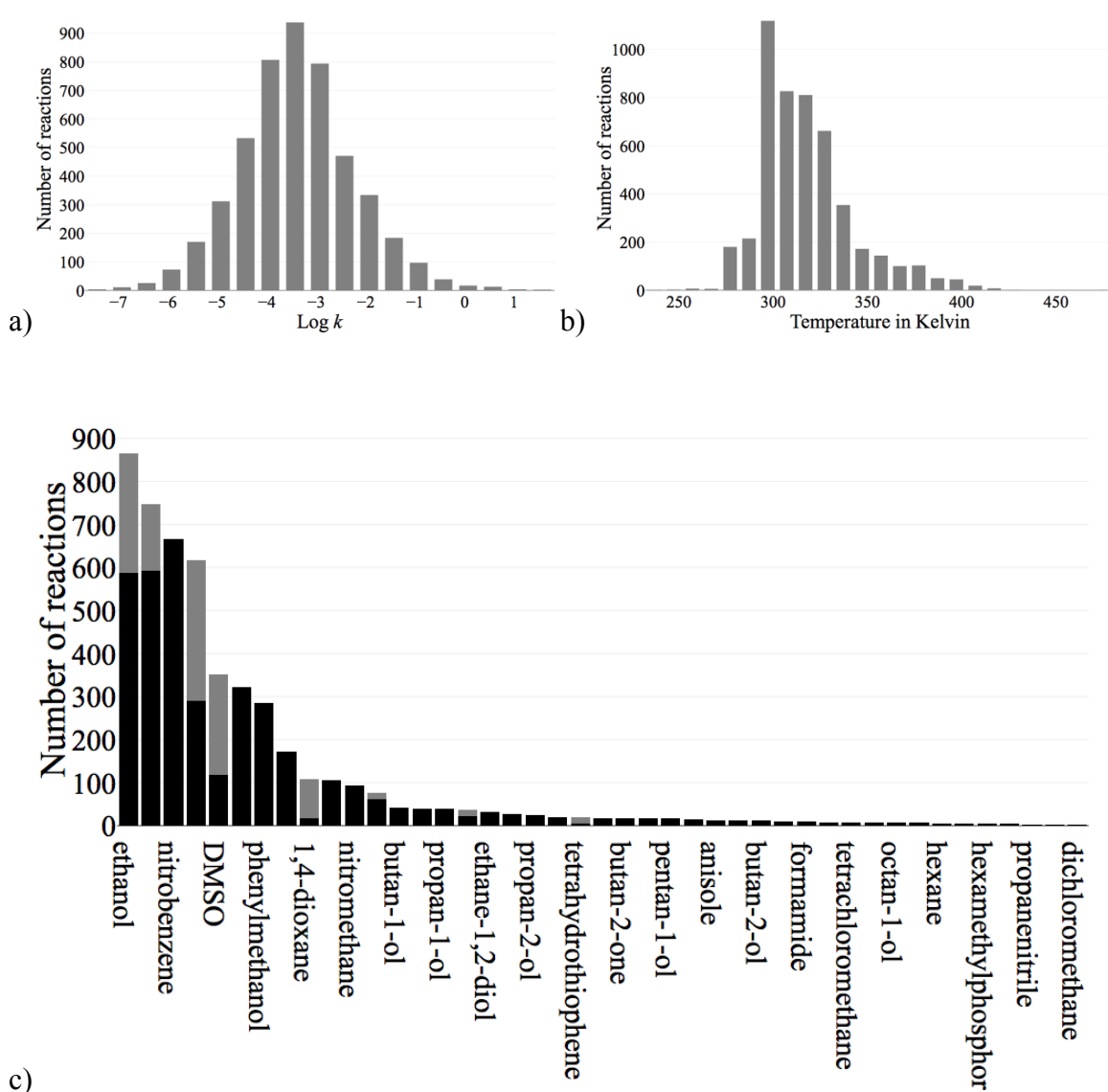


Figure 1. Histogram of standard deviation of $\log k$ values within a series of duplicated reactions.

The curated modeling set contains 4830 $\log k$ data points for 1383 unique transformations including 2882 reactions with neutral nucleophile and 1948 reactions involving anionic nucleophile in 43 different solvents and their mixtures with water (Figure 2c). The $\log k$ varied

from -7.68 to 1.65. The $\log k$ distribution resembles almost perfect Gaussian function with the peak at -3.5 (Figure 2a) whilst temperature distribution is highly skewed with expected cliff at 25°C (Figure 2b). The most popular solvents were ethanol, methanol, acetone often used in mixtures with water and nitrobenzene. Most of rate constants were measured at several different experimental conditions, only 551 reactions were studied at one condition only. For vast majority of reactions, less than 10 measurements of rate constant per transformation were reported (Figure 2d).



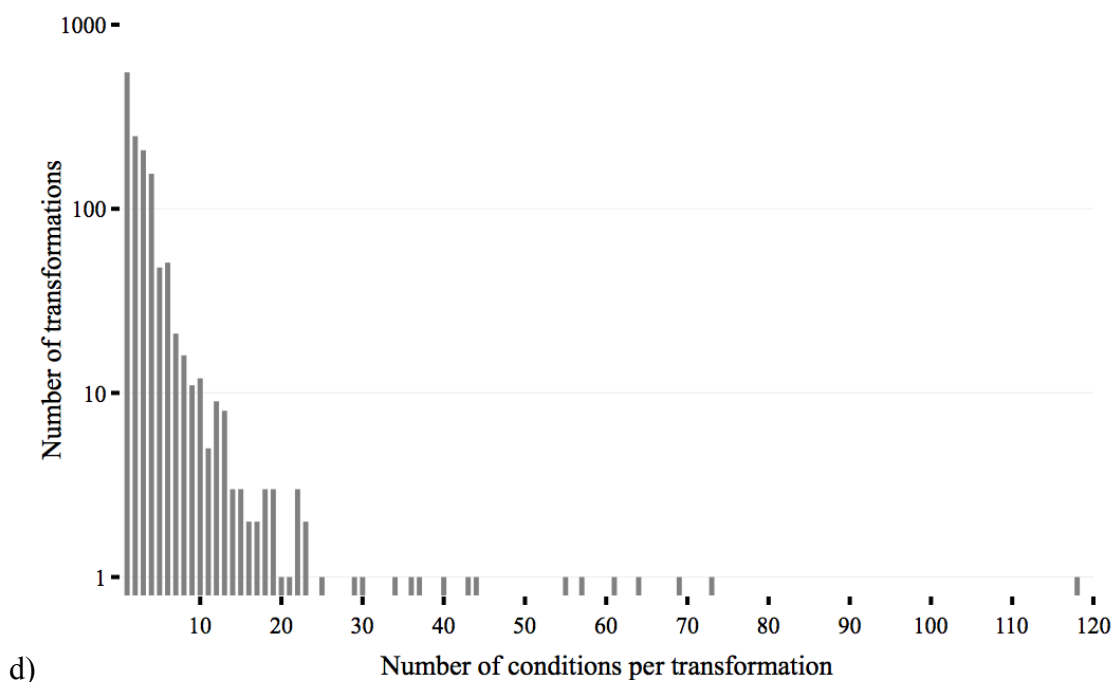


Figure 2. Data distribution with respect to (a) rate constant, (b) temperature, (c) solvent (dark part is for pure solvent one and light part is for mixture with water), (d) the number of experimental conditions per transformation.

2.2 Condensed Graphs of Reaction

The Condensed Graph of Reaction (CGR) approach [2, 11] was used to merge molecular graphs of all reactants and products into a single graph. In order to characterize chemical transformations CGR uses “dynamic bonds” corresponding to broken formed or transformed bonds, and “dynamic atoms”, characterizing changes of atomic charges upon the transformation (see Scheme 1). Previously this approach was used in the modeling of rate constants of various reactions (S_N2 [11, 12, 14], bimolecular elimination [5], Diels-Alder [17]), optimal condition prediction [7] as well as for detection of erroneous atom-to-atom mapping [18]. Technically, a CGR can be obtained from the reaction equation by superposing related atoms in the molecular graphs of reactants and products. Thus, an atom-to-atom mapping (AAM) procedure establishing these relations is required.

The CGR preparation workflow consists of the following steps: (1) all transformations were extracted in RDF format [19]; (2) atom-to-atom mapping was performed using the ChemAxon/Standartizer tool [16]; (3) the errors of mapping were fixed, first, using Indigo [20], then in *in-house* software [21]; and (4) CGRs corresponding to the transformations were built using the *in house* CGR-Designer tool and stored in SDF format.

2.3 Matched Reaction Pairs

Molecular Matched Pair (MMP) is defined for a pair of molecules, which are different with a respect of a single group [22]. Analysis of such differences across many pairs of molecules allows understanding of a trend with change of the analyzed property due to the appearance/disappearance of the respective group. The extension of MMPs to chemical reactions encoded by CGRs is straightforward since CGR represents a single molecular graph. Thus, instead of comparing a pair of compounds, one can compare a pair of reactions which we'll further call *Matched Reaction Pairs* (MRP), see Figure 4. The MRP analysis may help to understand how specific variations of reactants' structure affect kinetic or thermodynamic property of reaction.

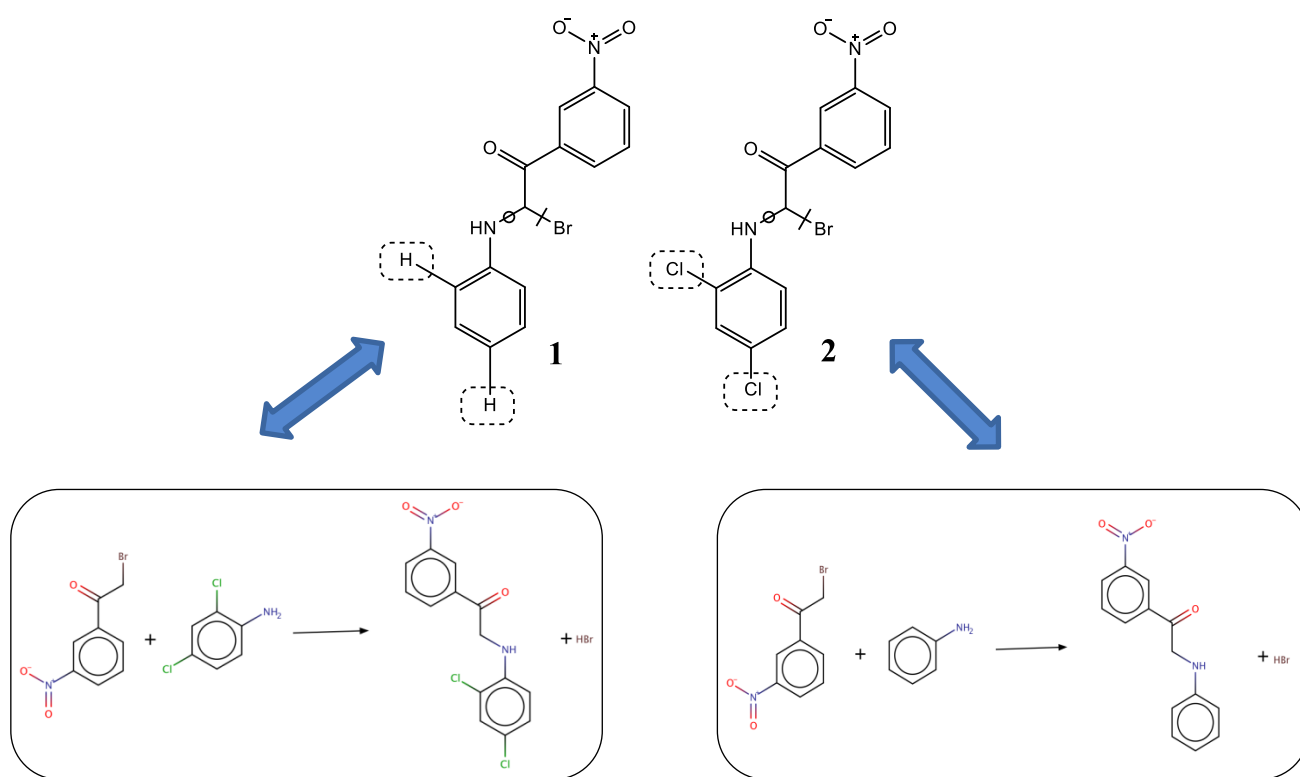


Figure 4. Example of Match Reaction Pairs (*top*) corresponding to replacement of two H atoms in **1** by two Cl atoms in **2**. Reactions used for CGR preparation are shown on the bottom

Since MRP reflects only structural factor, this analysis can be performed only for reactions proceeding in similar conditions. Therefore, a subset of 125 reactions carried out in pure methanol at ambient temperature has been prepared. All calculations were performed with the OCHEM tool [23] implementing the Hussain and Rea algorithm [22].

2.4 Data visualization with Generative Topographic Mapping

GTM constructs a 2D-dimensional map reproducing the data distribution of the initial D-dimensional space, defined by the descriptors [24, 25]. The map is framed by K nodes forming a perfect square grid. The nodes correspond to normal probability distributions centered on the GTM manifold, a flexible 2D envelope embedded into the D-dimensional data cloud.

Since we intended to analyze only structural diversity of reactions, the information about reaction conditions was not considered. The map was prepared for a set of 1383 unique transformations encoded by CGR. The atom-bond sequences of length from 2 to 4 containing at least one dynamic atom or bond were used as descriptors. In order to enhance the data analysis, the data points on 2D map were colored according to the type reaction signatures, substrate, and nucleophile nature (see Figure 3).

The calculations were performed with the ISIDA-GTM tool using default parameters for GTM construction [26].

2.5 Quantitative Structure-Reactivity modeling

2.5.1 Descriptors

The reactions were first rendered as Condensed Graphs of Reaction, created from the reaction RDF files using the in-house CGR Designer tool and stored in modified SDF format. This was directly processed by the in-house ISIDA Fragmentor software [2], in order to generate fragment descriptors. The length of monitored fragments varied from 2 to 14 for sequences and from 2 to 6 for atom-centered fragments. An important option regulating the amount of the overall generated CGR fragments is the ‘dynamic bond’ exclusive 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 center. That could be used to generate fragments that describe local environment of the reaction center exclusively. Overall, 616 descriptor sets have been generated for the preliminary SVR scanning.

Descriptors of the reaction conditions. Descriptors of reaction conditions included solvent descriptors and temperature. The solvent descriptors considered in this study include the values of polarity, polarizability, H-acidity and basicity: Catalan SPP[27], SA[28] and SB constants[29], Camlet-Taft α [30], β [31], and π^* [32] constants, 4 functions of dielectric constant ϵ (Born $f_B = \frac{\epsilon - 1}{\epsilon}$ and Kirkwood $f_K = \frac{\epsilon - 1}{2\epsilon + 1}$ functions, $f_1 = \frac{\epsilon - 1}{\epsilon + 1}$, $f_2 = \frac{\epsilon - 1}{\epsilon + 2}$), 3 functions of the refractive index n_D^{20} (denoted as n for the sake of simplicity in the following

formulae), $g_1 = \frac{n^2 - 1}{n^2 + 2}$, $g_2 = \frac{n^2 - 1}{2n^2 + 1}$, $h = \frac{(n^2 - 1)(\varepsilon - 1)}{(2n^2 + 1)(2\varepsilon + 1)}$. The final element of the condition descriptor features the reciprocal of reaction temperature, given in Kelvin⁻¹ (K).

2.5.2 Building and validation of SVR models

SVR models were built and validated using the ε -SVR algorithm implemented in the libSVM package^[43]. The modeling was performed using the evolutionary SVR optimizer^[44], which can be used to perform both descriptor space selection and optimization of the operational parameters (epsilon, kernel type, cost, gamma) of the SVR method. The procedure, applied to $\log k$ as a modeled property generated a total of 3000 models. Ten descriptors sets producing the SVR individual models of maximal robustness (estimated by Q^2 value obtained in 10 times repeated 5-fold cross-validation) have been selected. Resulting 500 models were used to predict the $\log k$ values for the reactions in the external test set. This consensus model is available for the users on our server <http://cimm.kpfu.ru> (see details in Supporting Information).

The predictive performance of the models has been estimated by root mean squared error (RMSE) and squared determination coefficient calculated in five-fold cross-validation (Q^2) repeated 10 times after the data reshuffling (10x5-CV), or on the external test set (R^2):

$$Q^2(R^2) = 1 - \frac{\sum_{i=1}^n (Y_{exp,i} - Y_{pred,i})^2}{\sum_{i=1}^n (Y_{exp,i} - \langle Y \rangle_{exp})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_{exp,i} - Y_{pred,i})^2}{n}} \quad (4)$$

Here Y_{exp} and Y_{pred} are, respectively, experimental and predicted $\log k$ values, n is the number of data points, while $\langle Y \rangle_{exp}$ is the mean of experimental values.

2.5.3 Applicability Domain of the models.

In order to improve the models performance on the external test set, several applicability domain (AD) definitions were examined. The first of them (“signature control”) retains the test set reactions which have reaction signatures the same as training set reactions. A “signature” represents a structural motif of CGR identifying a given type of reaction. Two types of signatures were considered. They correspond to the reaction center with its first one or two coordination spheres (“wide” and “narrow” signature controls, respectively).

A combination of fragment control and bounding box approaches [33, 34] were also used as AD for each individual model. The fragment control discards the objects possessing fragments

not present in the training set reactions. The bounding box AD discards a reaction for which, at least, one of its descriptors D_i , is outside of the range of D_i values computed for the training set reactions. When ensemble of individual models (“consensus” model) is simultaneously applied, a test set is retained if it is within fragment control and bounding box AD of, at least, of one individual model. Here, we’ll call this AD as “weak” consensus control. Additionally, a “firm” consensus control AD [35] was examined. It considers predictions unreliable if a given reaction is outside of AD for a certain percentage (50% by default) of individual models.

3. Data visualization

With the help of color code characterizing different types of transformations, substrates and nucleophiles, the 2D generative topographic map provides with clear view on chemical content of the studied reactions dataset. Four differently colored maps on Figure 3 show that different types of reactions are well separated which implicitly confirms the choice of fragment descriptors selected in genetic algorithm optimization.

One can see that the large majority of the data corresponds to the reactions between substrates containing C-Hal bonds with N- or O-containing nucleophiles (Figure 3a). The signature subsets separation is mostly governed by the substrate type: for a given halogen atom (Hal) the clusters corresponding to broken C-Hal bond and created C-O or C-N bonds tend to be located together. Figures 3b and 3c show, respectively, regions populated by reactions involving the most popular substrates and nucleophiles. The neutral nucleophiles occupy larger zone than anionic the ones which corresponds to their relative populations in the dataset (40% of anionic and 60% of neutral); these zones are well separated (Figure 3d).

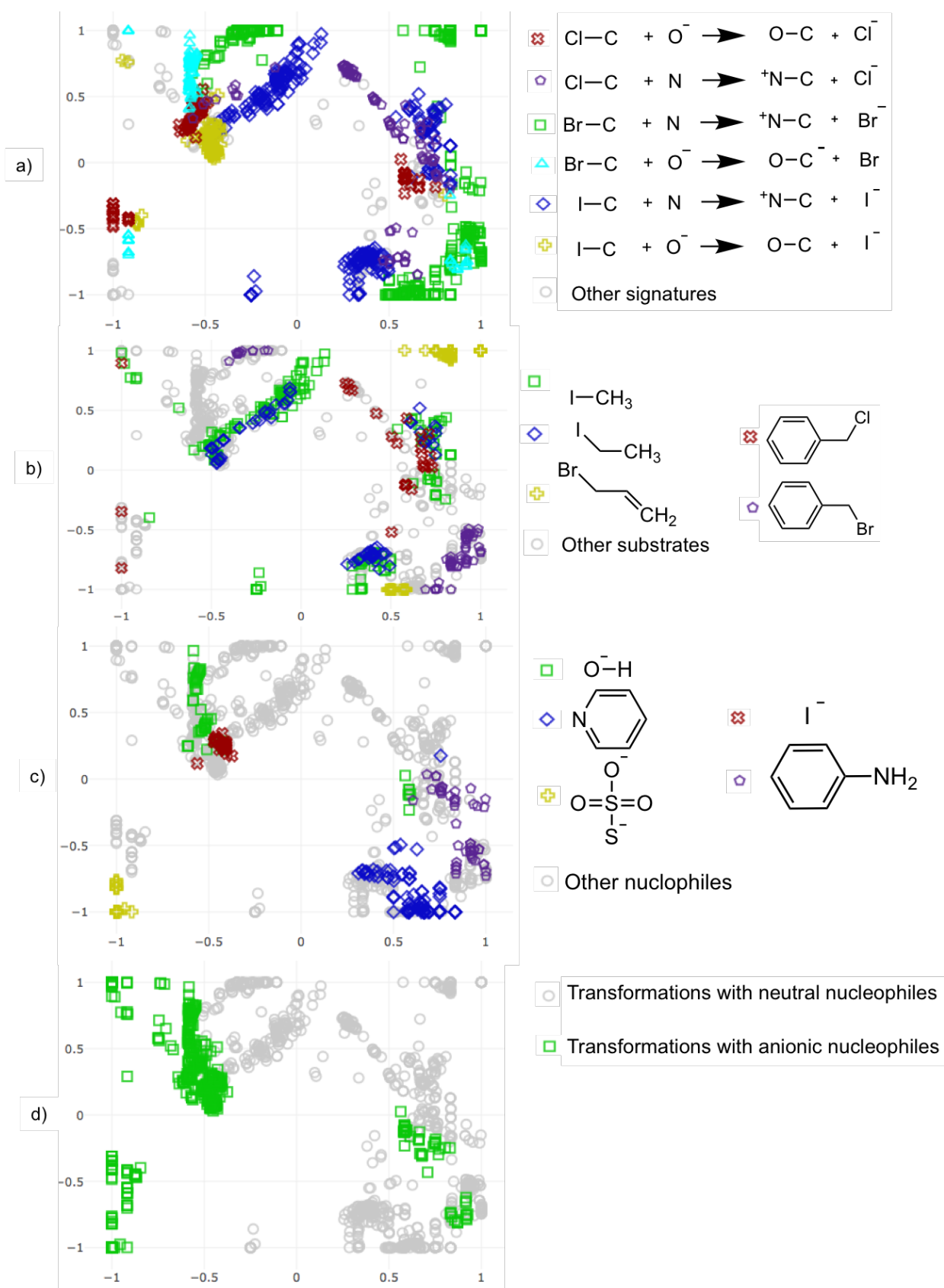


Figure 3. GTM map on 1383 unique transformations encoded by ISIDA fragments. Objects are colored according to a) reaction center signature (only reaction center atoms included), b) substrates, c) nucleophile structure, d) nucleophile type. The most popular signatures or molecules are shown explicitly.

4. Analysis of substituent effect using Matched Reactions Pairs (MRP)

For S_N2 reaction, substituents effects are conventionally considered in the framework of the reaction mechanism where an atom of nucleophile possessing lone electron pair or bearing negative charge attacks partially positively charged carbon atom which results in a leaving group replacement. Thus, electron donating substituents in nucleophile increase its reactivity and, hence, a reaction rate. Similarly act electron-acceptor substituents in substrate molecule which increase partial positive charge on reacting carbon atom.

In most of cases, the MRP analysis fully supports the known mechanism and follows conventional interpretation of substituents effects. For instance, substitution of hydrogen in nucleophile by chlorine (electron acceptor), as expected, slows the reaction down (Figure 5 *top*), whereas replacement of electron donating methoxy- to acceptor nitro-group in substrate molecule increases its speed (Figure 5 *middle*). However, MRP may also help to detect unexpected $\log k$ variations. For example, hydrogen / nitro-group replacements in the substrate (Figure 5 *bottom*) leads to either increase or decrease of the rate constant. In the right pair of reactions on Figure 5d, the H/ NO_2 replacement increases the reaction rate, which is fully in line with the conventional interpretation of substituent effects for S_N2 reactions. However, in 8 reaction pairs, one of which is shown in the left part of Figure 5c, this replacement leads to the decrease of $\log k$. This behavior is known for S_N1 reactions in which carbo-cationic intermediate is destabilized by electron-withdrawing substituents. On the other hand, it could also happen to an S_N2 reaction with late transition state and great charge separation where the bond with leaving group is strongly loosened. Then partial positive charge on carbon could be destabilized by electron acceptor and thus even in case of S_N2 reaction electron withdrawing group could slow the reaction down. Notice that this interesting observation can hardly be done without the MRP analysis.

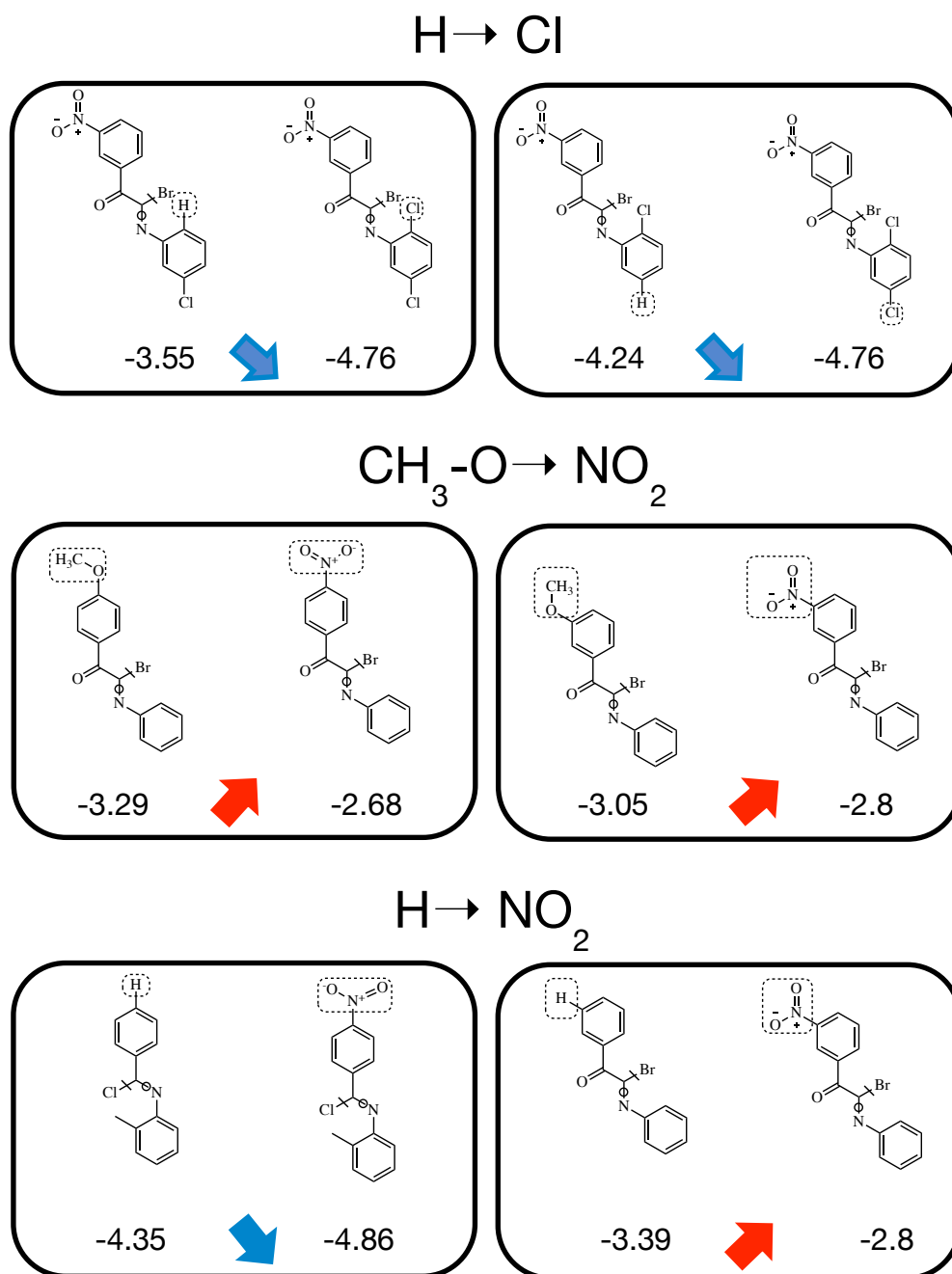


Figure 5. Examples of Molecular Reaction Pairs for S_N2 reactions in methanol at ambient temperature. In CGRs, formed and broken bond are shown as crossed and circled respectively. Blue and red arrows correspond, respectively, to the decrease and increase of $\log k$ induced by the given transformation.

5. Consensus model performance in cross-validation and outliers analysis

The performance of our SVR consensus model in cross-validation is pretty high: $Q^2=0.92$ and $RMSE=0.34 \log k$ units (see Figure 6) which is much better than that of previously reported SVR models by Honnakker et al. ($Q^2=0.53$, $RMSE=1.26$) [11] and Random Forest models by

Madzhidov et al ($Q^2 = 0.67$, $RMSE = 0.51$) [12] also based on the ISIDA descriptors. This could be explained by rigorous data curation protocol used in this study.

It should, however, be noticed that the above statistical parameters may not reflect a real predictive performance of the model since at the given CV fold the same reaction proceeding under slightly different conditions can be present both in training and test sets.

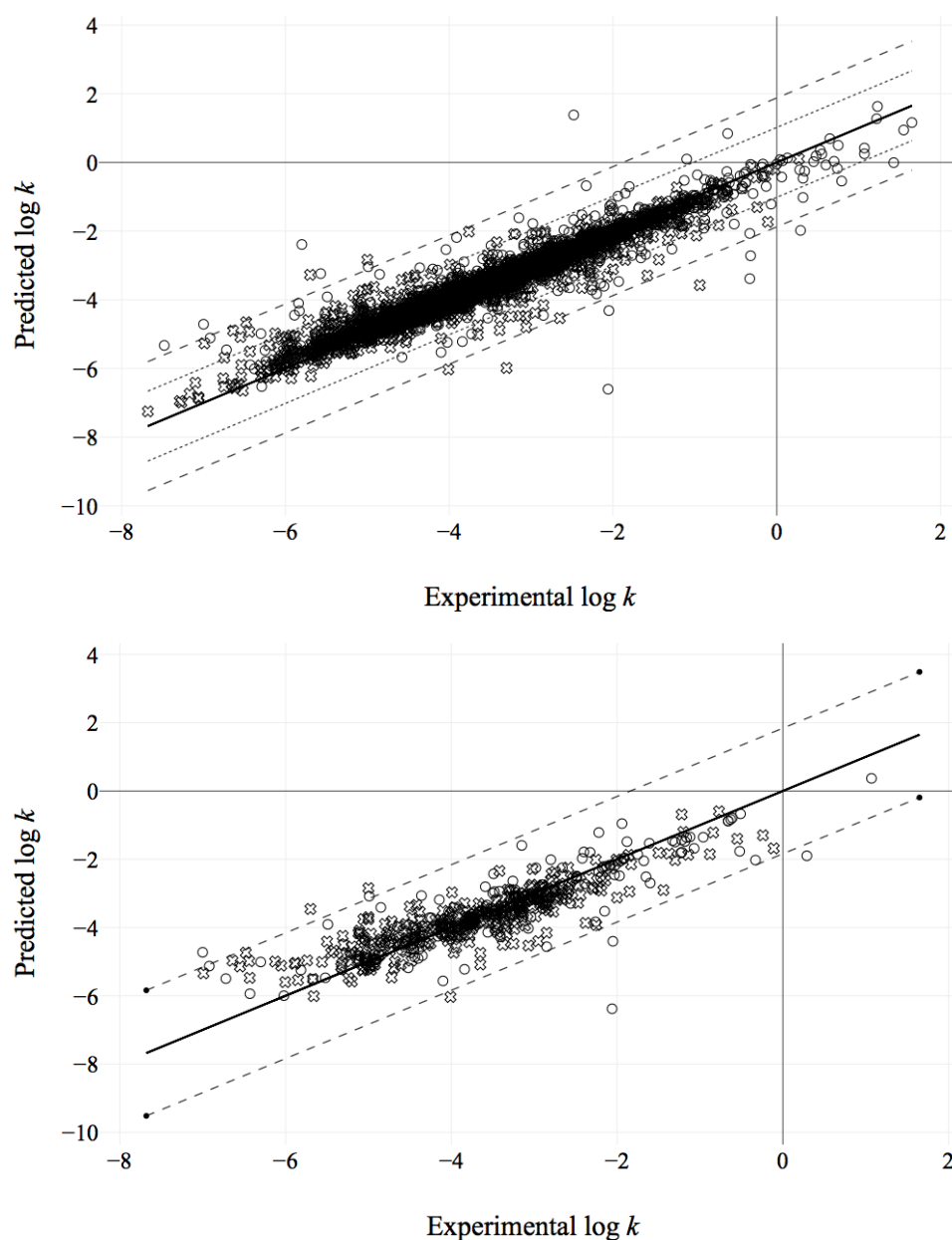


Figure 6. Performance of the global model in cross-validation: predicted vs experimental $\log k$ values for the entire data set (top) and for the for unique data points subset (bottom). Solid lines correspond to perfect predictions, dotted lines specify margin with values predicted within $3 \cdot RMSE$, dashed lines specify margin with values predicted within $3 \cdot RMSE_{UDP}$, crosses and circles represent reactions with neutral and anionic nucleophile reactions correspondingly. Examples of outliers are described in Table 2

In order to tackle this problem, the model's performance was assessed only on unique data points (UDP), i.e., reactions studied under one sole condition. Thus, predicted $\log k$ for 551 unique data points corresponding to 202 reactions with anion nucleophiles and 349 reactions with neutral nucleophiles were selected from cross-validation results obtained for the entire modeling set. The accuracy of predictions on UDP reactions was close to the experimental error: $\text{RMSE}_{\text{UDP}}=0.61$ and $Q^2_{\text{UDP}}=0.75$ (see Figure 6 *bottom*).

Examination of data points for which difference between predicted and experimental values exceeds $3 \cdot \text{RMSE}_{\text{UDP}}$ revealed several outliers, typical examples of which are shown in Table 2. They mostly resulted from (i) experimental data inconsistency, (ii) chemical complexity or (iii) some limitations of fragment descriptors.

Outliers **1** and **2** are typical examples of data inconsistency. Reaction **1** was carried out in phenyl-ethanol at 201 °C. Its experimental $\log k$ value doesn't follow expected 1/T dependence observed for the series of measurements at different temperatures: $\log k = -4.09$ (90 °C), -3.81 (100 °C), -3.68 (105 °C), -3.58 (110 °C), -3.42 (115 °C). Predicted $\log k = -1.36$ much better follows the 1/T trend than the experimental rate constant (-3.21).

Non-continuous variation of $\log k$ as a function of solvent mixture composition has been detected for reaction **2** carried out in pure DMSO. Analysis of the training set data shows that the rate constant increases with DMSO percentage in mixture with water: $\log k = 1.07$ (81% DMSO), 0.06 (65%), -1.01 (46%), -2.08 (30%), -2.93 (18%), -3.66 (8%), -4.09 (2%). Thus, $\log k = -2.48$ reported for pure DMSO looks too small and is clearly out of this trend.

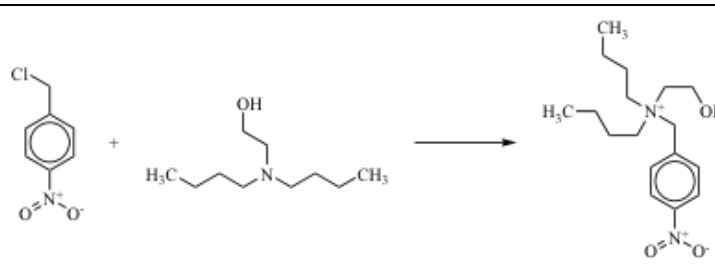
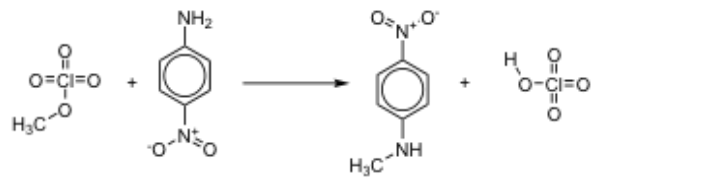
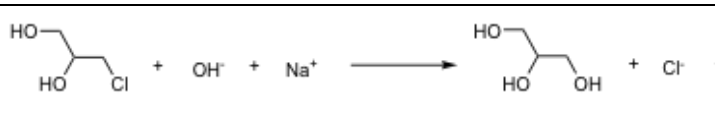

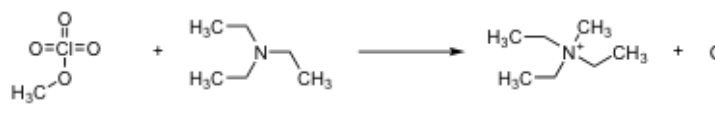

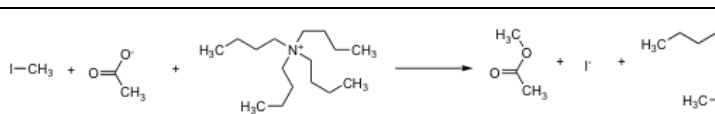
Some complex structural effects were not learned by the model due to the lack of examples in the modeling set. Thus, in para-nitroaniline (a nucleophile in reaction **3**) direct polar conjugation of amino- and nitro-groups in para-positions significantly reduces the nucleophilicity of NH_2 -group. Since training set didn't contain any close analogues of para-nitroaniline, the model overestimated $\log k$ (see Table 2). Similarly, large experimental $\log k$ for reaction **4** resulted from anchimeric assistance of OH group in β -position was not learned by the model due to the lack of data.

Reactions involving reagents with rarely occurred structural patterns can hardly be well predicted. This is a case of reaction **5** in which the substrate bears $\text{SO}_2\text{CH}_2\text{Cl}$ group, unique in the training set, and reaction **6** representing a rare transformation of tertiary to quaternary amine.

For the reactions involving small size reactants only small fragments are generated. These fragments are, sometimes, very common for the training set objects and therefore they are not

allowed to distinguish different classes of reactants. Thus this is not surprising the model fails to predict $\log k$ correctly for these reactions (see items 7, 8 and 9).

Table 2. Example of outliers detected in cross-validation.

N	Reaction	Experimental conditions	$\log k_{\text{exp}}$	$\log k_{\text{pred}}$
1		phenyl-ethanol, 201 °C	-3.21	-1.36
2	$\text{I-CH}_3 + \text{OH}^- + \text{Na}^+ \longrightarrow \text{-OH} + \text{I}^- + \text{Na}^+$	DMSO, 25 °C	-2.48	1.33
3		methanol, 0 °C	-5.00	-2.85
4		water, 0 °C	-2.06	-6.38
5		1,4-dioxane, 50 °C	-2.05	-4.40
6		methanol, 0 °C	-1.58	-3.47
7	$\text{I-CH}_2\text{Br} + \text{H}_3\text{C-O}^- + \text{Na}^+ \longrightarrow \text{Br-CH}_2\text{-O-CH}_3 + \text{I}^- + \text{Na}^+$	methanol, 50 °C	-4.99	-3.08
8		water, 25 °C	-5.80	-2.38
9		DMFA, 0 °C	0.29	-1.90

6. Local vs global models

The question arises whether the models built on selected subsets (“local” models) perform better than the “global” consensus model reported in previous section? To answer this question, the local models were built on the subsets containing only reactions with neutral and anionic nucleophiles or reactions proceeding in particular solvents (nitrobenzene, methanol, ethanol, acetone, water, benzene). The modeling workflow was similar to that described in section 2.5 and only unique data points (UDP) were used. Obtained results show that the local models built on “nucleophiles” subsets performed similarly to the global model: $RMSE_{UDP}(\text{local}) = 0.72$ and 0.59 and $RMSE_{UDP}(\text{global}) = 0.68$ and 0.57 for anionic and neutral nucleophiles, respectively.

The models built on the subsets corresponding to particular solvents involved fragment and temperature descriptors. Results given in Figure 8 show that for 4 solvents (nitrobenzene, ethanol, water and benzene) RMSE values of global and local models are similar. On the other hand, global models performed better for acetone and worse in methanol. Notice that accuracy of predictions vary as a function of solvent type: prediction error observed for nitrobenzene and ethanol subsets is smaller than for other solvents.

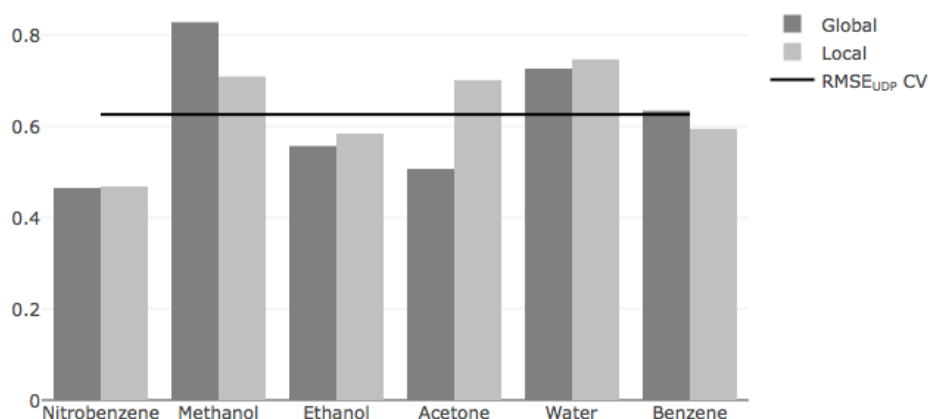


Figure 8. Cross-validated RMSE of global and local models on the subsets corresponding to particular solvents.

7. Consensus model validation on the external test set

An external data set containing 105 Menshutkin reactions was collected from the papers published in 1990-2010 and it didn't overlap with the training set reactions. Prediction performance of the models on this test set was slightly worse ($RMSE=0.8$ and $R^2=0.64$) than that observed in cross validation for UDP.

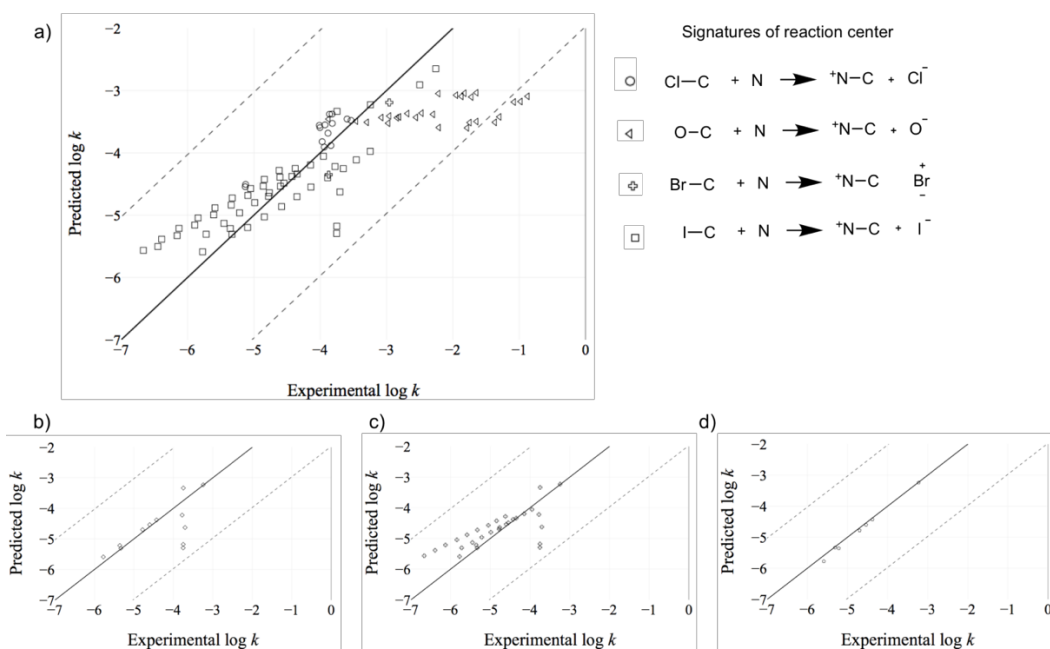
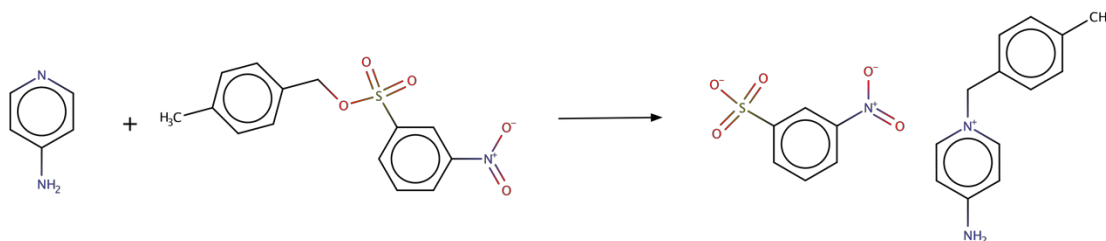


Figure 9. Validation of the global model on the external set: predicted vs experimental $\log k$. Solid line corresponds to perfect prediction. Dotted lines are located by $3 \times \text{RMSE}_{\text{UDP}}$ away from the perfect prediction line. (a) Datapoints are labelled according to reaction signature types. (b) Only reactions having reaction signatures similar to training set ones are shown. (c) only reactions within bounding box AD for, at least, one individual model are retained. (d) only reactions for which 50% of individual models were retained by the bounding box AD are shown

All detected outliers contain substituted phenylsulphonate leaving group (see example on Scheme 2). Their experimental $\log k$ values varying from -3.5 to -0.8 are much larger than predicted $\log k$ varying from -3 to -4. This could be explained by the fact that for 34 training set reactions of similar type, $\log k$ never exceeded -2.5.



Scheme 2. Example of $\text{S}_{\text{N}}2$ reactions with the substituted phenylsulphonate leaving group

In order to improve the models performance on the external set, several applicability domain definitions were examined (see section 2.5.3). Results how given in Table 3 that none of them are efficient. Indeed, the “wide” signature control AD discards no test set reactions and, therefore, it doesn’t impact the models statistical parameters. The “narrow” signature control and “weak” consensus control lead to significant decrease of both prediction performance and the data coverage. On the other hand, the “firm” consensus control significantly improves the

model's performance. However, it covers only 7.5% of data which doesn't allow us to recommend this AD in further applications.

Table 3. Performance of consensus model on the external test set as a function of Applicability Domain

AD definition	R ²	RMSE	Data coverage (%)
“wide” signature control	0.64	0.8	100
“narrow” signature control	0.22	0.69	13
“firm” consensus control	0.98	0.1	7.5
“weak” consensus control	0.5	0.61	32.3

Conclusions

Representation of a chemical transformation by a single Condensed Graph of Reaction opens a possibility to apply to chemical reactions mining variety of chemoinformatics approaches developed for individual molecules. In particular, this concerns invented in this work the Matched Reactions Pairs approach (an analogue of well-known Matched Molecular Pairs) which being applied to CGRs provides an interesting insight into the substituents effects on chemical reactivity. Being applied to a set of S_N2 reactions, this method helped us to detect transformations with late transition state and great charge separation.

A CGR can be encoded by fragment descriptors which, in turn, can serve as an input in the tools of data visualization and modeling. If necessary, these descriptors can be concatenated with the parameters describing experimental conditions. Here, this approach has successfully been used to visualize large set of SN2 reactions using Generative Topographic Method and to develop predictive SVR models for the logarithm of reaction rate. The performance of the developed consensus model in cross-validation (RMSE_{UDP}=0.61 log*k* units) and on the external test set (RMSE=0.8) is not far from the experimental error (0.5 log*k* units). Four different definitions of model's applicability domain for reactions have been examined, but none can be recommended because of either low ability to discard the outliers or low data coverage.

The developed consensus SVR model is freely available for the users on our server cimm.kpfu.ru.

Supporting Information contains some information about implementation of SVR consensus model on our server <http://cimm.kpfu.ru>

Acknowledgement. This study was supported by Russian Science Foundation, grant No 14-43-00024. TG thanks the IDEX program of the University of Strasbourg for the PhD fellowship.

References

1. Kubinyi H (2008) QSAR: Hansch Analysis and Related Approaches. VCH, Weinheim
2. 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* 19:693–703. doi: 10.1007/s10822-005-9008-0
3. Gimadiev TR, Madzhidov TI, Marcou G, Varnek A (2016) Generative Topographic Mapping Approach to Modeling and Chemical Space Visualization of Human Intestinal Transporters. *Bionanoscience* 6:464–472. doi: 10.1007/s12668-016-0246-5
4. Madzhidov TI, Gimadiev TR, Malakhova DA, et al (2017) Structure–reactivity relationship in Diels–Alder reactions obtained using the condensed reaction graph approach. *J Struct Chem*. doi: 10.1134/S0022476617040023
5. Madzhidov TII, Bodrov AV V., Gimadiev TRR, et al (2015) Structure-reactivity relationship in bimolecular elimination reactions based on the condensed graph of a reaction. *J Struct Chem* 56:1227–1234. doi: 10.1134/S002247661507001X
6. Gimadiev TRR, Madzhidov TII, Nugmanov RII, et al (2018) Assessment of tautomer distribution using the condensed reaction graph approach. *J Comput Aided Mol Des* 32:401–414. doi: 10.1007/s10822-018-0101-6
7. Lin AI, Madzhidov TI, Klimchuk O, et al (2016) Automatized Assessment of Protective Group Reactivity: A Step Toward Big Reaction Data Analysis. *J Chem Inf Model* 56:2140–2148. doi: 10.1021/acs.jcim.6b00319
8. Leach AG, Jones HD, Cosgrove DA, et al (2006) Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J Med Chem* 49:6672–6682. doi: 10.1021/jm0605233
9. Smith MB, March J (2007) March's advanced organic chemistry: reactions, mechanisms, and structure, Sixth edit. John Wiley & Sons, Hoboken
10. Kravtsov AA, Karpov P V, Baskin II, et al (2011) Prediction of Rate Constants of SN2 Reactions by the Multicomponent QSPR Method. *Dokl Chem* 440:299–301.
11. Hoonakker F, Lachiche N, Varnek A, Wagner A (2011) Condensed Graph of Reaction: considering a chemical reaction as one single pseudo molecule . *Int J Artif Intell Tools* 20:253–270.
12. Madzhidov TI, Polishchuk PG, Nugmanov RI, et al (2014) Structure-reactivity relationships in terms of the condensed graphs of reactions. *Russ J Org Chem* 50:459–463.
13. Kuz'min VE, Artemenko AG, Muratov EN (2008) Hierarchical QSAR technology based

- on the Simplex representation of molecular structure. *J Comput Aided Mol Des* 22:403–421. doi: DOI 10.1007/s10822-008-9179-6
14. Nugmanov RI, Madzhidov TI, Khaliullina GR, et al (2014) Development of “structure-property” models in nucleophilic substitution reactions involving azides. *J Struct Chem* 55:1026–1032. doi: 10.1134/S0022476614060043
 15. ChemAxon. (2015) InstantJChem 15.7.27.0.
 16. ChemAxon. (2015) Standardizer, JChem 15.8.3.0.
 17. Madzhidov TI, Gimadiev TR, Malakhova DA, et al (2017) Structure-reactivity relationship in Diels-Alder reactions obtained using the condensed reaction graph approach. *J Struct Chem*. doi: 10.15372/JSC20170402
 18. Muller C, Marcou G, Horvath D, et al (2012) Models for identification of erroneous atom-to-atom mapping of reactions performed by automated algorithms. *J Chem Inf Model* 52:3116–3122.
 19. Dalby A, Nourse JG, Hounshell WD, et al (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J Chem Inf Model* 32:244–255. doi: 10.1021/ci00007a012
 20. EPAM Systems. (2015) Indigo.
 21. Madzhidov TI, Nugmanov RI, Gimadiev TR, et al (2015) Consensus approach to atom-to-atom mapping in chemical reactions. *Butlerov Commun* 44:170–176.
 22. Hussain J, Rea C (2010) Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J Chem Inf Model* 50:339–348. doi: 10.1021/ci900450m
 23. Sushko I, Novotarskyi S, Körner R, et al (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* 25:533–554. doi: 10.1007/s10822-011-9440-2
 24. Bishop CM, Svensén M, Williams CKII (1998) GTM: The Generative Topographic Mapping. *Neural Comput* 10:215–234. doi: 10.1162/089976698300017953
 25. Gaspar HA, Marcou G, Horvath D, et al (2013) Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J Chem Inf Model* 53:3318–3325. doi: 10.1021/ci400423c
 26. Horvath D, Marcou G, Varnek A (2017) Generative Topographic Mapping Approach to Chemical Space Analysis. pp 167–199
 27. Catalán J, López V, Pérez P, et al (1995) Progress towards a generalized solvent polarity scale: The solvatochromism of 2-(dimethylamino)-7-nitrofluorene and its homomorph 2-

- fluoro-7-nitrofluorene. *Liebigs Ann* 1995:241–252. doi: 10.1002/jlac.199519950234
28. Catalán J, Díaz C (1997) A Generalized Solvent Acidity Scale: The Solvatochromism of *tert*-Butylstilbazolium Betaine Dye and Its Homomorpho, *o'*-Di-*tert*-butylstilbazolium Betaine Dye. *Liebigs Ann* 1997:1941–1949. doi: 10.1002/jlac.199719970921
 29. Catalán J, Díaz C, López V, et al (1996) A Generalized Solvent Basicity Scale: The Solvatochromism of 5-Nitroindoline and Its Homomorph 1-Methyl-5-nitroindoline. *Liebigs Ann* 1996:1785–1794. doi: 10.1002/jlac.199619961112
 30. Taft RW, Kamlet MJ (1976) The solvatochromic comparison method. 2. The α -scale of solvent hydrogen-bond donor (HBD) acidities. *J Am Chem Soc* 98:2886–2894. doi: 10.1021/ja00426a036
 31. Kamlet MJ, Taft RW (1976) The solvatochromic comparison method. I. The β -scale of solvent hydrogen-bond acceptor (HBA) basicities. *J Am Chem Soc* 98:377–383. doi: 10.1021/ja00418a009
 32. Kamlet MJ, Abboud JL, Taft RW (1977) The solvatochromic comparison method. 6. The π^* scale of solvent polarities. *J Am Chem Soc* 99:6027–6038. doi: 10.1021/ja00460a031
 33. Mathea M, Klingspohn W, Baumann K (2016) Chemoinformatic Classification Methods and their Applicability Domain. *Mol Inform* 35:160–180. doi: 10.1002/minf.201501019
 34. Varnek A, Fourches D, Horvath D, et al (2008) ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr Comput Aided-Drug Des* 4:191–198. doi: 10.2174/157340908785747465
 35. Tetko I V, Sushko I, Pandey AK, et al (2008) Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* 48:1733–46. doi: 10.1021/ci800151m