# Cross-validation strategies in QSPR modelling of chemical reactions

Assima Rakhimbekova[a], Tagir N. Akhmetshin[a,b], Guzel I. Minibaeva[a], Ramil I. Nugmanov[a], Timur R. Gimadiev[c], Timur I. Madzhidov[a]*, Igor I. Baskin[a,d] and Alexandre Varnek[c,d]

[a] *A.M. Butlerov Institute of Chemistry, Kazan Federal University, Kazan, Russia;* [b] *Laboratory of Chemoinformatics, UMR 7140 CNRS, University of Strasbourg, Strasbourg, France;* [c] *Institute for Chemical Reaction Design and Discovery, Hokkaido University, Sapporo, Japan;* [d] *Faculty of Physics, Moscow State University, Moscow, Russia*

*T.I. Madzhidov Timur.Madzhidov@kpfu.ru

# Cross-validation strategies in QSPR modelling of chemical reactions

In this article, we consider cross-validation of the quantitative structure-property relationship models for reactions and show that the conventional k-fold cross-validation (CV) procedure gives an 'optimistically' biased assessment of prediction performance. To address this issue, we suggest two strategies of model cross-validation, 'transformation-out' CV, and 'solvent-out' CV. Unlike the conventional k-fold cross-validation approach that does not consider the nature of objects, the proposed procedures provide an unbiased estimation of the predictive performance of the models for novel types of structural transformations in chemical reactions and reactions going under new conditions. Both the suggested strategies have been applied to predict the rate constants of bimolecular elimination and nucleophilic substitution reactions, and Diels-Alder cycloaddition. All suggested cross-validation methodologies and tutorial are implemented in the open-source software package CIMtools (https://github.com/cimm-kzn/CIMtools).

Keywords: validation; QSPR; chemical reactions; rate constant prediction; reaction rate; structure-reactivity modelling

## Introduction

Nowadays the external validation is considered as an integral component of any Quantitative Structure-Activity/Property Relationship (QSAR/QSPR) model, irrespective of the nature of the chemical objects under investigation [1–10]. It was several times pointed out that the correct validation of QSAR/QSPR models is essential [1–3,5,6,11]. It is common to distinguish internal validation, which is used for model selection, and external validation, used for the quality assessment [12]. In most cases, external validation is performed on a dedicated test set [2,11]. Its drawback is possible to bias due to random fluctuations, arbitrary or unfair selection of molecules to test set, which can be overcome by the rational division of the data set into training and test sets

[13–18]. Tetko et al [3] proposed external cross-validation as a more stable alternative to single the external test set. In this procedure, a part of the parent dataset is randomly placed to the external (outer) set used for assessing the predictive performance of the approach, while remaining objects are used for the model building including hyperparameters selection based on internal cross-validation. Outer set loops over the whole data set guarantee involvement of all data points in external prediction. In machine-learning, this technique is also known as a nested cross-validation. Since an external (outer) test set is by no means used for model building, it is free from model selection bias [3,19] and, hence, can be used for assessment of model performance.

Data sets of complex chemical objects, such as chemical reactions, polymers, compounds' mixtures often include similar elements (e.g., the same reactant participating in different reactions). Therefore, correct external validation of QSPR models for these objects is much less straightforward compared to 'classical' QSAR/QSPR models for individual molecules [7–9,20,21]. Each of these cases requires designing a specific validation strategy. For the models predicting properties of binary mixtures, 'Mixtures Out', 'Compounds Out', 'Points Out', 'Everything Out' cross-validation strategies have been suggested [7,8]. The 'Donor Out', 'Acceptor Out', and 'Both out' strategies have been successfully applied to validate QSPR models for dissociation free energy of H-bond donor-acceptor complexes [9].

Nowadays, the growing attention is attracted to statistical models predicting the kinetic or thermodynamic properties of chemical reactions [22–24]. Chemical reactions represent a complex object because they involve several molecular species of two types (reactants and products) and their properties depend on experimental conditions (solvent, catalyst, temperature, etc). In most of the studies, a random split or cross-validation (CV) techniques were used for the validation of models for chemical reactions [25–33].

However, recently we published some studies demonstrating the flaws of conventional validations techniques for reaction characteristics modelling [20,21]. We demonstrated on the bimolecular rate constant data set that the error estimated in conventional CV (RMSE = 0.3 log*k* units) was lower than both estimated experimental error (about 0.5-1.0 log*k* units) and that computed for the external test set [20]. This can be explained by the fact that the data set included the same structural transformations of reactants to products studied under slightly different conditions. Since the reaction rate values of these reactions were often very close, they behaved as duplicates. Their presence in both training and test sets in CV loops inevitably leads to very optimistic estimates of predictive ability [20,21]. To overcome this problem, the model's performance was assessed using a subset of reactions studied under one sole condition (called unique data points, UDP) for which such kind of bias is not possible [20]. However, such validation is not a panacea: its results strongly depend on the fraction of UDP in the data set, which can potentially vary from 0 (all reactions were studied at different conditions) to 100 percent (all reactions were studied at one condition only). Thus, such type of validation has limited applicability.

Polishchuk et al. [21] suggested the 'product-out' cross-validation strategy in which all reactions with the same main product are placed either in the training or test set for a particular fold. It has been shown that the ranking of models on different descriptors based on "product-out" validation is significantly different in comparison with conventional cross-validation called 'reaction-out' CV.

Validation for QSAR modelling means the assessment of the predictive power of the model on novel datapoints. For molecular QSAR/QSPR studies, a test set on each cross-validation fold consists of molecules absent in the training set. However, the novelty issue is more complex in the case of reactions modelling and should account for both either

chemical transformation or experimental conditions. Therefore, here we propose two validation strategies that assess predictive performance for a particular novelty type:

- "transformation-out" which estimates the ability to predict characteristics of chemical reactions with a novel reactant-product pair (hereafter we call it transformation),
- "solvent-out" which assesses the quality of prediction for reactions proceeding in a new solvent.

Note that in "transformation-out" validation reactions of the test set proceed under the same conditions the reactions of the training set. In "solvent-out" validation, the test set includes chemical transformations present in the training set, but solvents are different. The presence of completely new reactions, having both new transformations and solvents is more challenging for the model. In principle, such data points can be used for "both-out" validation. However, they were very scarce or absent in the considered data sets.

As reactions rates are often measured under several conditions and at several temperatures and the number of unique transformations is much lower than the number of reactions (see <Place for Table ), in conventional cross-validation, the test set reactions may have close neighbors in the training set which explains its higher $Q^2$ and lower RMSE compared to "transformation-out" and "solvent-out" validation.

Proposed validation methodologies have been applied to the modelling of the logarithm of $S_N2$, E2, and Diels-Alder reactions rate constants (log$k$). They have been implemented in the open-source CIMtools software package (https://github.com/cimm-kzn/CIMtools) developed for reaction modelling.

## Materials and methods

### *Data set and descriptors*

Three data sets for the following types of reactions were used in this study: bimolecular nucleophilic substitution ($S_N2$), bimolecular elimination (E2), Diels-Alder reactions (DA). These data sets were collected in our previous studies [20,30,31]. An external data set of 90 Menshutkin reactions ($S_N2$) were also used. The data set was collected and curated in our previous study [20]. Note that this data set comprised novel transformations and the same solvents as in the training set. Thus, it corresponds closely to "transformation-out" validation strategy. The data sets characteristics are presented in <Place for Table .

<Place for Table 1>

*Descriptors of the reactions.* The descriptor vector for each reaction was resulted from a concatenation of structural descriptors and parameters describing experimental conditions (solvent and temperature) as proposed in paper [34]. The chemical transformations 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 [35,36]. Thus, CGR represents the whole transformation, i.e. reactant-product pair, as a single molecular graph. CGRtools library was used to generate CGRs [37]. ISIDA fragment descriptors were computed for CGR using the ISIDA Fragmentor [38] 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, sequences

of atoms and bonds containing from 2 to 4 atoms were considered. This fragmentation was successfully used in our previous modelling studies [20,30,31,33].

*Descriptors of the reaction conditions*. Each solvent was described by 15 descriptors that represent polarity, polarizability, H-acidity, and basicity: Catalan SPP [39], SA [40], and SB constants [39], Camlet–Taft constants α [41], β [42], and π* [43], four functions depending on the dielectric constant, three functions depending on the refractive index as shown in paper [34]. 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).

### *Model building and validation*

*The general procedure of modelling*. The models were built using Random Forest (denoted as RF) approach implemented in the scikit-learn library [44]. 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 rest 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, three strategies of such split have been tested: conventional random 5-fold cross-validation (denoted as ‘reaction-out’ CV), ‘transformation-out’ CV, and ‘solvent-out’ CV

(see below). A hyperparameter (the number of features to consider when looking for the best split) of the Random Forest regression 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 (determination coefficient, $R^2$, corresponding to $Q^2_{F2}$ formulae in [5], and RMSE) were calculated. Notice, that performance on 'reaction-out' CV, 'transformation-out' CV, and 'solvent-out' CV reflects predictive ability on the external test set.

*Strategies for Model External Validation in QSPR for reaction datasets.* Three types of cross-validation strategies were applied. In the cross-validation procedure, the initial data set was divided into a given number of subsets, corresponding to the desired number of folds. Each subset was sequentially considered as an external test set, while the rest was used as the modelling set. The schematic representation of 'reaction-out', 'transformation-out', and 'solvent-out' CV is shown in

<Place for Figure 1>

.

'Reaction-out' CV approach was simply a regular five times repeated five-fold CV. The sizes of test sets in it are almost equal.

'Transformation-out' CV approach was implemented as a *k*-fold cross-validation (

<Place for Figure 1>

). In the ‘transformation-out’ procedure, all reactions having the same CGR were placed into the same fold (different shapes in

<Place for Figure 1>

). The implemented splitting algorithm tried to make the folds approximately equal in size. Therefore, a fold might contain a group of several CGRs (as presented in fold 2,

<Place for Figure 1>

). Moreover, the test set contained reactions proceeding in the solvents presented also in the training set, see

<Place for Figure 1>

. This allowed to avoid a bias due to the presence of new solvents in the test set. Thus, unique reactions (corresponding to a unique combination of transformation and solvent) were always placed into the training set, see

<Place for Figure 1>

. Since in ‘transformation-out’ validation all reactions having the same reactants and products were placed into the same subset (

<Place for Figure 1>

), this only showed how well reactions with new reactants and products, i.e. CGRs, were predicted.

<Place for Figure 1>

'Solvent-out' validation approach was implemented as a leave-one-solvent-out (

<Place for Figure 1>

), as the number of solvent types is usually low (less than 50). An additional hurdle to the application of k-fold validation was caused by a great imbalance in the number of reactions corresponding to one solvent. In 'solvent-out' validation, all reactions carried out in the same solvent were placed into the same subset which is sequentially used as the test set. Each reaction in the test set should have a counterpart in the training set with the same CGR but proceed in a different solvent. Unique reactions (in this case, reactions measured in a single solvent) were always included in the training set and never used in the test set. This method avoids underestimating of the model performance, because such reactions represent new CGR and new solvent for the trained model. It is worth noting that in this study we have grouped the folds by solvents only, but the developed algorithm can group following several conditions.

**Results and discussion**

The models were built using Random Forest and fragment descriptors of CGRs and validated using three strategies described above, i.e., 'reaction-out' CV, 'transformation-out' CV, and 'solvent-out' CV. The performance of models for three data sets obtained using 'reaction-out' CV, 'transformation-out' CV, and 'solvent-out' CV strategies are shown in **Erreur ! Source du renvoi introuvable.**. In all three cases, the prediction performance metrics for 'reaction-out' CV validation were better than for 'transformation-out' CV. In 'reaction-out' CV strategy reactions having the same reactants and products were simultaneously present in both training and test set. Two reactions with very similar conditions can be present in the training and test set, and thus

the prediction performance is overoptimistic. In ‘transformation-out’ CV strategy, all reactions with the same structural transformation were present in either the training or the test sets, but not both simultaneously. Therefore, RMSE values are bigger than for ‘reaction-out’ CV strategy (**Erreur ! Source du renvoi introuvable.**), but they remain on the acceptable levels.

<Place for Table 2>

The independent external set consisting of Menshutkin reactions ($S_N2$) was then used for model validation as well. Prediction on the external set of reactions was used for comparison with ‘reaction-out’ and ‘transformation-out’ validation strategies. The data set contained 48 Menshutkin reactions which had new transformations in comparison with the training set, but were carried out in known solvents (corresponding to the training set). As expected, the results on the ‘transformation-out’ validation are close to the one of external validation ($R^2$ = 0.51, RMSE = 0.99 log*k* units on the external test set), confirming that ‘transformation-out’ validation is a more rigorous and unbiased approach for validating QSPR models of reactions, and, therefore, reliably assesses the accuracy of predictions for novel structural transformations of reactions.

The prediction performance metrics in ‘solvent-out’ validation were quite different in all three cases. Reasonable statistical parameters were observed for the Diels-Alder reactions data set ($R^2$ = 0.76 and RMSE = 0.94 log*k* units, respectively). Surprisingly, 'solvent-out' validation for bimolecular nucleophilic substitution reaction and bimolecular elimination reaction data sets led to much worse statistical parameters than for DA data set. We suggest that such results in the case of Diels-Alder reactions are due to the fact that most of the reactions in the data set (86%) were carried out in non-polar solvents, moreover, 74% of the reactions were carried out in toluene, benzene, and chlorobenzene (53%, 11%,

and 10%, respectively). In the remaining data sets, the reactions were carried out in various solvents of different nature. Moreover, since these reactions have a non-polar transition state, the solvent effect should affect its rate much.

The obtained results in 'solvent-out' external validation on the bimolecular nucleophilic substitution reaction data set are discussed in more detail. The data set of $S_N2$ reactions consists of 43 different solvents where the most popular solvents were ethanol (17.9%), methanol (15.4%), nitrobenzene (13.7%), acetone (12.7%) often used in mixtures with water - more than half the database (59.7%), Figure 2. The total percentage of all reactions carried out in the rest 31 solvents that aren't shown in the diagram (Figure 2) was only 9.1%.

<Place for Figure 2>

The number of reactions carried out in particular solvents, which were used as a test set in 'solvent-out' validation, ranged from 1 to 422. Note that in 'solvent-out' validation the reactions are not included in the test set if there is no data on rate constant for the same transformations in other solvents. The values of RMSE for all solvents (folds of 'leave-one-solvent-out' external CV) are given in **Erreur ! Source du renvoi introuvable.** and discussed below. The same tables for DA and E2 data sets are given in Supplementary Materials (see Tables S1 and S2 in Supplementary materials).

<Place for Table 3>

Generally, we found no clear dependence between solvent type, use of water-organic mixtures as a solvent, size of the data set, and the prediction of RMSE for $S_N2$ reactions. Solvents, for which a lot of rate data exist, are characterized by a larger RMSE (from 0.59 to 1.52). At the same time, a poor prediction is common also for the least popular solvents. Among 8 solvents that are often used in mixtures with water 5 (methanol, ethanol,

dimethyl sulfoxide, acetone, acetonitrile) are poorly predicted but 3 (dioxane, sulfolane, oxolane) are predicted well. Prediction error in alcohols has some trend: heavy alcohols (more than 2 carbons) have low values of RMSE (from 0.09 to 0.45), ethanol has a greater error (0.590) and methanol have the greatest (0.82). It is worth noting that water has an even greater error (about 1.02).

We hypothesize that such specificity of 'solvent-out' validation can be explained by insufficient generalization ability of the RF approach, probably due to the application of a large number of weakly informative structural descriptors and a few highly informative solvent descriptors. To support this hypothesis, the descriptors used for trees' branching were analysed, Figure 3. One can see that RF pays a lot of attention to solvent descriptors: they are selected for branching in about 20-30% of trees, depending on the level of the tree (Figure 3), while solvent descriptors constitute only 5% of the descriptor set. So, poor prediction of solvents cannot be explained by ignorance of solvent features by the model. It can be seen from Figure 3 that RF is more likely to select solvent descriptors as branching criteria at first and second level of tree (i.e., at the root node of the decision tree or right after it) and to use fragment descriptors more frequently in levels 3 and lower (Figure 3).

<Place for Figure 3>

We believe such a selection of descriptors can be explained as RF tends to internally find correlations within a particular solvent type based on structural descriptors of transformation, creating a kind of family of submodels for each particular solvent type(s). Thus, model does not try to generalize solvent influence, and RF predicts reactions in new solvents poorer than with new transformations. It also explains the mentioned specificity of alcohols prediction. Heavy alcohols are predicted similarly by the same

implicit RF submodel based on the rest of alcohols but since methanol and ethanol are quite different from the rest, their prediction is poorer.

Our test on other machine-learning methods (support vector machines, neural nets) and other fragment descriptors showed similar results: models had a rather poor ‘solvent-out’ validation performance. It seems that the problem lies mainly in the application of fragment descriptors, but it would be a subject of specific research.

**Conclusion**

Validation of QSPR models is a very important aspect to understand the reliability of the models for the prediction of new objects not present in the data set. In the case of chemical reactions, test set objects can have new structural transformations or proceed under new conditions (new solvents, new additives, new catalysts, etc.). Thus, different types of novelties can be considered. Model performance on conventional cross-validation that does not have control over the constitution of the test set has unclear meaning in the case of chemical reaction modelling. To evaluate the ability of the model predicting property of chemical reaction, two strategies of model validation, ‘transformation-out’ and ‘solvent-out’, have been suggested. ‘Transformation-out’ validation provides an estimation of the predictive performance of the models for novel types of structural transformations in chemical reactions and ‘solvent-out’ CV – for reactions going under new solvents. It is worth noting that ‘solvent-out’ validation is a special case of 'condition-out' validation, i.e., we can group folds not only by solvents but also by other conditions (new additives, new catalysts, a combination of several conditions). It has been shown that performance on ‘transformation-out’ validation is similar to the external set one containing new transformations. Thus, the proposed two validation strategies are recommended to be used for an unbiased evaluation of reaction-property models.

On bimolecular nucleophilic substitution reaction and bimolecular elimination reaction data sets, we revealed that the reaction-property models better predict the rate constant for new structural transformations than the rate constants for reactions occurring in novel solvents. It was explained by the inability of applied machine learning methods (RF, support vector machines, neural nets) and different fragment descriptors to correctly generalize dependency of the reaction rate for wide solvent types. But it will require specific research that we plan to be done in the nearest future.

Proposed validation strategies and tutorial are implemented into open-source CIMtools library for structure-property modelling available on GitHub (https://github.com/cimm-kzn/CIMtools).

**Funding**

This work was supported by Russian Science Foundation grant No 19-73-10137.

**ORCID**

A. Rakhimbekova https://orcid.org/0000-0002-6820-6385

T.N. Akhmetshin https://orcid.org/0000-0002-2549-6431

G. I. Minibaeva https://orcid.org/0000-0001-7964-4842

R.I. Nugmanov https://orcid.org/0000-0002-8541-9681

T.R. Gimadiev https://orcid.org/0000-0001-5012-0308

T.I. Madzhidov https://orcid.org/0000-0002-3834-6985

I.I. Baskin https://orcid.org/0000-0003-0874-1148

A. Varnek https://orcid.org/0000-0003-1886-925X.

**References**


[1] M. Balls, B.J. Blaauboer, J.H. Fentem, L. Bruner, R.D. Combes, B. Ekwall, R.J. Fielder, A. Guillouzo, R.W. Lewis, D.P. Lovell, C.A. Reinhardt, G. Repetto, D. Sladowski, H. Spielmann and F. Zucco, *Practical Aspects of the Validation of Toxicity Test Procedures*, Altern. to Lab. Anim. 23 (1995), pp. 129–146.

[2] A. Tropsha, P. Gramatica and V. Gombar, *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models*, QSAR Comb. Sci. 22 (2003), pp. 69–77.

[3] I. V. Tetko, V.P. Solov'ev, A. V. Antonov, X. Yao, J.P. Doucet, B. Fan, F. Hoonakker, D. Fourches, P. Jost, N. Lachiche and A. Varnek, *Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure−Property Relationship Studies of Metal Complexation with Ionophores*, J. Chem. Inf. Model. 46 (2006), pp. 808–819.

[4] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C.P. Varghese and R.K. Agrawal, *Validation of QSAR Models - Strategies and Importance*, Int. J. Drug Discov. 2 (2011), pp. 511–519.

[5] K. Roy, S. Kar and R.N. Das, *Validation of QSAR Models*, in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Elsevier, 2015, pp. 231–289.

[6] P. Gramatica and A. Sangion, *A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology*, J. Chem. Inf. Model. 56 (2016), pp. 1127–1131.

[7] I. Oprisiu, E. Varlamova, E. Muratov, A. Artemenko, G. Marcou, P. Polishchuk,

V. Kuz'min and A. Varnek, *QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids*, Mol. Inform. 31 (2012), pp. 491–502.

[8] E. Muratov, E. V Varlamova, V.E. Kuzmin, A.G. Artemenko, N.N. Muratov, S. Mileyko, D. Fourches and A. Tropsha, *Everything Out Validation Approach for Qsar Models of Chemical Mixtures*, J Clin Pharm 1 (2014), pp. 1005.

[9] M. Glavatskikh, T. Madzhidov, V. Solov'ev, G. Marcou, D. Horvath and A. Varnek, *Predictive Models for the Free Energy of Hydrogen Bonded Complexes with Single and Cooperative Hydrogen Bonds*, Mol. Inform. 35 (2016), pp. 629–638.

[10] A.O. Aptula, N.G. Jeliazkova, T.W. Schultz and M.T.D. Cronin, *The Better Predictive Model: High q2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set?*, QSAR Comb. Sci. 24 (2005), pp. 385–396.

[11] A. Golbraikh and A. Tropsha, *Beware of q2!*, J. Mol. Graph. Model. 20 (2002), pp. 269–276.

[12] P. Gramatica, *Principles of QSAR models validation: internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.

[13] J. Gasteiger and J. Zupan, *Neural Networks in Chemistry*, Angew. Chemie Int. Ed. English 32 (1993), pp. 503–527.

[14] J. Huuskonen, *QSAR Modeling with the Electrotopological State: TIBO Derivatives*, J. Chem. Inf. Comput. Sci. 41 (2001), pp. 425–429.

[15] I. V. Tetko, V. V. Kovalishyn and D.J. Livingstone, *Volume Learning Algorithm*

*Artificial Neural Networks for 3D QSAR Studies*, J. Med. Chem. 44 (2001), pp. 2411–2420.

[16] M. Snarey, N.K. Terrett, P. Willett and D.J. Wilton, *Comparison of algorithms for dissimilarity-based compound selection*, J. Mol. Graph. Model. 15 (1997), pp. 372–385.

[17] A. Golbraikh, *Molecular Dataset Diversity Indices and Their Applications to Comparison of Chemical Databases and QSAR Analysis*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 414–425.

[18] C. Szántai-Kis, I. Kövesdi, G. Kéri and L. Örfi, *Validation subset selections for extrapolation oriented QSPAR models*, Mol. Divers. 7 (2003), pp. 37–43.

[19] D. Baumann and K. Baumann, *Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation*, J. Cheminform. 6 (2014), pp. 47.

[20] T. Gimadiev, T. Madzhidov, I. Tetko, R. Nugmanov, I. Casciuc, O. Klimchuk, A. Bodrov, P. Polishchuk, I. Antipin and A. Varnek, *Bimolecular Nucleophilic Substitution Reactions: Predictive Models for Rate Constants and Molecular Reaction Pairs Analysis*, Mol. Inform. 38 (2019), pp. 1800104.

[21] P. Polishchuk, T. Madzhidov, T. Gimadiev, A. Bodrov, R. Nugmanov and A. Varnek, *Structure–reactivity modeling using mixture-based representation of chemical reactions*, J. Comput. Aided. Mol. Des. 31 (2017), pp. 829–839.

[22] O. Engkvist, P.-O. Norrby, N. Selmi, Y. Lam, Z. Peng, E.C. Sherer, W. Amberg, T. Erhard and L.A. Smyth, *Computational prediction of chemical reactions:*

*current status and outlook*, Drug Discov. Today 23 (2018), pp. 1203–1218.

[23] H. Gao, T.J. Struble, C.W. Coley, Y. Wang, W.H. Green and K.F. Jensen, *Using Machine Learning To Predict Suitable Conditions for Organic Reactions*, ACS Cent. Sci. 4 (2018), pp. 1465–1476.

[24] I.I. Baskin, T.I. Madzhidov, I.S. Antipin and A.A. Varnek, *Artificial intelligence in synthetic chemistry: achievements and prospects*, Russ. Chem. Rev. 86 (2017), pp. 1127–1156.

[25] D.T. Ahneman, J.G. Estrada, S. Lin, S.D. Dreher and A.G. Doyle, *Predicting reaction performance in C–N cross-coupling using machine learning*, Science. 360 (2018), pp. 186–190.

[26] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *A Structure-Based Platform for Predicting Chemical Reactivity*, Chem 6 (2020), pp. 1379–1390.

[27] J.A. Kammeraad, J. Goetz, E.A. Walker, A. Tewari and P.M. Zimmerman, *What Does the Machine Learn? Knowledge Representations of Chemical Reactivity*, J. Chem. Inf. Model. 60 (2020), pp. 1290–1301.

[28] A.A. Kravtsov, P. V. Karpov, I.I. Baskin, V.A. Palyulin and N.S. Zefirov, *Prediction of rate constants of SN2 reactions by the multicomponent QSPR method*, Dokl. Chem. 440 (2011), pp. 299–301.

[29] A.A. Kravtsov, P. V. Karpov, I.I. Baskin, V.A. Palyulin and N.S. Zefirov, *Prediction of the preferable mechanism of nucleophilic substitution at saturated carbon atom and prognosis of SN1 rate constants by means of QSPR*, Dokl. Chem.

441 (2011), pp. 314–317.

[30] T.I. Madzhidov, A. V. Bodrov, T.R. Gimadiev, R.I. Nugmanov, I.S. Antipin and A.A. Varnek, *Structure–reactivity relationship in bimolecular elimination reactions based on the condensed graph of a reaction*, J. Struct. Chem. 56 (2015), pp. 1227–1234.

[31] T.I. Madzhidov, T.R. Gimadiev, D.A. Malakhova, R.I. Nugmanov, I.I. Baskin, I.S. Antipin and A.A. Varnek, *Structure–reactivity relationship in Diels–Alder reactions obtained using the condensed reaction graph approach*, J. Struct. Chem. 58 (2017), pp. 650–656.

[32] 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, *Structure-reactivity relationships in terms of the condensed graphs of reactions*, Russ. J. Org. Chem. 50 (2014), pp. 459–463.

[33] T.R. Gimadiev, T.I. Madzhidov, R.I. Nugmanov, I.I. Baskin, I.S. Antipin and A. Varnek, *Assessment of tautomer distribution using the condensed reaction graph approach*, J. Comput. Aided. Mol. Des. 32 (2018), pp. 401–414.

[34] 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, *Structure-reactivity relationships in terms of the condensed graphs of reactions*, Russ. J. Org. Chem. 50 (2014), pp. 459–463.

[35] A. Varnek, D. Fourches, F. Hoonakker and V.P. Solov'ev, *Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures*, J. Comput. Aided. Mol. Des. 19 (2005), pp. 693–703.

[36] F. Hoonakker, N. Lachiche, A. Varnek and A. Wagner, *Condensed Graph of*

*Reaction: considering a chemical reaction as one single pseudo molecule*, Int. J. Artif. Intell. Tools 20 (2011), pp. 253–270.

[37] R.I. Nugmanov, R.N. Mukhametgaleev, T. Akhmetshin, T.R. Gimadiev, V.A. Afonina, T.I. Madzhidov and A. Varnek, *CGRtools: Python Library for Molecule, Reaction, and Condensed Graph of Reaction Processing*, J. Chem. Inf. Model. 59 (2019), pp. 2516–2521.

[38] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. Tetko and G. Marcou, *ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors*, Curr. Comput. Aided-Drug Des. 4 (2008), pp. 191–198.

[39] J. Catalán, V. López, P. Pérez, R. Martin-Villamil and J.-G. Rodríguez, *Progress towards a generalized solvent polarity scale: The solvatochromism of 2-(dimethylamino)-7-nitrofluorene and its homomorph 2-fluoro-7-nitrofluorene*, Liebigs Ann. 1995 (1995), pp. 241–252.

[40] J. Catalán and C. Díaz, *A Generalized Solvent Acidity Scale: The Solvatochromism ofo-tert-Butylstilbazolium Betaine Dye and Its Homomorpho,o'-Di-tert-butylstilbazolium Betaine Dye*, Liebigs Ann. 1997 (1997), pp. 1941–1949.

[41] R.W. Taft and M.J. Kamlet, *The solvatochromic comparison method. 2. The .alpha.-scale of solvent hydrogen-bond donor (HBD) acidities*, J. Am. Chem. Soc. 98 (1976), pp. 2886–2894.

[42] M.J. Kamlet and R.W. Taft, *The solvatochromic comparison method. I. The .beta.-scale of solvent hydrogen-bond acceptor (HBA) basicities*, J. Am. Chem. Soc. 98 (1976), pp. 377–383.

[43] M.J. Kamlet, J.L. Abboud and R.W. Taft, *The solvatochromic comparison method. 6. The .pi.* scale of solvent polarities*, J. Am. Chem. Soc. 99 (1977), pp. 6027–6038.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Scikit-learn: Machine Learning in Python*, (2012).

Table 1. The data sets characteristics

| Data set | Number of reactions | Number of unique transformations | Number of unique solvents | Ref. |
|---|---|---|---|---|
| $S_N2$ | 4830 | 1382 | 43 | [20] |
| E2 | 1820 | 934 | 21 | [30] |
| DA | 1866 | 812 | 21 | [31] |
| $S_N2$ (external) | 90 | 48 | 3 | [20] |

Table 2. Coefficient of determination ($R^2$) and RMSE for different external validation strategies of QSPR for three reaction data sets

| Data set | Strategy of validation | | | | | |
|---|---|---|---|---|---|---|
| | 'Reaction-out' | | 'Transformation-out' | | 'Solvent-out' | |
| | $R^2$ | RMSE, logk units | $R^2$ | RMSE, logk units | $R^2$ | RMSE, logk units |
| $S_N2$ | 0.83 | 0.48 | 0.58 | 0.76 | 0.39 | 0.98 |
| E2 | 0.73 | 0.77 | 0.56 | 0.98 | 0.14 | 1.29 |
| DA | 0.85 | 0.73 | 0.71 | 1.04 | 0.76 | 0.94 |

Table 3. Results of external 'solvent-out' validation of developed QSPR models. Each fold is characterized by one solvent. Only solvents used in 10+ reactions were shown

| folds/solvents | number of reactions in the test set | number of unique transformations | $R^2$ | RMSE, log*k* units |
|---|---|---|---|---|
| methanol* | 422 | 167 | 0.406 | 0.847 |
| acetone* | 327 | 84 | 0.540 | 0.854 |
| ethanol* | 305 | 89 | 0.676 | 0.593 |
| nitrobenzene | 284 | 87 | -0.644 | 1.481 |
| dimethyl sulfoxide * | 173 | 25 | -0.380 | 1.371 |
| benzene | 137 | 53 | 0.520 | 0.745 |
| water | 113 | 36 | -0.146 | 1.039 |
| 1,4-dioxane* | 80 | 27 | 0.632 | 0.542 |
| dimethylformamide | 74 | 36 | 0.518 | 1.206 |
| acetonitrile* | 56 | 41 | -0.193 | 0.973 |
| nitromethane | 40 | 14 | 0.140 | 0.617 |
| oxolane* | 34 | 14 | 0.741 | 0.415 |
| phenylmethanol | 28 | 12 | 0.481 | 0.434 |
| chlorobenzene | 26 | 15 | 0.658 | 0.507 |
| butan-1-ol | 26 | 11 | 0.819 | 0.331 |
| propan-2-ol | 23 | 16 | 0.788 | 0.431 |
| propan-1-ol | 17 | 9 | 0.828 | 0.318 |
| bromobenzene | 16 | 10 | 0.645 | 0.442 |
| butan-2-one | 15 | 7 | 0.725 | 0.194 |
| toluene | 15 | 9 | 0.250 | 0.623 |

| folds/solvents | number of reactions in the test set | number of unique transformations | $R^2$ | RMSE, log*k* units |
|---|---|---|---|---|
| cyclohexane | 14 | 12 | -2.215 | 1.491 |
| sulfolane* | 14 | 5 | 0.175 | 0.231 |
| anisole | 11 | 7 | 0.699 | 0.186 |
| heptan-1-ol | 11 | 4 | 0.851 | 0.177 |
| 3-methylbutan-1-ol | 10 | 6 | 0.126 | 0.538 |

*mixtures with water

Figure 1. The schematic representation of ‘reaction-out’, ‘transformation out’, and ‘solvent out’ cross-validation. Each shape characterizes a CGR, and different colours indicate the solvent used. The red star is a unique reaction. The unique reactions are excluded from the test set and always complement the training set

Figure 2. The percentage of reactions carried out in the most popular solvents for the $S_N2$ data set

Figure 3. The percentage of solvent and fragment descriptors used for branching at particular depth levels in decision trees inside trained on $S_N2$ reactions RF model. The root node in the decision tree was considered as having depth 0. Notice that only a few trees had depth greater than 40 which caused spurious fluctuations. Temperature descriptor was ignored