Predictive models for kinetic and thermodynamic parameters of reactions

Timur Gimadiev

UNIVERSITE DE STRASBOURG

ECOLE DOCTORALE DES SCIENCES CHIMIQUES

[UMR 7140]

THÈSE présentée par :

Gimadiev Timur

Soutenue le : **11 juillet 2018**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**
Discipline/ Spécialité : Chimie/Chémoinformatique

Modèles prédictifs pour les paramètres cinétiques et thermodynamiques des réactions chimiques

THÈSE dirigée par :

VARNEK Alexandre	Professeur, Université de Strasbourg
MADZHIDOV Timur	Docteur, Université Fédérale de Kazan

RAPPORTEURS :

ERTL Peter	Docteur, Novartis , Suisse
PETITJEAN Michel	Professeur, INSERM, Université de Paris

EXAMINATEUR :

KELLENBERGER Esther	Professeur, Université de Strasbourg

GIMADIEV Timur

Modèles prédictifs pour les paramètres cinétiques et thermodynamiques des réactions chimiques

Résumé

Ce travail est consacré à la modélisation QSPR des propriétés cinétiques et thermodynamiques des réactions chimiques à l'aide de l'approche Graphe Condensé de Réaction (CGR). Le CGR permet de coder des structures de réactifs et de produits en un seul graphe moléculaire pour lequel des descripteurs moléculaires peuvent être générés. Une base de données contenant plus de 11000 réactions collectées manuellement a été développée puis utilisée dans la modélisation. Les modèles prédictifs ont été construits pour les constantes de vitesse de réactions Diels-Alder, S_N2 et E2 ainsi que pour les constantes d'équilibre des transformations tautomères. Ils sont rendus publics via un portail WEB. Une partie de la thèse concerne une étude de mécanique quantique des réactions entre des sydnones et des alcynes contraints pour lesquels la taille du jeux de données n'était pas suffisante pour produire des modèles statistiquement significatifs.

Résumé en anglais

This work is devoted to QSPR modeling of kinetic and thermodynamic properties of chemical reactions using the Condensed Graph of Reaction (CGR) approach. CGR allows encoding structures of reactants and products into one sole molecular graph for which molecular descriptors can be generated. A comprehensive database containing some 11000 manually collected reactions has been developed then used in the modeling. Predictive models were built for rate constants of Diels-Alder, S_N2 and E2 reaction as well as for equilibrium constants of tautomeric transformations. They are available for the users via WEB portal. A part of the thesis concerned quantum mechanics studies of reactions between sydnones and strained alkynes for which the size of the dataset was not sufficient to produce statistically meaningful models.

Table of contents

Acknowledgements

I would like to express my gratitude to my academic advisors, Prof Alexandre Varnek and Dr Timur Madzhidov for their help in academic challenges that I have faced. I would like to thank all the members of the jury, Dr Michel Petitjean, Dr Peter Ertl and Prof Esther Kellenberger for accepting to judge and review my work.

I would also like to express my sincere gratitude to all colleagues and co-authors from both laboratories of Chemoinformatics from Kazan and Strasbourg: Dr. Dragos Horvath, Dr. Gilles Marcou, Dr. Pavel Polishchuk, Dr. Ramil Nugmanov, Dr. Fanny Bonachera as well as, former and current PhD students Dr. Helena Gaspar, Grace Delouis, Pavel Sidorov, Marta Glavatskih, Arkadii Lin, Iuri Casciuc for their help in the work, advice and support. Special thanks to Dr. Olga Klimchuk and Soumia Hnini for their help on accommodation and administration tasks.

I am thankful to all people that helped to collect experimental data from the literature in Kazan Federal University: N. Khafizov, N. Shalin, A. Bodrov, D. Malakhova, A. Petrovski, F. Bekmuratova, G. Khaliullina, and some others.

List of Abbreviations

QSAR – Quanitative Structure-Activity Relationship

InChI – International Chemical Identifier (alphanumerical string for molecular structure representation)

GTM – Generative Topographic Mapping

CGR – Condensed Graph of Reaction

MMP – Matched Molecular Pairs

SAR – Structure-Activity Relationship

QSPR – Qantitative Structure-Property Realtionship

AAM– Atom-to-atom mapping

DB – Database

SQL – Structured Query Language

SVM – Support Vector Machine

SVR – Support Vector Regression

ISIDA – In SIlico Design and data Analysis

SMIRKS– Linear representation for reactions

RInChI – International Chemical Identifier for Reactions (alphanumerical string for reaction structure representation)

LFER – Linear Free Energy Relation

IUPAC – International Union of Pure and Applied Chemistry

SMILES – simplified molecular-input line-entry system

DFT – Density Functional Theory

HPLC – High-performance liquid chromatography

SiRMS – SImplex Representation of Molecular Structure

MRP – Matched Reaction Pair

KFU – Kazan Federal University

LS – Latent Space

CRDB – Comprehensive Réaction Database

S_N2 – Bimolecular Nucleophilic Substitution reactions

E2 – Bimolecular Elimination reactions

DA – Diels-Alder reactions

TAU – Tautomeric equilibria

RBF – Radial based function

BCN – bicyclo-[6.1.0]-nonyne

TMTH – 3,3,6,6-tetram-ethylthiaheptyne

BARAC – biarylazacyclooctynone

DIBAC – dibenzoazacyclooctyne

TCO – trans-cyclooctene

DIFO – difluorocyclooctyne

Cp(1,3) – 1,3- dimethylcyclopropene

Cp(3,3) – 3,3- dimethylcyclopropene

Résumé en français

Introduction

La plupart des approches chémoinformatiques pour l'analyse de données, la visualisation et la modélisation sont développées pour des molécules individuelles codées par des vecteurs de descripteurs. Dans ce contexte, une réaction chimique est un objet complexe, car il implique plusieurs structures moléculaires de deux types - les réactif(s) et les produit(s). Le rendement et les paramètres cinétiques et thermodynamiques de la réaction dépendent des conditions expérimentales (solvant, température, catalyseur, etc…) qui doivent aussi être pris en compte dans la modélisation. Dans ce projet, deux méthodologies différentes sont appliquées pour réduire ces complexités. Afin de réduire la complexité structurelle, on a utilisé l'approche des Graphes Condensés de Réaction (Condensed Graph of Reaction (CGR)), qui combine les structures des réactifs et des produits dans un graphe moléculaire unique, une sorte de pseudo-molécule (Figure 1), pour laquelle des descripteurs moléculaires peuvent être calculés. De plus, quelques descripteurs pertinents pour le solvant et la température ont été proposés afin de réduire la complexité des conditions réactionnelles. Ces méthodologies ont été appliquées ici pour construire des modèles prédictifs des constantes de vitesse pour trois classes de réactions ainsi que pour les constantes d'équilibre de 11 classes de réactions tautomériques.

Malgré tous nos efforts, aucun modèle statistique robuste n'a été obtenu pour les constantes de vitesse des réactions dites "bioorthogonales" entre les sydnones et les iminosydnones avec des cycloalcynes. Par conséquent, la méthode quantique DFT a été appliquée pour accéder aux états de transition des réactions et aux énergies d'activation associées.

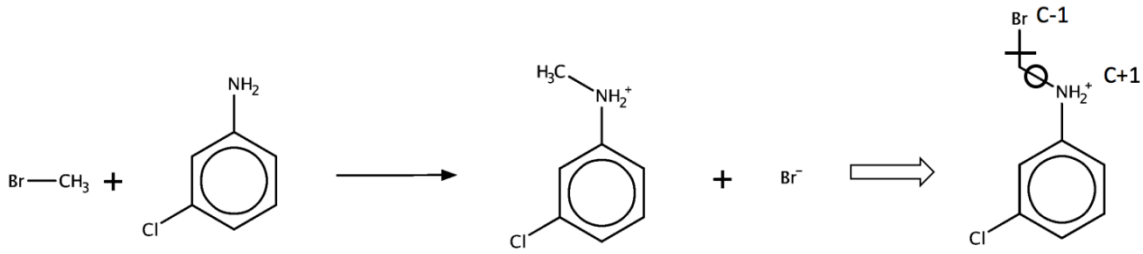


Figure 1. Exemple d'une réaction chimique (à gauche) et du Graphe Condensé associé (à droite). Les labels ⊖ et + correspondent à des liaisons créées et cassées, respectivement. Les labels "C-1" et "C+1" décrivent les charges dynamiques caractérisant les atomes dont la catégorie de charge (neutre, positive, négative) varie au cours de la transformation chimique.

Cette thèse consiste en 11 chapitres. Le chapitre d'introduction décrit différents types de codages de réactions et passe en revue les publications antérieures sur la modélisation des relations structure-activité. Le chapitre 2 apporte des informations à propos des approches et des outils de la chémoinformatique et de la chimie quantique utilisées dans cette étude. Le chapitre 3 décrit la procédure d'obtention et de validation des modèles ainsi que les méthodes d'apprentissage automatique utilisées dans ce travail. Le chapitre 4 décrit la base de données exhaustive de réactions (Comprehensive Réaction Database (CRDB)) créée dans cette étude, les procédures de standardisation des données, de nettoyage, ainsi que de nouveaux modes de représentation des réactions à l'aide des signatures CGR et des atomes dynamiques. Le chapitre 5 décrit les modèles prédictifs pour la constante de vitesse des réactions de type S_N2. Le chapitre suivant (6) est consacré à la modélisation des réactions E2 et aux études comparatives des descripteurs ISIDA par rapport aux descripteurs classiques et au nouveau type de descripteurs pour les réactions, les SIRMS. Le chapitre 7 décrit la modélisation de constante de vitesse de réactions de cycloaddition en solution. Le chapitre 8 est dévolu à la modélisation SVR et quantique (quantum chemical (QC)) des constantes d'équilibre tautomérique. Dans le chapitre 9, nous décrivons la modélisation SVR

et QC de la constante de vitesse de plusieurs réactions bioorthogonales impliquant des sydnones. The chapter 10 summarize conclusions for all work.

Data collection, cleaning, representation and storage in comprehensive Reaction Database (CRDB) :

La CRDB contient des informations à propos de plus de 15.000 entrées, incluant la structure des réactifs et des produits, les conditions expérimentales (solvants, température, catalyseur) et les constantes de vitesse pour la substitution nucléophile bimoléculaire (S_N2), l'élimination bimoléculaire (E2), les réactions de Diels-Alder (DA) et la constante d'équilibre pour certaines réactions tautomériques (TAU). Ces données ont été collectées manuellement à partir d'articles de recherches ainsi que de thèses de doctorat et de thèses d'habilitation défendues à l'Université Fédérale de Kazan. Toutes les structures sont vérifiées, organisées et standardisées en utilisant un protocole développé dans le cadre de ce travail.. Chaque réaction contient une information à propos de la transposition atome à atome et est encodée par une empreinte particulière aux réactions. Les données sont conservées dans un fichier SQL et il est possible d'y effectuer des recherches via l'outil IJChem/ChemAxon ou en utilisant des requêtes SQL. Ainsi, la CRDB est une base de données unique, qui contient des informations claires à propos des réactions chimiques, souvent manquantes dans les bases de données largement utilisées comme la CAS ou la base de données Reaxys.

Toutes les réactions ont été standardisées suivant le processus basé sur la représentation linéaire des CGRs, appelé « signatures CGR ». Au total, 10998 entrées de données brutes de réactions de substitution nucléophilique bimoléculaire (S_N2), d'élimination bimoléculaire (E2), de Diels-Alder (DA) et d'équilibre tautomérique (TAU) ont été collectées (Table 1). La colonne « données collectées » dans la Table 1 représente le nombre total de réactions collectées. La colonne « Données standardisées » montre le nombre de réactions après standardisation structurale, correction AAM et suppression des duplicats complets, c'est-à-dire des entrées où tous les champs coincident. Cette dernière raison, étant majeure, est

causée par le fait que ce jeu de données a été collecté simultanément par plusieurs personnes, et parfois, par erreur, une même réaction a pu être extraite deux fois. Les erreurs dans la structure ou l'abscence de certains champs obligatoires ont été les secondes raisons de la suppression de points de données. La colonne « Jeu modèle » contient des réactions sélectionnées pour la modélisation. La procédure de standardisation de données implique dans ce cas de standardiser, d'identifier les duplicats et de moyenner les propriétés des duplicats. Le nombre total de types de transformation (combinaison de réactifs et de produits) est donné dans le champ « transformation dans le jeu modèle ». La différence entre la taille du jeu modèle et le nombre de transformations reflète le fait que les propriétés ont été mesurées dans plusieurs conditions pour certaines transformations.

Table 1. Données expérimentales utilisées dans ce travail

Types de réactions	Données collectées	Données standardisées	Jeu modèle	Transformation dans le jeu modèle	Sources bibliographiques
S_N2	7848	7544	4830	1382	[1]
E2	1431	1389	1043	843	[1]
CA	1178	1130	880	679	PhD thesis defended in KFU
TAU	905	840	782	367	[1]

Modélisation structure-activité des constantes de vitesse des réactions S_N2, E2 et Diels-Alder.

Pour chaque réaction, le vecteur de descripteurs résulte de la concaténation des descripteurs caractérisant les structures chimiques, le solvant, et la température.

Les structures chimiques ont été encodées par des descripteurs ISIDA, correspondant à des sous-graphes CGR de différentes tailles et topologies. Chaque solvant a été caractérisé par 15 descripteurs physico-chimiques. La température inverse (1/T) a été aussi utilisée comme descripteur. Pour chaque jeu de données, le flux de travail a impliqué les étapes suivantes : (1) préparation des CGRs, (2) génération des descripteurs ISIDA [2], (3) préparation des vecteurs de descripteurs de réactions comme une combinaison des descripteurs ISIDA, des descripteurs de solvants et de l'inverse de la température, (4) construction et validation des modèles individuels en utilisant la méthode de la Régression à Vecteurs de Support [3], (5) sélection des 10 meilleurs modèles individuels et (6) estimation du consensus des 10 modèles au sujet de la propriété modélisée. Dans l'ensemble, 616 types de fragmentation ISIDA différents ont été utilisés; chacun d'eux a conduit à un modèle SVR individuel dont la performance a été évaluée par une validation croisée en 5 paquets répétée 10 fois. Les hyper-paramètres de la SVR et les meilleurs types de fragmentation ont été sélectionnés par un algorithme génétique optimisant le coefficient de détermination estimé en validation croisée. Les performances des modèles consensus de l'estimation des constantes de vitesse (log*k*) sont données dans la Table 1. On constate que la précision des prédictions de log*k* est proche de l'erreur expérimentale estimée (0.7-1.0 unités de log).

Une attention particulière a été donnée au jeu de données S_N2 qui assemble les réactions effectuées dans 44 solvants différents et impliquant à la fois des nucléophiles anioniques et neutres (voir Table 2). En essayant d'améliorer la performance des prédictions de log*k*, certains modèles "locaux" correspondant à un solvant particulier ou impliquant un type de nucléophile particulier ont été préparés. Leurs performances étaient néanmoins similaires à celles du modèle "global" (voir Table 2).

Table 2. Paramètres des jeux de données et performances des modèles [a]

Dataset	N_{tot}	N_{react}	N_{unique}	RMSE	$RMSE_{unique}$	R^2	R_{unique}^2
S_N2	4830	1382	554	0.37	0.65	0.9	0.73
E2	1043	843	395	0.72	0.87	0.75	0.58
DA	880	679	279	0.95	1.37	0.8	0.58
TAU	782	367	267	0.74	0.94	0.79	0.84

[a] Pour chaque jeu de données, les paramètres suivants sont donnés : le nombre de données (N_{tot}), le nombre de réactions (N_{react}), le nombre de réactions "uniques" (N_{unique}) pour lesquelles seulement une mesure est disponible, et les coefficients de détermination de validation croisée (R^2 et R_{unique}^2) et les erreurs quadratiques moyennes (RMSE and $RMSE_{unique}$), estimés, respectivement, pour le jeu de données entier et pour son sous-ensembe de données "uniques".

La visualisation et l'analyse de l'espace de réactions avec l'approche de Cartographie Topographique Générative (« Generative Topographic Mapping », GTM) a été réalisée avec les descripteurs fragmentaux. La puissance des cartes dans l'analyse de données est due à la possibilité de colorer les objets en fonction de différents critères. Ainsi, nous avons coloré les réactions par rapport à leur cœur, aux substrats, et à la nature nucléophile (Figure 2). On voit bien que les cartes séparent bien différentes classes de réactions.

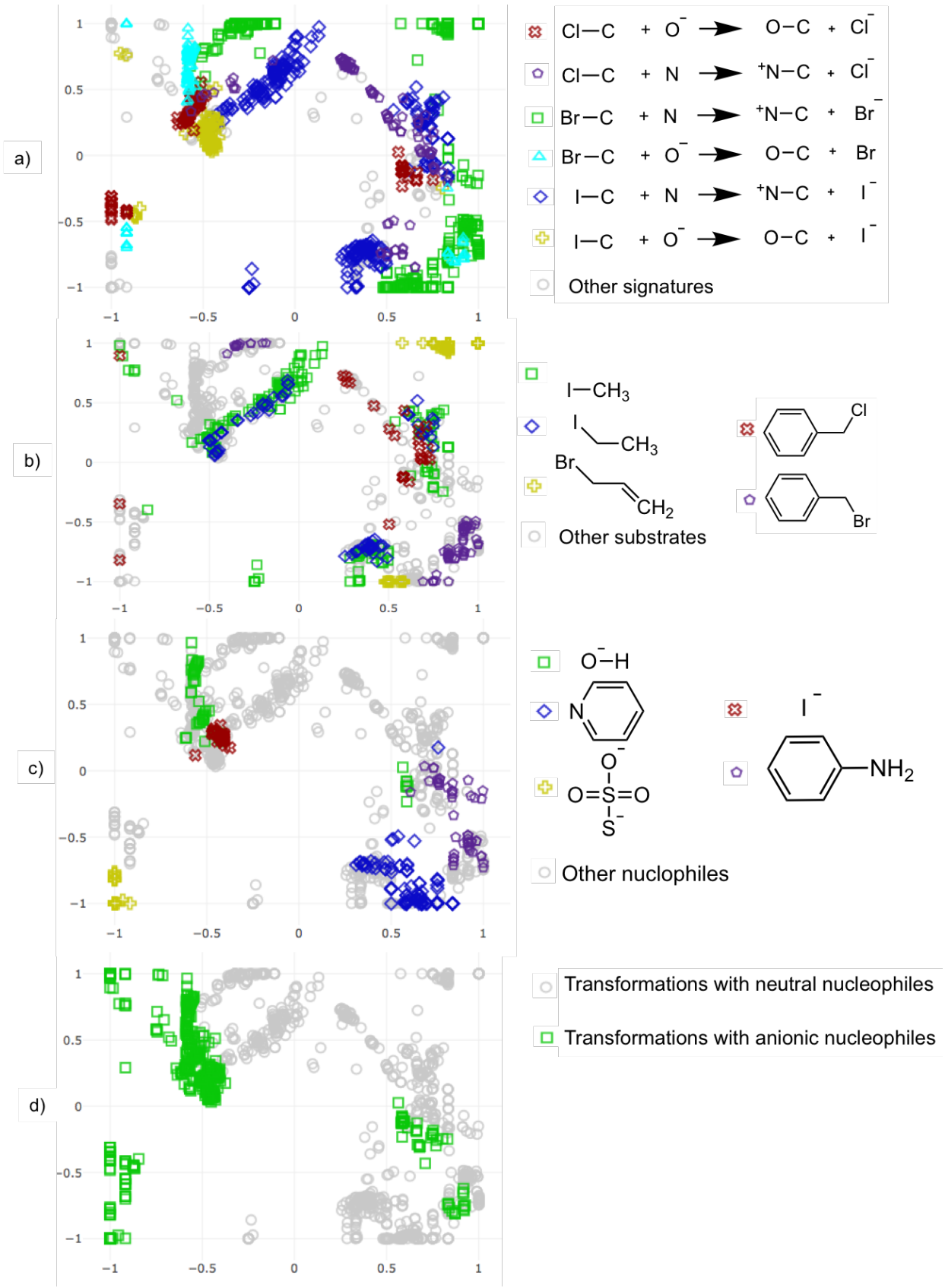


Figure 2. Carte GTM sur le jeu de données de 1394 transformations de réactions S_N2, encodées par des fragments ISIDA, sans prendre en compte les descripteurs de conditions. Les objets sont colorés en fonction a) de la signature du

centre de réaction (seuls les atomes du centre de réaction sont inclus), b) des substrats, c) des nucléophiles, d) du type de nucléophiles. Les signatures ou les molécules les plus populaires sont explicitement montrés.

QSPR et modélisation par chimie quantique des constantes d'équilibre tautomérique

Par convention, le logarithme de la constante d'équilibre tautomérique ($logK_T$) est évalué comme la différence entre les valeurs de pKa d'une paire structures tautomères. En prenant en compte l'effet cumulatif des erreurs, ceci pourrait significativement affecter la précision des prédictions de $logK_T$ estimés à partir de modèles de pKa. Ici, pour obtenir des modèles pour le $logK_T$, un équilibre tautomérique a été encodé par un Graphe Condensé de Réaction, ce qui nous a permis d'appliquer la procédure de modélisation décrit dans la section 2.2. Le jeu de données de modélisation contenait 11 types d'équilibres (kéto/énol, amino/imino, azo/hydrazone, pyridol/pyridone, pyridinoid/pyridonoid, phenol-imine/keto-amine, thione-enol/keto-thiol, amine-thione/imine-thiol, nitro/acide, forme neutre/zwitterion et cycle/chaîne) mesurés à différentes températures et dans différents solvants. La précision de la prédiction obtenue en validation croisée (RMSE = 0.94 unités de log, Table 2) est similaire aux erreurs expérimentales.
Dans un but de comparaison, des calculs de chimie quantique ont été effectués sur deux jeux de données de test (TEST1 et TEST2) contenant respectivement 23 et 24 équilibres tautomériques. Les valeurs de $logK_T$ ont été calculées comme la différence des énergies libres de formation des tautomères estimées par des calculs DFT dans B3LYP/6-311++G(d,p) implémentés dans le logiciel Gaussian09. Les effets de solvant ont été évalués en utilisant le formalisme IEF-PCM du modèle de solvant continu polarisable de Tomasi [4]. Les paramètres SMD [5] pour la part non-électrostatique de l'énergie de solvatation ont été utilisés. Nos modèles SVR surpassent significativement ces calculs DFT (Table 3). Il a été aussi montré que

nos modèles sont plus performants que l'outil commercialisé par la société ChemAxon, qui fait référence actuellement.

Table 3. Prédiction des constantes d'équilibre tautomérique ($logK_T$) en solution. Performance des calculs DFT et des modèles SVR appliqués à deux jeux de données externes TEST1 et TEST2.[a]

Method	Dataset	N_{eq}	RMSE	R^2	MT,%
DFT	TEST1	23	1.62	0.5	74
	TEST2	24	5.8	-1.2	48
SVR	TEST1	23	1.2	0.73	74
	TEST2	24	2.6	0.45	52

Nombre de points de données (N_{eq}), coefficients de détermination (R^2), erreurs quadratique moyenne (RMSE en unités de log), et taux de succès de prédiction du tautomère dominant (MT, %).

Cinétique des réactions "bioorthogonales" impliquant des sydnones.

Dans le cadre du projet ANR ClickReal, nous devions construire des modèles pour les constantes de vitesse de réaction (*k*) des sydnones (SYD) avec des cycloalcynes. Ces réactifs pouvaient potentiellement être utilisés pour des réactions bioorthogonales. Le jeu de données contenait 18 valeurs de *k* mesurées par nos partenaires. Nous ne sommes pas parvenus à construire un modèle statistique satisfaisant pour log*k*, quels que soient les descripteurs ou la méthode d'apprentissage machine utilisés. Par conséquent, une série de calculs de DFT en phase gazeuse a été effectuée afin d'identifier les états de transition des réactions et les énergies d'activation associées. L'ensemble de fonctionnelles de Perdew–Burke–Ernzerhof (PBE) et l'ensemble de base 3z implémentés dans le programme Priroda14 ont été utilisés. En accord avec les données expérimentales, le chemin réactionnel contenait 2 états de transition : l'un d'entre eux (TS1, Figure 3) définissait la vitesse de réaction. Les énergies libres d'activation calculées ($\Delta\Delta G$)

corrélaient bien avec les observations expérimentales de log*k*, en utilisant l'équation d'Arrhenius (Figure 4). La précision des prédictions de ΔΔG is 1.97 kcal/mol (RMSE) pour SYD.

Table 4. Reactions of sydnones (A,B,C) with bicyclo-[6.1.0]-nonyne (BCN) and 3,3,6,6-tetram-ethylthiaheptyne (TMTH)

No.	**R**	**X**	**Cycloalkyne**	**Rate constant (tolerance), $M^{-1}*sec^{-1}$**	**Reference**
A	p-Me C_6H_4	H	BCN	0.032 (0.001)	[6]
B	p-Me C_6H_4	F	BCN	42	[6]
C	p-Me C_6H_4	F	TMTH	1500	[6]

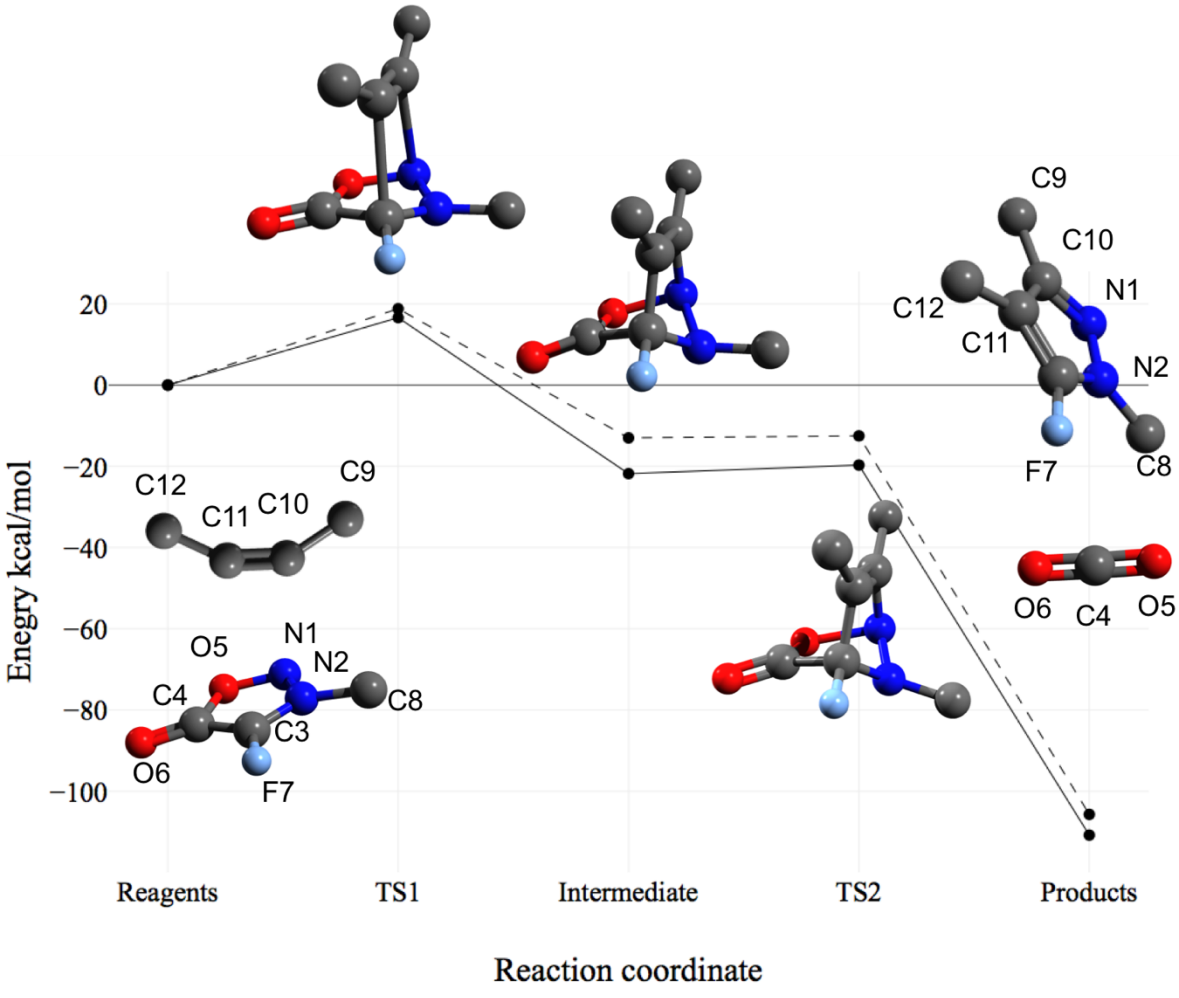


Figure 3. Voie réactionnelle pour la réaction A (ligne en pointillés rouges), B (ligne noire pleine), et C (ligne en pointillés bleue), selon la Table 4. Les énergies libres relatives à 298K des molécules par rapport aux réactifs sont montrées. La structure des réactifs, les états de transition et intermédiaires pour la réaction B sont montrés. Le substituant R du sydnone ainsi que presque tous les atomes du BCN ont été omis pour des soucis de clarté.

la Figure 3 montre que l'étape limitante des réactions est le premier état de transition (TS1), l'intermédiaire étant plus bas en énergie libre que les réactifs de 21 kcal/mol. Selon notre calcul, l'intermédiaire est très instable étant donné qu'il est séparé des produits par un état de transition (TS2) avec une petite barrière (environ 1 kcal/mol). La décomposition intermédiaire est exergonique d'environ 85 kcal/mol. La même chose est vraie pour la réaction beaucoup plus lente A. L'effet du solvant calculé par le modèle IEF-PCM [4] n'a presque aucune influence sur la barrière de réaction, à cause de la compensation. Dans la réaction C entre le fluorosydnone et le TMTH, l'état de transition n'a pas été localisé du tout.

Cette petite barrière d'énergie est pathologique d'une décomposition totalement dominée par les effets entropiques très difficiles à estimer (illustrée Figure 3).

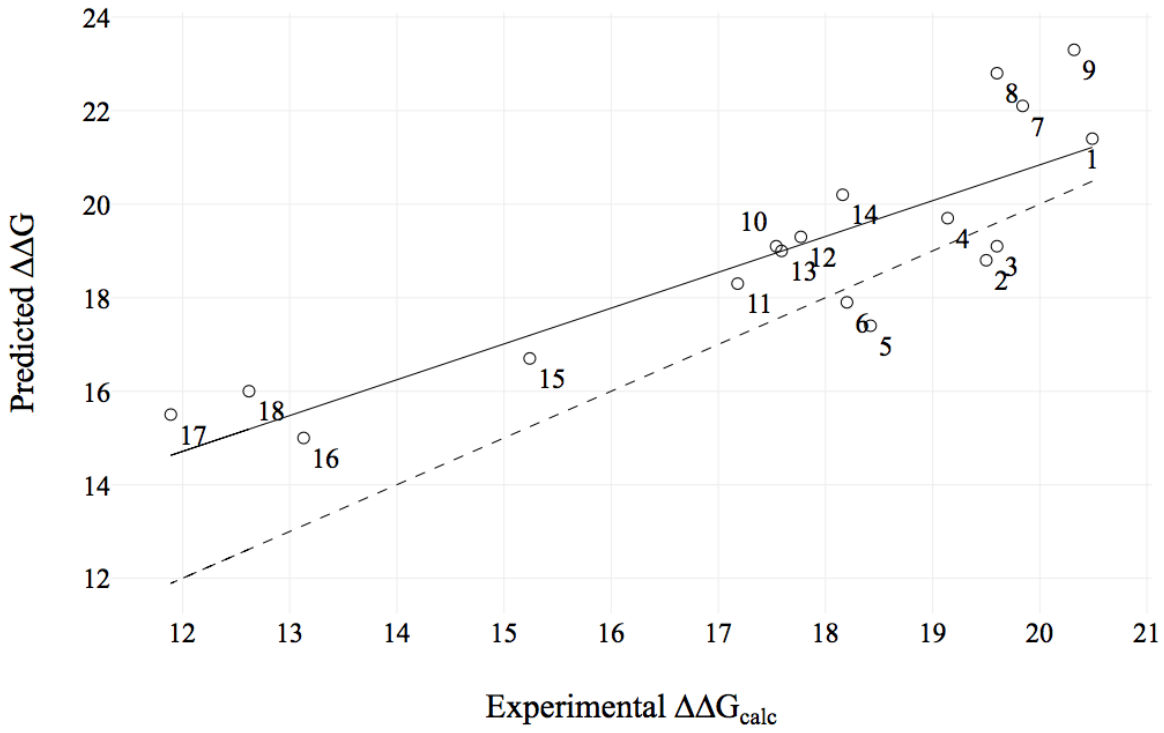


Figure 4. Corrélation entre les énergies libres d'activation calculées (Predicted ΔΔG) et les observations expérimentales (Experimental $\Delta\Delta G_{calc}$) pour les réactions entre les sydnones et les iminosydnones avec des alcynes cycliques. Les ΔΔG expérimentales sont calculées à partir des valeurs mesurées de log*k* en utilisant l'équation d'Arrhenius. Les coefficients de détermination (R^2), les erreurs quadratiques moyennes (RMSE) et le nombre de points de données (n) sont affiché dans le graphique.

2.5 Implémentation des modèles.

Les modèles développés pour les réactions S_N2, E2, DA et TAU ont été intégrés à un service WEB et rendus disponibles pour les utilisateurs à l'adresse suivante : https://cimm.kpfu.ru/predictor

Conclusions:

1. La *Comprehensive Reaction Database* contenant des informations sur plus de 15.000 réactions chimiques a été créée. Elle inclut la structure des réactifs et

des produits, les conditions expérimentales (solvant, température, catalyseur) ainsi que des paramètres cinétiques et thermodynamiques pour quatre classes de réactions.

2. La méthodologie des Graphes Condensés de Réactions a été étendue par l'introduction du concept de charges dynamiques décrivant les atomes dont la catégorie de charge (neutre, positive, négative) varie au cours de la transformation chimique.
3. Une procédure originale de modélisation des paramètres cinétiques et thermodynamiques des réactions chimiques en solution a été proposée. Elle inclut : (1) la transformation de l'ensemble des réactifs et produits en Graphes Condensés de Réactions, (2) la préparation des vecteurs de descripteurs de réactions comme une combinaison des descripteurs fragmentaux, des descripteurs du solvant et de l'inverse de la température, (4) la construction et la validation des modèles en utilisant la méthode de Régression à Vecteurs de Supports (SVR).
4. Des modèles prédictifs pour la constante de vitesse pour la substitution nucléophilique bimoléculaire (S_N2), l'élimination bimoléculaire (E2) et la réaction de Diels-Alder (DA) ont été préparés. La précision de la prédiction est comparable à l'erreur expérimentale.
5. Des modèles prédictifs pour les constantes d'équilibre de 11 types différents de réactions tautomériques (TAU) ont été construits. Il a été démontré qu'ils présentent de meilleures performances que les calculs DFT de haut niveau ou que l'outil commercial de ChemAxon.
6. Les modèles développés pour les réactions S_N2, E2, DA et TAU sont mis à disposition des utilisateurs à l'adresse https://cimm.kpfu.ru/predictor
7. Une série de calculs DFT sur 18 sydnones et sur les états de transition de leurs réactions avec des cycloalcynes a été réalisée. Les énergies libres d'activation calculées *in vacuo* corrèlent bien avec les vitesses expérimentales de réaction en solution.

Chapter 1.

Introduction

Reactions are the main tool for the chemist. New substances cannot be produced without being involved into this process of matter transformation. There are a lot of empirical rules that should be memorized by a chemist in many various examples of transformations. Consequently, this process makes every chemist to a very focused specialist in a narrow area of chemistry. The creation of a chemoinformatic tool helping chemist to extract and store these rules from known experimental data will improve synthesis flexibility of research labs. Chemoinformatics deals with information management and prediction of different properties of various chemical systems [7]. A lot of articles are devoted to modeling of different features of molecules, but much less has been written about predicting of properties of reactions like rate constant, conditions, selectivity, yield, etc.

Recently, a gain of interest of chemoinformatics specialists for chemical reaction modeling is observed. It could be related to the fact that, first, a wealth of data in chemical reaction has been accumulated by now and, second, chemoinformatics technologies were developed enough to address new challenges. Several reviews were published [8, 9]. It was stated that organic synthesis is a rate-limiting factor in drug discovery and the usage of artificial intelligence tools could revolutionize medicinal chemistry [10].

The development of chemoinformatics approaches for reactions is limited by the fact that it is a complex process: it involves reagent(s) and product(s), depends on their concentrations, reaction conditions (like temperature, pressure, radiation), This makes representation of reactions in electronic databases much more problematic than the storage of simple molecules, requiring specific procedures for standardization. Another rising problem is incompleteness of experimental data. There is a lot of information about reactions, but popular databases like Reaxys and SciFinder provide very heterogeneous information. Often, only one major component of reaction is reported. The information about kinetics of reactions is rare and usually comes from only one source, so there is no cross-validation of values between various experimental methods.

Currently, the main tool of reactions studies is quantum chemistry. The main advantage of this method is that the whole energetic profile of a reaction can be generated by calculations. It means that in principle all possible ways of reaction can be

determined with rather high detail resolution of the electron transferring processes. Quantum chemistry is still the only one way for *ab initio* prediction (e.g. not based on previous knowledge of related examples) of energetic characteristics like rate constants or activation barriers. But it is limited by intrinsic inaccuracy of polarizable continuum solvent models. Additionally, fast methods like DFT must first be calibrated on certain sets of molecules and usually does not achieve "chemical" accuracy (about 1 kcal/mol) [11]. The accuracy *versus* computing effort ratio is much too low to qualify these approaches as large-scale predictors of properties of large series of reactions. So, we decided to use another strategy, which is empirical, not mechanistic and, in exchange, extremely fast. The proposed approach implies application of QSAR/QSPR modeling, a machine-learning technique that learns a model of a property from known examples and extrapolates property values of new items based on this model. The Condensed Graph of Reaction (CGR) approach helped exploiting rich base of Chemoinformatics methods to predict properties of reactions. CGR representation (as it shown at Figure 5) allows the encoding of a reaction as pseudo molecule and to apply all QSPR techniques for modeling of its properties.

The thesis consists of 10 chapters. Chapter 2 provides information about the approaches and tools of chemoinformatics and quantum chemistry used in this study. The 3rd chapter describes different types of reactions encoding in chemoinformatics and reviews publications on structure-reactivity modeling and provides with some information about chemoinformatics and quantum chemistry approaches and tools used in this study. The 4th chapter describes the Comprehensive Reaction Database (CRDB) created in this study and procedures of data curation, cleaning and new way of representation of reactions with the help of CGR signatures and dynamic atoms. The chapter 5 describes predictive models for the rate constant for S_N2 reactions. The Next chapter 6 is devoted to modeling of E2 reactions and benchmarking of ISIDA descriptors in comparison with classical ones and new type of mixture SIRMS descriptors for reactions. In the 7th chapter application of the approach to cycloaddition reactions in solution is described. Chapter 8 is devoted to the SVR and quantum chemical modeling (QC) of the tautomeric equilibrium constants. In Chapter 9, the SVR and QC modeling of the rate constant of

several bioorthogonal reactions involving sydnones is described. The Chapter 10 summarize conclusions for the work.

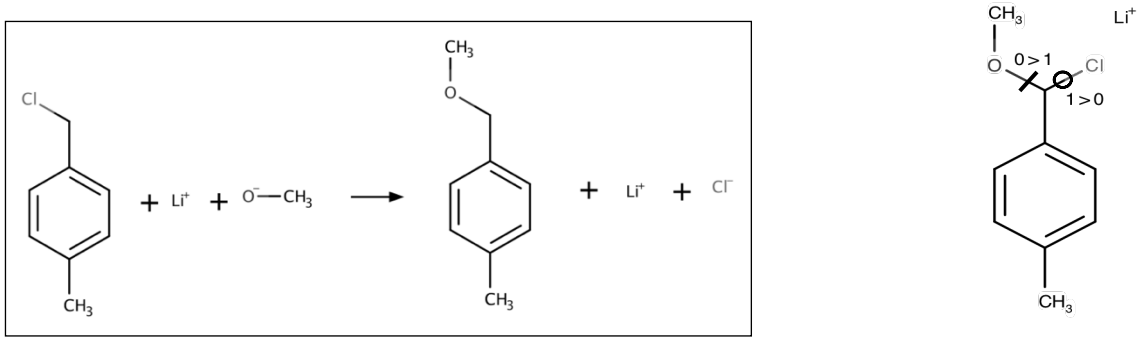


Figure 5. Example of a chemical reaction (left) and related Condensed Graph (right). Labels "0>1" and "1>0" correspond to created and broken bonds, respectively.

Chapter 2.

Review on chemical reaction modeling studies

2.1 Thermodynamics and kinetics of reaction

A **chemical equation** is a special symbolic representation of a chemical reaction. The reactant entities are given on the left-hand side and the product entities on the right-hand side.

$$xA + yB + \cdots = nC + mD + \cdots$$

The numbers next to the formulae of entities represent number of molecules participating reaction act are called stoichiometric coefficients [12]. The first-ever chemical equation was diagrammed by Jean Beguin in 1610 in first edition of "Tyrocinium Chymicum" book [13].

The reaction is a complex instance that requires a description of physical conditions and chemical environment of compounds along with description of reagents and products. Chemical environment includes compounds such as solvents, catalysts, catalytic poisons, additives, etc, which do not participate in reaction according to its chemical equation, but can influence it. Physical conditions include physical factors that are significant for reaction: temperature, pressure, irradiation, etc. All of them influence the macroscopic parameters characterizing the reaction process : yield, rate, selectivity, equilibrium constant etc. From chemists' point of view, the most important endpoint of reactions is the yield of desired compound. The latter depends on the speed of product and byproducts accumulation (reaction rate) and the equilibrium constant, the key thermodynamic parameter of a chemical process.

Properties of reactions can be divided into thermodynamic and kinetic ones. Thermodynamic properties reflect the position of equilibrium, heat of reaction and work against external forces. Kinetic properties reflect speed of reagents conversion or product accumulation.

Among the most important thermodynamic properties of chemical reaction are enthalpy, entropy and free energy of reaction. The enthalpy, H, comprises a system's internal energy, which is the energy required to create the system, plus the amount of work required to make room for it by displacing its environment and establishing its volume and pressure. The difference between enthalpy of products and reagents is equal to the heat released or absorbed in reaction, provided that system has constant pressure:

$Q = -\Delta H$. Entropy, S, is a thermodynamic function reflecting the number of microscopic configurations that a thermodynamic system can have in a state with defined macroscopic variables. In other words, it measures the degree of disorder in the thermodynamic system. The difference between enthalpy (total energy) and entropy (amount of useless energy accounting for disordered movement) defined Gibb's free energy $G = H - TS$, reflecting the amount of work that a thermodynamic system can perform (where T is absolute temperature of the system, in Kelvin). The standard change of Gibb's free energy in a reaction (Eq. 1), under isothermal and isobaric conditions is related with the equilibrium constant, equation (Eq. 2), K, as shown in equation (Eq. 3).

Eq. 1 $$\Delta G = \Delta H - T\Delta S$$

Eq. 2 $$K = \frac{[Product1][Product2]\ldots}{[Reagent1][Reagent2]\ldots}$$

Eq. 3 $$\Delta G = -RT \cdot \log K$$

where R is the gas constant.

The equilibrium constant defines the relative concentrations of products and reagents at chemical equilibrium. Thus, equation (2) and (3) used for calculation of reagent and product concentration. While direct experimental measurement of ΔG is rather challenging, equilibrium constant could be readily calculated from (2) if composition of mixture at equilibrium is known.

Chemical kinetics deals with the study of reaction evolution in time. The most essential characteristics studied by kinetics are rate of compound formation (for product) or conversion (for reagent). Both of them are usually called reaction rate but one should take into account that in case of competitive or consecutive reactions rate of product formation and reagent conversion can differ. Reaction rate depends on concentration of reagents involved in elementary reaction step. If the reaction rate depends on concentration of one, two, etc compounds (reagents in elementary step), it is called mono-, bimolecular, etc respectively. The rate constant, k, is obtained dividing the rate by concentrations of compound. The constant is insensitive to reagent concentration, but still depends on temperature (via Arrhenius equation (Eq. 4)), solvent, pressure, and other conditions.

Eq. 4 $$k = Ae^{-\frac{E^a}{RT}}$$

where A is pre-exponential factor, and E^a is activation energy of reaction.

Thermodynamics and kinetics are closely connected with each other, e.g. rate constants of forward, k_f, and backward reaction, k_b, can be used to calculate equilibrium constant for this process, $K = k_f / k_b$.

Competitive reversible reactions could proceed under thermodynamic or kinetic control. A reaction selectivity will highly depend on the dominating type of control that itself depends on reaction condition. Here (Figure 6), product **P1** has lower activation energy for reaction of its formation (greater reaction rate), but **P2** is more stable (dominate at equilibrium). In this case in the beginning of reaction the product **P1** is formed more rapidly but over time the amount of **P2** rises and finally ~~it~~ becomes dominating.

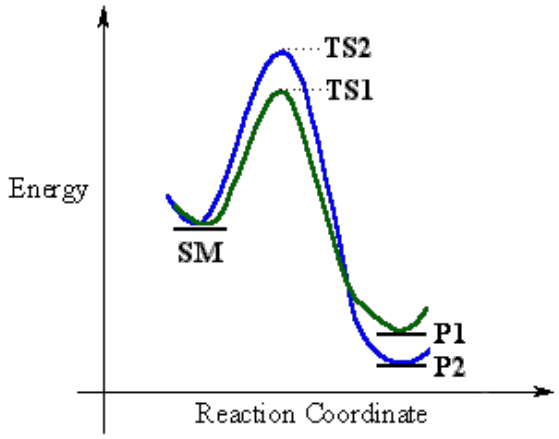


Figure 6. Thermodynamic and kinetic reaction control.

For example, high temperature and catalysts shorten the time required for reaching equilibrium and favor the thermodynamically controlled product. With properly selected reaction time, the kinetically controlled product can dominate in mild conditions. Knowing reaction rate constants and equilibrium constant or ΔG of reaction, one could shift reaction in desired direction and maximize the yield of desired product.

Unfortunately, the measure of kinetic and thermodynamic data for reactions is quite expensive, and prediction tools are either too computationally demanding (like quantum chemical prediction), or insufficiently developed (like chemoinformatics approaches), or have limited applicability (like LFER approaches).

Prediction of reaction characteristics for synthesis optimization is an unsolved problem. Quantum chemical methods could hardly be used as *a priori* calculation tools mainly due to their time- and resource-consummation. However, in 2013 scientists proposed an approach for selection of the best solvent in bimolecular substitution reaction, combining quantum mechanical computations of the reaction rate constant in a few solvents with linear regression based on solvent descriptors [14]. According to authors, it took several days to make one prediction.

Although properties of pure compounds are widely predicted by means of Quantitative Structure-Property (Activity) Relationship (QSPR, QSAR) approach [1] that favorably differs from quantum chemical calculations in terms of required effort, only few works were devoted to QSPR modeling of reactions. The main reason is that reaction is a more complex ITEM for modeling than a molecule [9]. First, a reaction involves several molecules. Next, a reaction has a sense, which must be accounted for in predictions: the backward process from products to reagents is characterized by $\Delta G_{back}=-\Delta G$ and $K_{back}=1/K$. Moreover, some chemical compounds such as solvents, catalysts and additives influence the reaction but they are usually omitted from the reaction equation. And finally, physical conditions (temperature, pressure) are very important for some properties. Hence, account for all this effect is required in QSAR/QSPR study of reactions and special approaches for handling reaction complexity need to be introduced.

In spite of the huge amount of known reactions, only a tiny portion has any kinetic or thermodynamic data and this information has not been transferred from primary sources to electronic databases. There is no publicly available database of kinetic characteristics of diverse reactions in common conditions (there are some small databases of gas phase reactions [15, 16]). The largest databases like Reaxys and CAS REACT annotate only the fact that kinetics of thermodynamic data were measured but not the values.

2.2 Reaction representation

As already mentioned, reaction description has to include information about reagents (structural information), and products, temperature, solvent, catalyst, respectively (condition descriptors).

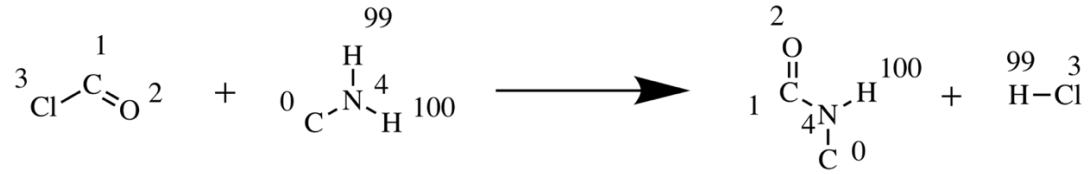
There are several ways to represent structural information of reactions, based on graph representation (usually represented as connection tables), linear notations systems (reaction is represented by an alphanumerical array) or descriptor representation calculated usually on the basis of previous two. The numeric encoding of its characteristics under the form of a vector of descriptors is key feature of chemoinformatics.

Generally, one can define 3 main ways to represent a reaction in chemoinformatics: (i) as ordered set of reagent and product molecules, (ii) reaction center based representation and (iii) product-reagent difference. In our work we proposed the special type of reaction representation [17] in which reagent and product part of reaction is treated as mixture with given composition [18].

2.2.1 Reagent-product representation

Reagent-product representation of chemical reaction is based on consideration of common reaction equation: reagents in reaction are enumerated and followed by products. Using special approaches reagents are explicitly separated from products to avoid confusion.

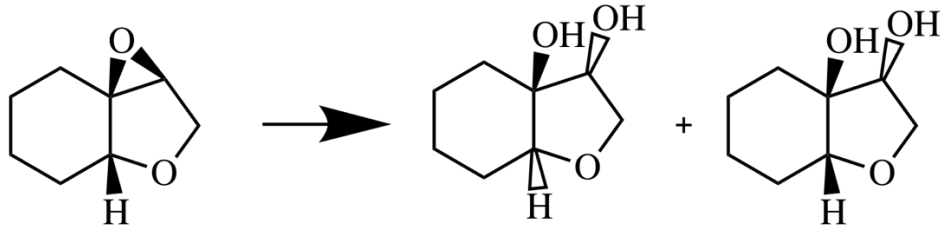
One of the most popular system for molecule representation is the simplified molecular-input line-entry system (SMILES), developed by D. Weininger [19]. It is an ASCII string that is compiled according to specification that encodes the structure of chemical species. The main advantage of this type of representation is saving of space for storage, machine and human readability. An algorithm of canonicalization [20] is used for generation of a unique SMILES representation of molecule. As a development of SMILES, the SMIRKS [21] represents reaction as transformation of reagents SMILES into products SMILES. Reagents and product molecules are separated by dots as usually for disconnected components in SMILES. After list of reagents ">" symbol opens list of additives which follows another ">" symbol, after which products are specified. In the absence of additives, the ">>" symbol separates the lists of reagents and products. Atoms and their correspondence (atom-to-atom mapping, see below) could be specified using numbers (see Figure 7).



[C:1](=[O:2])[Cl:3].[H:99][N:4]([H:100])[C:0]>>[C:1](=[O:2])[N:4]([H:100])[C:0].[Cl:3][H:99]

Figure 7. Reaction (top) and corresponding SMIRKS (bottom). Atom-to-atom mapping is specified by numbers.

The other widely used molecule line notation system was developed within IUPAC initiative. International Chemical Identifier(InChI) is a textual identifier for chemical substances, designed to provide a standard way to encode molecular information and to facilitate the search for such information in databases and on the web [22]. Widely distributed and open source standard software for InChI generation is the main advantage of it. Thus, InChIs are unique by default. This format is still machine and human readable, but it is by far less human readable than SMILES. The extension of InChI for reactions was called RInChI [23] and was developed within IUPAC project 2009-043-2-800 finished in 2017 when the first version was released [24]. In RInChI molecules of reagents or products are represented by InChI separated by the "//" symbol; products follow reagents after "///" symbol.



(I) (II) (II)

RInChI = 0.02.1S/C8H12O2/c1-2-4-8-6(3-1)9-5-7(8)10-8/h6-7H,1-5H2/t6-,7+,8-/m1/s1///C8H14O3/c9-6-5-11-7-3-1-2-4-8(6,7)10/h6-7,9-10H,1-5H2/t6-,7+,8+/m1/s1//C8H14O3/c9-6-5-11-7-3-1-2-4-8(6,7)10/h6-7,9-10H,1-5H2/t6-,7,8+/m1/s1/d+

Figure 8. RInChI for a given concurrent reaction.

Chemical table file (CT File) formats developed by Molecular Design Limited [25] represent the most popular file formats for storage and exchange by structural information. For reaction representation, the RXN file format was developed. It is based on enumeration of specified number of reagents followed by enumeration of products. Molecules are represented in MOL format. RDF file format is an extension of RXN providing functionality to store factual data associated with a given reaction in text, numeric and structure (for reagents, catalysts, etc) fields. Other popular file format based on Chemical Markup Language has an extension for reaction representation [26].

2.2.2 Reaction center representation

Reagent-product representation is not an effective way for reaction storage, since every atom is indeed represented twice – in reagents and in products. Reaction center type representations avoid this drawback, encoding the changes occurring in reaction. This type of representation is tightly connected with the notion of reaction center, *i.e.* the atoms incident to edges associated with bond order changes). With a few exceptions, changes in reaction are mostly associated with bond formation, cleavage and bond order changes.

Detection of reaction center is a crucial step required for this type of representations. Atom-to-atom mapping (AAM) or bond-to-bond mapping (BBM) is the essential procedure used for identification of reaction center [27]. AAM/BBM establishes one-to-one correspondence between atoms/bonds of reagents and products assigning each atom a unique label (Figure 9), AAM/BBM allows identification of atoms with altered environment and thus enables automated reaction center detection. For reaction on Figure 9, three atoms comprise reaction center Br^1, C^2, N^9. Reaction center can be annotated as the simplest possible reaction of a given type: Br-C + N = CH-N + Br^- (implicit hydrogens are omitted). Automated, computer-based AAM is one of the key problems in reactions processing. However the procedure itself is error prone and rather computationally intensive [27].

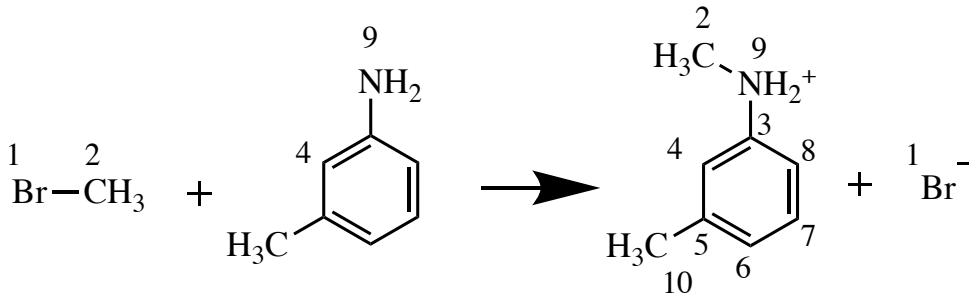


Figure 9. Common representation of chemical reaction with S_N2 mechanism. Numbers represent atom-to-atom mapping.

Several reaction center-based representations were proposed. Initially they were proposed for reaction classification rather than for information storage and retrieval. Reaction centers are indeed the "signatures" that differentiate reaction [28]. They were used for creation of an ontology of chemical reactions and creation of unique reaction identifiers, like the IUPAC nomenclature for chemical compounds. The first reaction classification scheme based on electron redistribution in pericyclic reactions including 6 atoms in reaction center was proposed by Balaban [29]. Arens proposed an approach for textual representation of bond types valid both for cyclic (when reaction center forms cyclic graph) and linear reaction centers (without cycles) [30–32], Figure 10. Hendrickson developed Balaban's approach and created coherent classification system of pericyclic reactions based on bonds redistribution within 4-, 5- and 6-membered rings [33]. The latter was further extended to other reaction types and called "comprehensive system for classification and nomenclature of organic reactions" [34], Figure 10. Zefirov and Tratch developed a hierarchical system for reaction representation and a classification that considers not only the reaction center, but its higher level of generalization [35, 36]. Albeit universal and widely accepted methodologies of reaction classification haven't been developed, the efforts were not made in vain and finally led to progress in reaction data mining. One of the most promising approaches in this emerging field of application is the Condensed Graph of Reaction methodology.

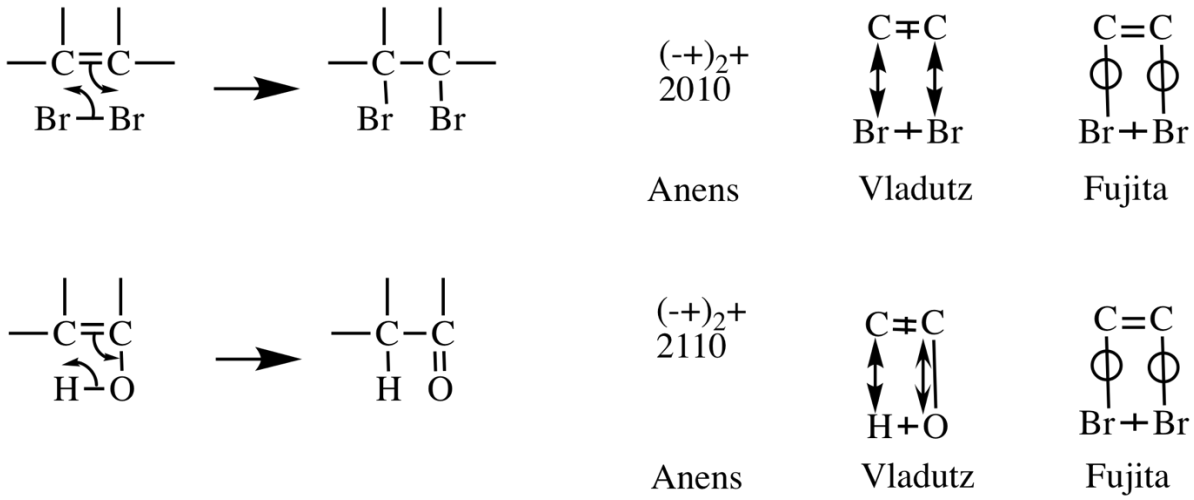


Figure 10. Comparison of different reaction representations.

The Condensed Graph of Reaction approach is based on the early work of Vladutz [37], who used it for reaction information storage and classification. He was the first to give the definition of reaction center as group of atoms in which changes of bonds take place during the reaction and proposed a method for representing the reaction center in terms of bond changes [38]. He proposed to merge reaction center atoms from reactant and product graphs, and specially mark changes in the bond orders, (Figure 10). He called the obtained graph the "skeletal scheme of reaction". The latter is common to reactions of the same type and encodes the changes taking place in the reaction center. Kiho proposed to construct these graph notations not only for the reaction center, but also for the whole reaction [39], however his paper was unnoticed and the same approach was proposed independently by S. Fujita [40]. He called the superimposed graph of reagents and products Imaginary Transition State [40]. The latter had three types of bonds: in-bonds, out-bonds and par-bonds standing for the bonds that are formed, cleaved and unaffected in the reaction, respectively. Thus the whole reaction is represented by a single graph with different labels on the edges that trace bond transformation. Further development of this approach was made in the group of Kauffman, who renamed Imaginary Transition State representation as Condensed Graph of Reaction and distinguished only two types of bonds: ordinary bonds (single, double, etc.) and dynamic bonds (single broken, single created, etc.) [41]. To additionally represent changes in stereochemistry and atom formal charges in reaction he added pseudo-atoms standing for atomic charges, stereochemistry, etc. Thus any changes in the reaction were encoded by bond reorganization.

Any atom-to-atom-mapped reagent-product representation of a reaction can be easily converted into CGR, Figure 11. A CGR can be easily represented as any molecular 2D-sketch. The bonds, which are created, broken, or modified during the reaction are called dynamic bonds and represented by specific notations denoting the change, e.g. a single bond broken by the reaction, Figure 11. The bonds that are not changed in reaction (single, double, aromatic, etc.) are represented conventionally.

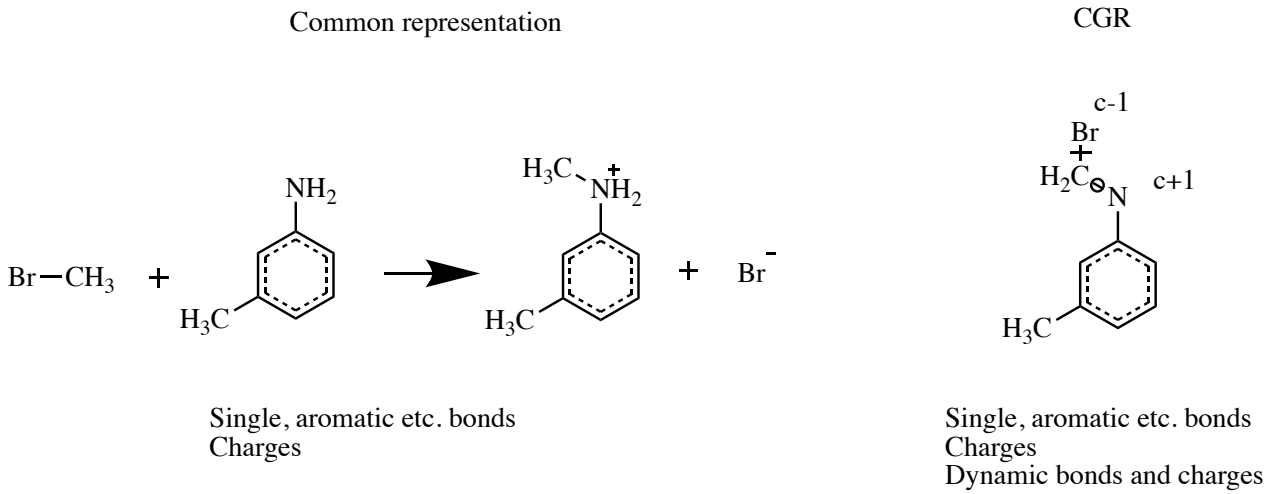


Figure 11. Common representation and CGR representation of reaction with addition features (dynamic bonds and dynamic charges). Bond with circle – dynamic bond corresponding to formed single bond, crossed bond represent cleaved single bond. "c+1" and "c-1" labels are used to represent dynamic atoms: increase or decrease of atomic formal charge by one, respectively

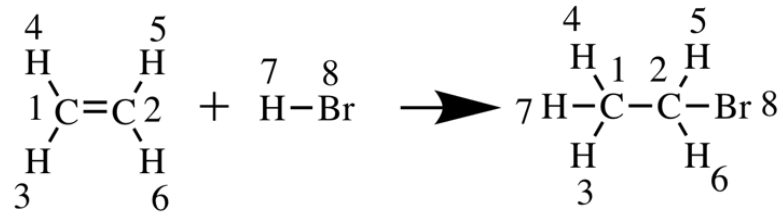
Recently we proposed discontinue support for pseudo-atoms in CGR and instead introduce "dynamic atoms" to represent changes in atomic properties [42].

2.2.3 Representation based on difference in structure of reagents and products.

While detection of reaction center requires AAM, the changes in reaction could be detected subtracting features or reagents from products. This idea is used in difference reaction representation.

Ugi and Dugundji [43] proposed to represent reaction as difference of bond-electron matrices of products (E matrix) and reagents (B matrix). B and E-matrices of

reagents and products, correspondingly, are symmetric matrices where elements of main diagonal represent number of valence electrons on the lone pairs of certain atoms and off-diagonal elements are bond orders (Figure 12).



	1	2	3	4	5	6	7	8
1		2	1	1				
2	2				1	1		
3	1							
4	1							
5		1						
6		1						
7								1
8							1	6

B

	1	2	3	4	5	6	7	8
1		-1					1	
2	-1							1
3								
4								
5								
6								
7	1							-1
8		1					-1	

R

	1	2	3	4	5	6	7	8
1		1	1	1			1	
2	1				1	1		1
3	1							
4	1							
5		1						
6		1						
7	1							
8		1						6

E

Figure 12. Ugi-Dugundji R matrix as representation of reaction.

A reaction is represented as a new difference matrix R = E - B. R gives a description of changes taking place in the reaction center and has some other interesting mathematical properties [43]. The Ugi - Dugundji formalism was used to plan organic synthesis and predict reaction pathways [44] and to search for the shortest distance between the reactant and product [45]. AAM is required for R-matrix creation. This type of representation is useful for many applications but has not been used for reaction characteristics modeling.

2.3. Reaction descriptors

Graph representations of chemical objects are mainly used for primary structural information storage but they need further preprocessing in order to encode information leading the way to numerical descriptors that are perfect for in silico processing. Representation of chemical items as vectors of numerical descriptors allows for fast information retrieval and can be used in conjunction with most machine learning methods i.e. QSAR/QSPR modeling. However, descriptor representation is not always invertible, and the structure of the item usually may not be always be restored from

descriptor values – which implies that the encoding process may trigger some loss of information.

Descriptors are values that otherwise represent features of a chemical item. Evolution of QSPR/QSAR modeling resulted in numerous descriptors of different types. The comprehensive compilation of descriptors were made by Todescini and Consonni [46]. They can be categorized into several types. For example, according to origin, descriptors could be classified as:

- Measured physicochemical properties such as "logP(o/w)," the logarithm of the octanol/water partition coefficient, molecular polarizability and refraction.
- Calculated descriptors are mathematical models of some kind, from the simplest counting of carbon atoms and summing up of the molecular mass, to descriptors representing predicted properties according to QSPR equations, themselves relying on simpler descriptors.

According to dimensionality of the molecular representation required for calculation of descriptors one can distinguish:

- One-dimensional (1D) descriptors that are calculated from the composition of compounds,
- 2D descriptors that are calculated from a planar graph representation of a molecule,
- 3D descriptors such as "molecular volume" or "surface area" that are derived from molecular conformation, and
- 4D descriptors that require ensemble of conformations or molecular dynamics trajectories.

According to application typically the following descriptor types are distinguished:

- Fragment descriptors that are usually represented by binary and integer vectors monitoring the existence or the frequency of fragment occurrence in a structure[47]. The main advantage of fragment descriptors is their universality [48].
- Topological indices or connectivity indices are planar molecular graph invariants used for numerical representation of graph topology [49].

- Physicochemical descriptors can be obtained from experimental measurements of compounds physicochemical properties. The most frequently used descriptor of this type is the logarithm of the octanol/water partition coefficient, a measure of the hydrophobic character of a molecule. Nowadays, these characteristics are usually calculated using QSPR models.
- Quantum-chemical descriptors are characteristics that obtained from approximate solution of Schrodinger's equation for molecules [50]. Energetic, molecular orbital and electron distribution descriptors are the most widely used.
- Descriptors of molecular fields are the values that approximate interaction of the molecule with some virtual probe. This descriptors are often used in 3D-QSAR modeling of biological activity, for example, in CoMFA approach [51].
- Pharmacophore descriptors show occurrence of pharmacophore-labelled fragments (usually pharmacophore pairs or triplets with defined distance between centers) in a molecule [52]. A pharmacophore is the ensemble of steric, electronic and other physico-chemical properties that are necessary to ensure optimal supramolecular interactions with a specific biological target structure.
- Substituent constants, first introduced by Hammet [53], reflect electronic or sterical influence of a given substituent on the core molecule.
- Descriptors of molecular similarity report the molecular similarity with respect to some common set of reference compounds.

The descriptors best suited to encode the structural information of reactions are fragment descriptors of reagents/products/CGR. Since fragment descriptors represent *per se* a vast chapter of possible fragmentation schemes and strategies to capture specific chemical information, this category was the only one exploited in this work (see below). Quantum-chemical descriptors and substituent constants are often used in conjunction with linear regression to model some reaction properties (usually kinetics) or to reveal reaction mechanism.

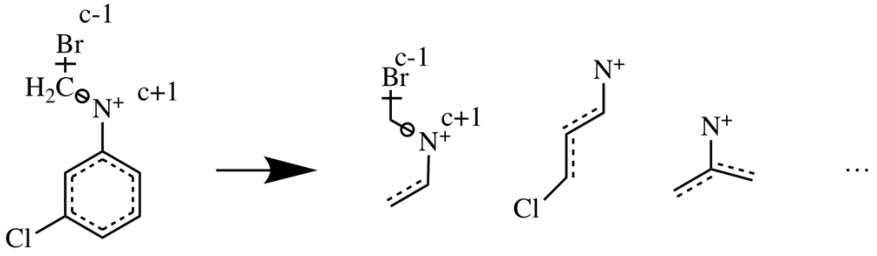
2.3.1. Fragment Descriptors

An important advantage of fragment descriptors is related to the simplicity of their calculation, storage and interpretation [2, 54, 55]). They belong to information-based descriptors [56], which encode the information stored in molecular structures. This contrasts with knowledge-based (or semi-empirical) descriptors derived from consideration of the mechanism of action. Owing to their versatility, fragment descriptors can efficiently be used to build structure–property models, perform similarity searches, virtual screening and in silico design of chemical compounds with desired properties.

The following types of fragment descriptors can be distinguished [57]: Simple Fixed Types Fragments [58], WLN and SMILES Fragments [59], Sequences [2], Atom-centered Fragments [60, 61], Bond-centered Fragments [62, 63], Atom Pairs and Topological Multiplets [52, 64, 65], Substituents and Molecular Frameworks [66–68], Basic Subgraphs [69], Mined Subgraphs [70, 71], Random Subgraphs [72], Library Subgraphs [73]. Despite many different type of fragment descriptors being proposed, only some of them became widely used. Below, two most universal approaches for fragment descriptor generation are discussed: ISIDA and SiRMS fragments.

2.3.1.1. ISIDA descriptors

ISIDA (shorthand for In SIlico Data Analysis) fragments – simple way for encoding molecules and reactions. For both molecules and CGR, the ISIDA Fragmentor [74] produces a vector of integers counting the occurrences of molecular fragments of different topology. There are two types of the ISIDA descriptors: sequences of atoms and/or bonds, and augmented atoms (circular fragments) representing a selected atom with its closest environment, Figure 13.



2	1	1	…

Figure 13. Vector of numbers after ISIDA fragmentor.

There are several main parameters controlling ISIDA fragment descriptor generation:

- Fragmentation scheme. Three fragmentation schemes are supported: sequences, augmented atoms (circular fragments) and topological triplets. User can specify whether to take atom and bond labels in the generated fragments into account. For example, in Figure 13 one can see values for sequence descriptors.
- Minimal and maximal length of sequence or augmented atom radius. All sequences having specified length or augmented atoms having given topological distance from central atom are generated. This parameter influences the balance between long and short fragments. If short fragments are too general such description will be non-informative. Long fragments could be too selective and numerous and thus could introduce noise in the model.
- Do_all_ways - flag switches off default calculation of shortest path sequences and fragments generated by all possible detours on the graph. It is relevant only for structures with cycles.
- Formal charge option adds the information of the formal charge on an atom into fragment description. This option is useful to differentiate protonation states of molecules. In case of CGR ISIDA fragment descriptors are universal and were used for many different QSAR/QSPR tasks: prediction of H-bond stability [75, 76], halogen bonding [77], metal complexation [78], and many other biological and physico-chemical properties of molecules [2, 74, 79], if to name a few.

2.3.1.2. SiRMS descriptors

SiRMS (shorthand for SImplex Representation of Molecular Structure) are topological multiplets (see 2.3.1). The descriptor vector encodes frequency of presence of all possible atomic multiplets (atom combinations) having given number of vertices. MultipletSSS could contain atoms that are not bound, but unlike other topological multiplets the topological distance between atoms is not specified in SiRMS, Figure 14. Initially, SiRMS contained only so called simplexes (tetratomic multiplets) but later

fragments with any number of centers (from two to six, as an option) became used [80]. The latter are still called simplexes due to historical reasons. Description of simplexes could contain bond orders, stereochemical configuration of simplex, atomic property labels (like charge, lipophilicity, polarizability, etc.). Optionally, simplexes containing atoms not bound to other atoms of a given fragment could be discarded.

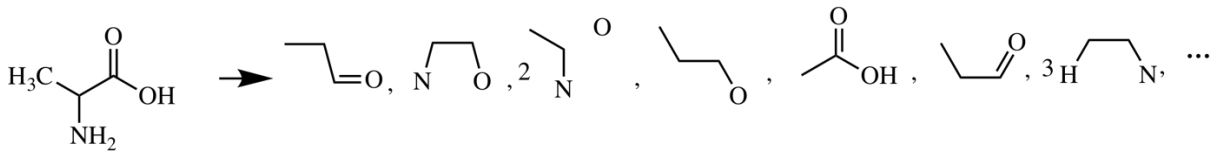


Figure 14. Generation of SiRMS descriptors for chemical compound.

SiRMS descriptors were shown to be generally applicable for prediction of many different properties: biological activity of compounds [81, 82], environmental toxicity [83], solubility [84] and other physico-chemical properties [85]. One of their specific features is the possibility to correctly represent stereochemistry-related properties [86]. Quite recently, mixture SiMRS descriptors [18] incorporating calculation of simplexes for chemically disconnected molecules were proposed. For a mixture, one fragment could contain parts coming from several molecules. This approach was successfully used for prediction of properties of binary mixtures [87]. The "Quasi-mixture" approach, in which every compound is represented as a mixture with itself, was shown to be superior over the single molecule approach in physical properties prediction [88].

2.3.2. Particular types of reaction descriptors

The reaction descriptor vector should encode two essential features of chemical reaction –the transformation (i.e. structure of reagents and products) and reaction conditions. To simplify the situation, modeling usually addresses either a set of reactions under common conditions, or a single reaction under variable conditions. In this case, the constant part of descriptor vector is neglected.

In this part we will describe some techniques for reaction descriptor calculation.

2.3.2.1. Reagent-product descriptors

The simplest way to describe reaction is to consider it as a combination of reagents and products. Reagents and products in this approach are considered as molecular graphs. Thus this approach is based on molecular descriptors making it rather general since it does not require special reaction-oriented descriptors.

2.3.2.1.1. Reagent descriptors

Let us consider the case of a series of reactions of the same type. In this case, the structure of product is fully determined by reagents and thus only reagents could be considered.

The Linear Free Energy Approach (LFER) is based on this assumption. Moreover, within LFER, reagents are congeneric, i.e. they differ by substituent. Thus changes in predicted value could be explained only by electronic and steric influence of the substituents. The first substituent descriptor was introduced by Hammett [53] and was used for prediction of acidity. Later, Taft [89] introduced the steric factor of substituents. Later, a lot of different substituent constants were proposed to describe inductive, mesomeric and other electronic and steric effects. Substituent constants [90] descriptors were used for prediction of reaction rate or equilibrium constants k using linear equation of the type $\frac{k}{k_0} = \sigma\rho$, where k0 is reaction rate with hydrogen as substituent, σ is substituent constant and ρ is reaction and scaffold type dependent constant. Often, multiparametric LFER equations were built to describe dependence of property under study on several substituent constants or solvent parameters. Since these descriptors have clear physical meaning the dependencies are used for interpretation and to study reaction mechanism. Although the usage of LFER approach for predictive purpose has mostly historical importance, it is still used for mechanism elucidation and interpretation [90, 91] Usage of the approach with predictive purpose is limited since it has a very narrow applicability mainly due to two facts: (i) only congeneric reactions could be predicted and (ii) descriptors used are obtained experimentally. Nevertheless, from time to time predictive models based on LFER approach are proposed. For example, in the work [14] LFER equation was used to correlate quantum-chemically predicted activation energies for a given reaction in different solvents with descriptors of solvents to predict optimal conditions for the reaction.

The idea that reagent descriptors could be used to predict reaction properties is sometimes exploited in the field of QSRR. Various descriptors [92] of compounds including topological indices, information indices based on charge distribution in molecules, fragment descriptors and others were used for prediction of rate of homolysis of nitrocompounds. In this case, focusing the description on anything else but the reagents is impossible, since the mechanism of reaction is unknown and reaction results in complex mixture of products.

In the work of Marcou et al different descriptor creation strategies were benchmarked [93] in order to predict Michael reaction feasibility under given conditions (solvent type and catalyst type). They compared reagent based descriptors (420 CDK descriptors, MOLMAP [63, 94], ISIDA fragments and EED descriptors [95]) and came to conclusion that the models employing them have the same predictive performance that the one based on descriptors with explicit reaction center encoding. They came to conclusion that within the considered training set, condition specific structural "patterns" or "signatures" can be established even in absence of explicit knowledge of the reaction center itself.

2.3.2.1.2. Concatenated descriptors of reagents and products

In case when for a given reagent several products can be formed, and modeling reaction property depends on the product formed, the reagent based approach described above can no longer be used. In this case, the reaction descriptor should explicitly encode both reagent and product features. The natural idea to solve this problem is to concatenate reagent and product descriptor vectors. This approach is quite universal and can be used for any chemical reaction, but leads to rather long descriptor vectors. Large descriptor vectors could introduce noise and spurious correlations into a model. If several reagents and products are involved, they should be represented in some specific order (reaction descriptor would be concatenation of descriptors for A, B, C and D molecule). Thus, careful curation of the dataset is required.

Kravtsov et al [96] used concatenated representation of reagents and products based on topological, physicochemical and quantum chemical descriptors for the modeling reaction rate of S_N2 reaction. The transformation descriptor vector was concatenated with vector of solvent descriptors and reaction temperature. Using

descriptor selection on the basis of Fast Stepwise Multiple Linear Regression and artificial neural networks as machine learning method the first chemoinformatics model able to predict reaction rate in different conditions were built. Similar approach was used in the work of the same authors for classification of preferable mechanism of nucleophilic substitution (S_N1 or S_N2) reactions [97].

2.3.2.1.3. Difference descriptors

The last approach that could be used for generation of reaction transformation descriptors without explicit consideration of reaction center is based on arithmetic difference between vectors of reagents and products. The approach is based on the idea that if reagents and products are represented by fragment descriptors, difference between them represents fragments that exist in one part of the equation but absent in the other. In such a reaction center is implicitly represented without AAM.

In case of binary fingerprint usage there are two options how the difference fingerprint could be calculated. First, the difference could be calculated as element-wise subtraction, in this case final fingerprint will contain "1", "0" and "-1" and no longer a binary number. This approach was used to classify enzymes according to reactions they catalyze [98]. The other option is to calculate difference fingerprint as bitwise logic OR operation, resulting fingerprint will be still binary but will coincide for forward and backward reaction [99].

The drawback of the approach is that only fragment descriptors can be used (otherwise their meaning is unclear), and reaction should be perfectly balanced. If one reagent of product is absent descriptor vector will contain meaningless values.

Schneider et al. [100] implemented the idea of reaction difference fingerprints, using the following procedure: authors calculated the bitstrings for each molecule involved in the reaction and the descriptor representation of the reaction was found as the difference between the sum of product bitstrings and the sum of reactant bitstrings. Resulted vector was summed with descriptors of additives to enhance the ability of the model. It allows classifying reactions by their types.

Another type of difference descriptors has been proposed by Ridder and Wagener [101] who calculated the reaction descriptors as the difference between the frequencies of occurrence of some fragment in the products and in the reactants. A similar approach,

but using fragment descriptors, which were called atomic signatures (sort of augmented atoms), was applied by Faulon et al [102]. The MOLMAP reaction descriptors calculated as the difference between the product and reactant MOLMAP descriptors, which are calculated as Kohonen's map of bond centered descriptors for a given molecule. They were used several times for classification tasks [63, 93, 94] involving chemical reactions.

2.3.2.1.4. CGR descriptors

Explicit representation of reaction center could overcome many problems related with descriptor generation: first, making emphasis on reaction center potentially enhance predictive ability and reduce chance to find spurious correlations; second, this method is insensitive to molecule ordering and reaction balancing. In LFER approach atoms belong to reaction center were usually selected manually choosing appropriate substituent descriptor or calculating some quantum chemical property of atoms causing interest.

Reaction center graph representation could be used as a basis for descriptor calculation. Varnek [2] noticed that CGR can be regarded as a pseudo-molecular graph and its fragment descriptors could be readily calculated using approaches designed for application to molecules, Figure 15.

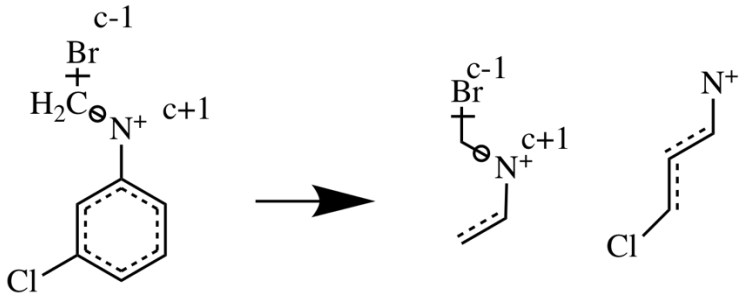


Figure 15. Example of fragment descriptors generation for CGR. One could notice that the procedure is very similar to the one used for molecules descriptor generation, given on Figure 13.

For CGR-based fragment descriptors calculation for chemical reactions one can generate fragments that contain only dynamic bonds (no ordinary bonds), or fragments containing at least one dynamic atom\bond, or all fragment descriptors. The first two options introduce special attention on reaction center that can be efficiently used for quantitative reaction similarity assessment or applied in QSRR studies. Corresponding options were introduced in ISIDA Fragmentor descriptor generation software.

Resulted descriptor vectors for chemical reactions could be compared using well-known similarity metrics and applied in reaction similarity search [103]. De Luca et al [104] used neighborhood behavior approach [105] with different types of CGR fragment descriptors and similarity indices to find optimal strategies to assess reaction similarity. The paper was based on the idea that the most similar reactions should correspond to the same type. Best descriptors selected by neighborhood behavior were used for mapping reactions using Kohonen's self-organizing map approach and it was shown that reactions of a given type fall into the same or nearby nodes. The latter was used to select reaction signatures corresponding to a given reaction center. Application of fragment descriptors for CGR in QSRR modeling will be described below.

2.3.2.2. Reaction condition descriptors

Reaction conditions are essential for prediction of reaction properties. Temperature, pressure, reactant concentration and solvent are the most important of them. Usually concentration independent characteristics are modeled (reaction rates, equilibrium constants, usually taken as logarithms) and thus reagents initial concentrations are neglected. Temperature and pressure are numerical values that could be easily added to the descriptor vector. Due to Arrhenius equation (Eq. 4) if logarithm of reaction rate is modelled it is more natural to introduce temperature descriptor as inverse of absolute temperature (1/T, where T is in Kelvin).

Traditionally, the mechanism of solvent action is associated with its ability to stabilize polar molecules in solution (called polarity), participation in H-bonds with solvent being H-donor or H-acceptor (called H-bond acidity or basicity) [90]. Moreover, a dissolved molecule could polarize solvent itself that changes the polarity of the latter. In modeling, solvent is represented as a vector of usual molecular descriptors (e.g., fragment descriptors and others) and by physicochemical parameters describing the main mechanism of solvent action on solute. In the work [96] authors included both type of descriptors however fast stepwise multiple linear regression used for descriptor selection picked only physicochemical parameters of solvents, that indirectly shows that the latter could better reflect solvent influence.

Among physicochemical parameters the most important are dielectric permittivity and refraction index. The former is related to solvent polarity and the latter with its polarizability. The dielectric permittivity influence on a dissolved molecule is a highly nonlinear. Reaction field theory (a.k.a Onsager theory) [106] gives some formulae that describe the dependence of solute energy on dielectric permittivity: Born and Kirkwood , Debye function [107]. Analogously polarizability of solvent could be described as a function of the refractive index measured by D-line of sodium spectra at 20 degrees Celsius n_D^{20} [107].

Several physicochemical scales for description of solvent effects (polarity-polarizability, H-acidity and basicity) were proposed: Catalan SPP, SA and SB scales [108–110], Camlet-Taft π^*, α and β scales [111–113], Koppel-Palm tetraparametric equations [114] for solvent effect modeling. Most of these scales are based of solvatochromic effects of the solvent, i.e. shifts in UV, visible or IR radiation absorption of some molecules in different solvents.

2.4. Computational approaches to reactivity modeling

One of the most universal approaches for modeling properties of chemical objects is based on application of machine learning tools to chemical systems. Generally this approach is called QSAR (Quantitative Structure–Activity Relationship) or QSPR (Quantitative Structure–Property Relationship), with QSAR being the most widespread term. The essence of the approach is based on finding mapping between a space represented by set of structural descriptors $\{X_1, X_2, \dots X_n\}$ and predicted property:

$$Y = f(X_1, X_2, \dots X_n)$$

This mapping is done in a way valid for most of objects. Thus model obtained on some limited portion of data called training set could be applied for the rest of objects.

Linear Free Energy Relationships [115] use very similar technique in general but interpretability of the model is a cornerstone of the approach. That is why in LFER linear regression and physicochemical descriptors are used almost exclusively. In SAR/QSAR/QSPR black box machine learning methods based on very flexible fitting functions $f(\boldsymbol{X})$ and many other type of descriptors are widely used. The main accent is done on general applicability and robustness of the model rather than its interpretability. At the same, the other methods are widely used for prediction of chemical objects

characteristics, first of all one need to mention quantum chemistry and molecular mechanics. These approaches use strict physical theories to predict desired properties of chemical systems.

Linear free energy approach dates back to 1937 when Hammett [53] proposed it to predict acidities of chemical compounds. The name arises because the logarithm of an equilibrium constant (at constant temperature and pressure) is proportional to a standard free energy (Gibbs energy) change, and the logarithm of a rate constant is a linear function of the free energy (Gibbs energy) of activation. The approach was extensively used to establish linear correlations between substituent or solvent descriptors and reaction characteristics such as reaction rate or equilibrium constants. There were a lot of investigations done in this field that are reviewed in the book of Palm [116]. In the end of XX century with the development of computers quantum chemical approaches have mostly substituted LFERs. Usually quantum chemical studies involve building simple correlations between parameters of molecules and their reactivities which are alike LFER, for example, linear equations between Diels-Alder reaction rate and conceptual DFT indices [117]. Reaction kinetics study sometimes still involves LFER to understand substituent or solvent effect and thus get some ideas on reaction mechanism.

In this part we will describe the application of different approaches to study chemical reactions.

2.4.1. Quantum chemical calculations

The quantum chemical methods are based on approximate solution of Schrodinger equation (Hartree-Fock and post-Hartree-Fock methods) or Kohn-Sham equations [118] within Density Functional Theory (DFT) [119]. One of earliest and most widely used method of approximate Schrodinger equation solution is self-consistent field (a.k.a. Hartree-Fock-Roothaan, or simply Hartree-Fock) method [120]. Post-Hartree-Fock methods use Hartree-Fock solution to decompose wavefunction into series of excited Slaters determinant that allows achieving more appropriate values of energy. MP2 [121], Multireference Configuration Interactions [122] and Coupled Cluster approach [123] are popular methods of this type. Less computationally expensive methods are based on Density Functional Theory, which uses a functional correctly reproducing exchange-

correlation energy. The list of available functionals is extremely wide [124], however B3LYP [125] and PBE [126] are the most well-known.

Energy and its derivatives are used to locate extremums on potential energy surface corresponding to stable chemical structures or transition states and in such a way used to assess molecular energy and geometry. Many other characteristics of molecular systems (including spectra) could be calculated in a similar manner.

2.4.1.1. Predicting of thermodynamic parameters

The quantum chemical calculations are the main way to predict thermodynamic parameters of molecules nowadays. Usage of quantum chemical approaches for prediction of different thermodynamic parameters is published each year in thousands of articles. Special approaches for precise estimation of thermodynamic properties of molecules taking into account all required energetic terms and usually some empirical corrections are proposed [127, 128].

Prediction of thermodynamics parameters for reagents and products could be used to assess reaction enthalpy and free energy. The latter could be used for prediction of reaction equilibrium constant. For example, the company "Schrodinger" proposed an accurate quantum chemical approach for assessing tautomeric equilibrium constants [129].

Despite quantum chemical approaches are the most used methods for prediction of reaction thermodynamics, they are rather slow and the need to estimate solvent effect sufficiently complicates this problem. The achievement of chemical accuracy in the prediction requires the performance of very sophisticated and resource-intensive calculations. And even with them the quality of data is far away from experimental accuracy [11]. Results of a blind competition performed in the frame of SAMPL2 Challenge [130] where various, mainly quantum chemical, approaches competed in the prediction of solvation energies and energy differences of tautomers have shown that computational accuracy they reach now (the mean square error of the energy calculation was ~2.5—3 kcal/mol) is insufficient for the adequate description of the energy and is much lower than the experimental one. Almost all methods used had fitting parameters

adjusted to a training set proposed in the competition that allows more appropriate estimation of desired property compared to *ab initio* calculation.

2.4.1.2. Mechanism of reaction

Usually mechanism of reaction is explained in terms of Transition State Theory, that was developed simultaneously by Eyring [131], Evans and Polanyi in 1935 [132]. A minimal energy pathway on potential energy surface that connects reagents and products is stated in the theory. This pathway describes alterations of geometry and energy in the system while reagents are transformed to products. Minimal energy pathway is the pathway that requires the least amount of energy. The movement along the pathway requires changing of one or more molecular coordinate and is called reaction coordinate. Each elementary step of reaction involves formation of Transition State (activated complex) – highest energy structure on reaction pathway. Difference between energies of transition state and reagents is called activation energy. Reaction could proceed if reagents have enough energy to jump over transition state.

Quantum chemical study of mechanism of reaction requires determination of transition state and all intermediate along reaction pathway. One also needs to prove that detected transition states belong to reaction path. Usage of Intrinsic Reaction Coordinate (IRC) following algorithm makes small steepest descent steps down from the transition state. If one could get reagents and product following IRC it means that transition state, reagents and products belong to the same reaction pathway, otherwise they belong to different domains on potential energy surface.

Eyring have shown that reaction rate depends on activation free energy:

Eq. 5. $$k = \frac{k_B T}{h}\, exp\left(-\frac{\Delta G^{\neq}}{RT}\right)$$

where $\Delta G^{\neq}$ is known as the Gibbs free energy of activation, is the standard molar Gibbs energy change for the conversion of reactants into transition state. A plot of standard molar Gibbs energy against a reaction coordinate is known as a Gibbs-energy profile. In principle the right-hand side of equation (Eq. 5) should be multiplied by a transmission coefficient, κ, which represents the probability that an activated complex

forms a particular set of products rather than reverting to reactants or forming alternative products.

In such a way, quantum chemical calculations could be used to understand mechanism of reaction and to find its rate constant [133]. Even for simple reactions quantum-chemical study of its mechanism and kinetics is computationally intensive, but nevertheless it is the most developed and widely used approach for reaction study and qualitative assessment of their characteristics [134].

2.4.2. QSAR in reaction modeling

Chemoinformatics approaches were rarely applied to study and predict reaction characteristics. Mainly it is related with reaction complexity and lack of data ready for modeling. Thus works related to reaction modeling appeared sporadically and no systematic studies were still published.

2.4.2.1. Reaction kinetics

The collection of data on reaction kinetics were mainly inspired by great interest in reaction mechanism study and popularity of linear free energy relationships in the XX century [116, 135, 136]. By the end of 1990th general workflow of QSAR modeling had been developed and thus first attempts to build models for reaction kinetics based on QSAR methodology was done.

The first works in reaction rate prediction were done in N. Zefirov's group and were based on application of artificial neural networks and molecular descriptors of different type to chemical reactions. In the work of Sukhachev et al [92] rate constants of nitro compounds decomposition in gas phase were predicted. Authors used ordinary topological and fragment descriptors ignoring conditions that were the same for all measurements.

In the work of Halberstam et al [137] a model for the rate constant of 2092 acid ester hydrolysis reactions was built. Since one of reagent was unchanged (water) the descriptors of chemical transformation were simply molecular descriptors of ester, quantum-chemical descriptors were used. To represent reaction conditions reaction temperature and Palms descriptors of solvents were concatenated with structural

descriptors. On validation set (209 reactions), RMSE of prediction was 0.34 log*k* units. Using the same dataset and workflow but fragment descriptors to represent ester structure Zhokhova at al [138] slightly increased the quality of prediction– RMSE dropped down to 0.31 log*k* units.

For bimolecular nucleophilic substitution reaction rate assessment Kravtsov et al [96] proposed approach where molecular descriptors of two substrate and product molecules were combined into the feature vector using for modeling by means of neural networks. Dataset comprised 3451 S_N2 reaction rate constant in pure solvents at different temperature. Authors used fragment, topological and quantum-chemical descriptors for encoding chemical transformation, Palm solvent descriptors to represent media. Significant descriptors were selected using the fast stepwise linear regression procedure. On 193 test set reactions RMSE was 0.58 log units. The approach was used as well for S_N2/ S_N1 reaction classification and S_N1 reaction rate prediction [97] with RMSE 0.61 log units on test set that is unbiased enough.

Further development of reaction rate modeling was related with CGR approach. Hoonakker et al [139] used dataset of 1014 reactions in water solvent proceeding at various temperatures, 3 various machine-learning techniques (SVR, M5P, MLR) and ISIDA fragment descriptors based on CGR representation of reactions for building model for log*k* prediction. The best model built using SVR and atomic sequences of topological length from 2 to 8 as descriptors shown quite good performance, Q^2=0.53, RMSE=1.26, but worse than that of Kravtsov et al [96]. The difference could arise due to different validation procedures and biased estimation of prediction error (see chapter 4.2). Hoonakker's model could only predict reaction rate in water solvent. In the next work [140] the dataset of 1041 reaction proceeding in different solvents and at different temperatures involving only neutral nucleophiles was collected. The model on this data was built using ISIDA and SiRMS descriptors of structure, 15 solvent descriptors characterizing polarity, polarizability, and proton-acceptor and proton-donor abilities of solvent, temperature of reaction. Random Forest regressor usage for model building allowed reaching RMSE on out-of-bag sample of 0.5 log*k* units that was shown to be comparable with experimental noise. Neutral nucleophiles were selected to exclude effects due to complication of reaction process in case of usage of salts as nucleophiles. In the next work [141] it was found that for particular case of S_N2 reaction using azide ion as

nucleophile addition of descriptors that encode concentration of substrate and nucleophile and the nature of counter ion increases predictive performance of the model. In this case RMSE achieved 0.98 vs 1.07 log*k* units in case of absence of counter ion descriptor. Such wise despite 4 models for S_N2 reactions were published by now, there was no model that was developed to predict reaction rate of any nucleophile in any solvent and solvent mixtures.

2.4.2.2. Reaction classification

Classification of reaction types implies the recognition of types of reactions present in databases or entered by a user, which is important for solving practical tasks of synthetic chemistry, for example, for search of similar reaction, optimal condition selection or synthesis planning (generation or selection of retrosynthetic rules). There is two approaches for reaction classification: model-based and data-based [28]. Model-based approaches use different reaction center representations schemes (see Chapter 2.2.2). Reaction center is specific for reaction of given type and that could be utilized for grouping, making reaction nomenclature and ontology [33, 34]. It is however not a universal approach since classification of reactions used in organic chemistry sometimes is not related to immediate environment of reactions center.

Similar approach that could be called rule-based works in opposite way: first, reaction ontology is manually created using concepts in organic chemistry and then rules for reaction extraction are manually proposed. For example, RXNO reaction ontology (http://github.com/rsc-ontologies/rxno) created on the basis of works [142, 143] was used in rule-based reaction classification NameRxn system [144]. This system was used to investigate the popularity of reactions of various types in medicinal chemistry [145].

Data-based classification approaches are based on application of machine learning tools to the dataset with known reaction classes to create the model able to classify new reactions. One of first approaches to classify reaction by type have used Kohonen's map and physicochemical descriptors of atoms [146]. It was shown that reactions could be quite efficiently classified; however the number of reaction types was quite small. Similarity based system for hierarchical reaction classification to additions, eliminations, and substitutions, followed by two successive subdivisions by number and types of reactive atoms was proposed by Sello and Termini [147]. Rules in mentioned NameRxn system lack generality and thus they sometimes skip reactions belonging to given type.

Thus in the work of Schneider et al [100] several machine learning techniques were used to make classification model trained on dataset of reactions, extracted from patents and annotated by type using NameRxn tool. Information on conditions could help with classification since reactions of a given type are usually performed in similar conditions. Thus, authors used very unusual fingerprints: they substracted from product fingerprint reagent one and summed result with fingerprint of small-molecule reactant. Several RDKit molecular fingerprints were used for reaction fingerprint creation [148]. The created model were used to annotate reactions that were skipped by rule-based approach.

2.4.2.3. Reaction conditions prediction

Prediction of optimal conditions for chemical reactions is very interesting but yet really insufficiently explored topic in reaction modeling. One of the first work in this direction was the one by Struebing et al [14] where optimal solvents for nucleophilic substitution reaction were predicted. Proposed approach used quantum chemical calculations of transition states of reactions in a given solvent with the following LFER-like equation that united activation energy and solvent parameters. The latter was used to find solvent that potentially enhances reaction rate and the procedure repeated until self-consistency reached. As the result authors reported about new found solvent for reaction that increased reaction rate on 40%. This approach was quite resource-consuming – one prediction required some days of calculation.

Chemoinformatics approach in this case could be much faster. Marcou et al [93] proposed a classification model that could predict general type of catalyst and solvent for Michael reactions. The model was built using manually annotated set of 193 reactions. Several descriptor types (reagent-based, difference fingerprints, CGR-based fragments) and machine learning techniques were benchmarked. The best model had balanced accuracy 0.85± 0.15. The model was successfully tested on external set and was capable to predict feasibility conditions for new 50 Michael reactions.

Recently, approach to assess optimal conditions for hydrogenation reactions was proposed by Lin et al [149]. The training set reactions were extracted from Reaxys database and curated in fully automatic manner. Optimal conditions were predicted using similarity based approach: for test set reaction system looks for most similar reaction

center environment in curated database of reactions where required transformation proceeds and where does not. The approach was validated internally and externally and has shown very good results.

2.4.2.4. Reaction yield prediction

Prediction of expected yield for reaction running under certain condition could be rather interesting for synthetic chemist to assess yield of multistep procedure or ensure that reaction conditions were selected properly. There are two very recent publications that dealt with yield prediction.

Very recently Skoraczyński et al [150] modeled 425 000 of reactions from Reaxys. The authors predicted yields and times of reactions with several types of descriptors. The author claimed that the results of their work were somewhat negative but tended to be thought-provoking. Error in prediction of yields and reaction times are 35% and 25%, correspondingly. Classification models have accuracy 65% for prediction of reactions of class with high yields (>65%) [150]. Such limited prediction was interpreted as consequence of imperfectness of descriptors but one should take into account that yield of reaction is a very noisy parameter.

In the work of Ahneman et al [151] high-throughput reaction screening approach for collection data on yield of Buchwald-Hartwig reaction was presented. The yield was predicted using different machine-learning techniques and quantum-chemical descriptors. Unlike aforementioned work, model shows very good performance on out-of-bag validation: RMSE was substantially lower than in previous work - 11% and R^2=0.91. Such a drastic difference between two articles could be partially explained by different dataset (big and noisy data from Reaxys vs relatively small automatically collected data).

Chemical reactions study and discovery is a central topic in synthetic and quantum chemistry. But as it comes from the review reaction modeling is really poorly explored topic in chemoinformatics. There are few sporadic publications without systematic efforts for the development of approach adapted for reaction modeling taking into account reaction specifics. The goals of present study immediately come from previous studies. In this PhD thesis we want to contribute to reaction modeling by (i) collection of big dataset of kinetic and thermodynamic properties of chemical reactions, (ii) development of workflow for modeling reaction characteristics using machine learning techniques, (iii) development of specific methodologies for reaction data curation, (iv) development of novel descriptors to model chemical reactions taking into account reaction conditions, (v) modeling reactions studied previously and a new one , creation of universal models predicting reaction characteristics of different types in variety of conditions, (vi) development of server for publication of reaction models.

Chapter 3.

Computational techniques used in the study

In this chapter technologies used in the work for reaction modeling will be reviewed. Moreover, some clarification of technical aspects of the workflows used in the work is given.

3.1. Quantitative Structure Reactivity Relationship (QSRR) modeling strategy

QSAR/QSPR methodology application to chemical reaction proposed and used in the work will be described in the following chapters. In order to discriminate it from other approaches we called it for convenience Quantitative Structure Reactivity Relationship (QSRR).

The general workflow of QSRR is the same as for QSAR study and starts with generation of descriptors for the objects of the dataset. The next step is selection of model validation procedure and machine learning method. Hyperparameters of machine learning method and fragment descriptors type maximizing model performance are selected using stochastic algorithms. The parameters shown best performance in cross validation are used for final consensus model building. All the steps will be discussed in chapters below.

3.1.1. Reactions Descriptors

Every modeling procedure starts from descriptors generation since machine learning methods usually require input represented as numerical vector. In this work two types of fragment descriptors were used for modeling of reactions – SiRMS and ISIDA.

The SiRMS descriptors were used for modeling of reaction for the first time in this work. The procedure of generation descriptors is fully described in the article on E2 reactions modeling in the chapter 6.2, as the procedure that was used only in that part of the work and not included in general workflow. SiRMS descriptors were calculated using in house software developed by Dr. Pavel Polishchuk available at https://github.com/DrrDom/sirms.

The ISIDA substructural fragments based on CGR were commonly used in the work. ISIDA Fragmentor tool with the following settings were used for the descriptor set generation (see chapter 2.3.1.1 for options description):

- Minimal length from 2 to 4;
- Maximal length from 3 to 8;
- Fragment types: sequences and augmented atom type descriptors including description of only atoms, only bonds or both atom and bonds;
- Formal charge on atoms were optionally explicitly included into fragment description;
- Shortest paths or all possible paths were optionally explored in sequence descriptor generation.

As a result, 616 fragmentation schemes were used. All fragmentations were supplemented by 13 descriptors of solvents. Each solvent was described by special descriptors that represent polarity, polarizability, H-acidity and basicity: Catalan SPP, SA, and SB constants [108–110], Kamlet–Taft constants α, β, and π* [111–113], four functions describing solvent polarity depending on the dielectric constant ε (Born $f_B = \frac{\varepsilon - 1}{\varepsilon}$ and Kirkwood $f_K = \frac{\varepsilon - 1}{2\varepsilon + 1}$ functions, $f_1 = \frac{\varepsilon - 1}{\varepsilon + 1}$, $f_2 = \frac{\varepsilon - 1}{\varepsilon + 2}$), three functions describing solvent polarizability depending on the refractive index measured by D-line of sodium spectra at 20 degrees Celsius n_D^{20} ($g_1 = \frac{n^2 - 1}{n^2 + 2}$, $g_2 = \frac{n^2 - 1}{2n^2 + 1}$, $h = \frac{(n^2 - 1)(\varepsilon - 1)}{(2n^2 + 1)(2\varepsilon + 1)}$, refractive index is denoted as n for simplicity). In order to describe water-organic solvent mixtures descriptor specifying the molar ratio of organic solvent in water was added. The inverse absolute temperature, 1/T (in Kelvin degrees) was also used as temperature descriptor.

3.1.2. Model Validation

The best parameters for machine learning methods are selected within 5-fold cross-validation procedure. Cross-validation is a model validation technique for assessing predictive power of the model. In 5-fold cross-validation procedure the whole dataset is divided into five almost equal parts. Four of them were used for model building (training set) and one part is used to test predictive performance (test set), see Figure 16. The procedure is repeated in such a way that every part once is used as test set. The

evaluation of model is done by taking into account only predictions for the reactions when they were in the test set.

	Model 1	Model 2	Model 3	Model 4	Model 5
Fold 1	TEST	TRAIN	TRAIN	TRAIN	TRAIN
Fold 2	TRAIN	TEST	TRAIN	TRAIN	TRAIN
Fold 3	TRAIN	TRAIN	TEST	TRAIN	TRAIN
Fold 4	TRAIN	TRAIN	TRAIN	TEST	TRAIN
Fold 5	TRAIN	TRAIN	TRAIN	TRAIN	TEST

predicted values for evaluation

TEST	Fold 1
TEST	Fold 2
TEST	Fold 3
TEST	Fold 4
TEST	Fold 5

Figure 16. Cross-validation procedure of machine learning parameters evaluation.

3.1.3. Machine-Learning methods

Machine learning algorithms can fit some flexible function to describe the data, and thus it could extract rules from observed data without being explicitly programmed [7]. Such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, based on sample inputs. Machine learning is widely used in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible.

3.1.3.1. Support Vector Machine

Among the most popular state-of-the-art algorithms is **Support Vector Machine (SVM)** - a supervised non-probabilistic learning method used for classification (called SVC or simply SVM) and regression analysis (SVR).

In classification task a set of training examples is given, each labelled as belonging to one of two classes, an SVM training algorithm assigns new objects some class. An SVM looks for hypersurface in some descriptor space that separate objects of different classes. The surface is adjusted in a way that the gap (called margin) between objects of opposite classes is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on definite side of the surface they fall on.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using so called kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Consider some training data D that are represented by a set of *n* objects with known class:

$$D = \{(\boldsymbol{x_i}, y_i) \mid \boldsymbol{x_i} \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^{n}$$

where the y_i is either 1 or -1, indicating the class to which the point $\boldsymbol{x_i}$ belongs. Each $\boldsymbol{x_i}$ is a *p*- dimensional real vector. Optimal classifier is found by selection of the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$, Figure 17. Any hyperplane can be written as the set of points $\boldsymbol{x}$ satisfying condition

$$\boldsymbol{w} \cdot \boldsymbol{x} - b = 0$$

where $\boldsymbol{w}$ is the (not necessarily normalized) normal vector to the hyperplane. The parameter $\frac{b}{||\boldsymbol{w}||}$ determines the offset of the hyperplane from the origin along the normal vector [152], Figure 17. Maximization of margin was shown to be equivalent to maximization of $\frac{2}{||\boldsymbol{w}||}$. One could notice that correct classification of objects by plane with given $\boldsymbol{w}$ and *b* means that

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} - b) \geq 1$$

Thus the following task of quadratic programming is posed:

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \rightarrow \min$$

$$y_i(\boldsymbol{w} \cdot x_i - b) \geq 1$$

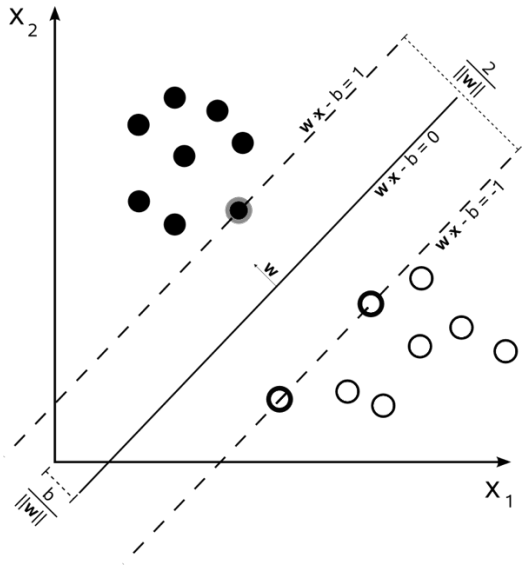


Figure 17. Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

A support vector regression(SVR) was proposed in 1996 by V. Vapnik, et al [3] as a development of SVM method for continuous variables prediction (Figure 18). To find the original regression $y = \boldsymbol{w} \cdot \boldsymbol{x} + b$ for points T={($\mathbf{x_1}$,y_1),($\mathbf{x_2}$,y_2),...} is to find $\boldsymbol{w}$ and b minimizing

$$\frac{\|\boldsymbol{w}\|^2}{2} + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

where C is the cost that define the severity of the penalty for points whose predicted value deviates from experimental one more than ε, Figure 18:

$$y_i - (\boldsymbol{w} \cdot \boldsymbol{x_i}) - b \le \varepsilon + \xi_i$$
$$(\boldsymbol{w} \cdot \boldsymbol{x_i}) + b - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i \ge 0$$
$$\xi_i^* \ge 0$$

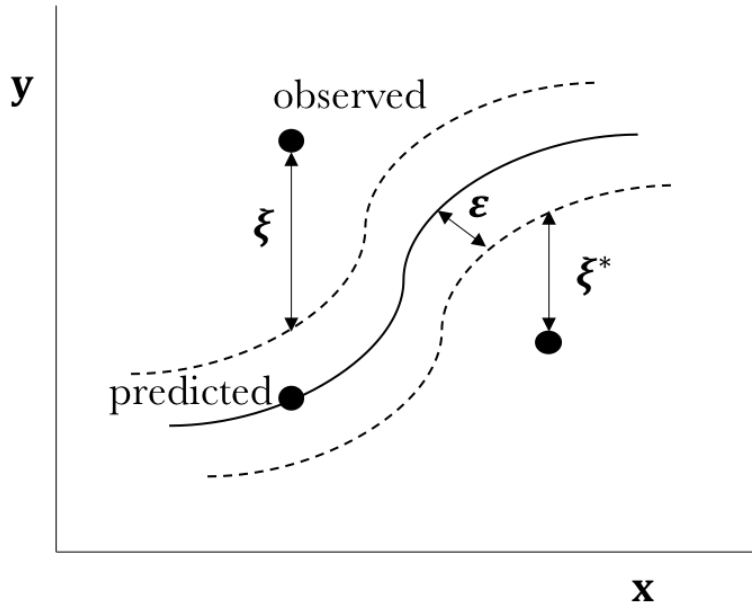


Figure 18. Illustration of parameters of SVR method.

Kernel trick based implemented using dual formulation of SVR optimization task is utilized to build non-linear regression model.

3.1.3.2. Random Forests

Another popular machine learning method is **Random Forest (RF)**. Random forests are a combination of tree predictors such that each tree is built on random subset of objects sampled independently and with the same distribution for all trees in the forest [153]. This method is evolution of simple decision tree [154], where each node of tree is a rule based on value of some descriptor that selects subset enriched with objects of given class. Thus a tree is an ensemble of rules that leads to decision about class attribution of an object.

In random forest tree predictors are united in a way that every tree gives independent prediction and final decision about class attribution is based on simple majority of votes. For random forest regressor trees for prediction of continuous variables are used and averaging predicted values of trees makes decision. Each tree is built on subset of objects selected by bootstrapping [155]. Moreover, during tree learning (growing) procedure, when algorithm looks for feature giving best split in branching, random subset of descriptors is considered rather than all descriptors. The main hyperparameters in RF is number of trees and number of descriptors in random subset.

The more trees are generally better for this machine learning method, but it also leads to increase in resources and time consummation. Optimal number of descriptors in random subset is usually adjusted using cross-validation or on the basis out-of-bag sample prediction.

3.1.3.3. GTM

As a visualization tool **Generative Topographic Mapping (GTM)** was used.

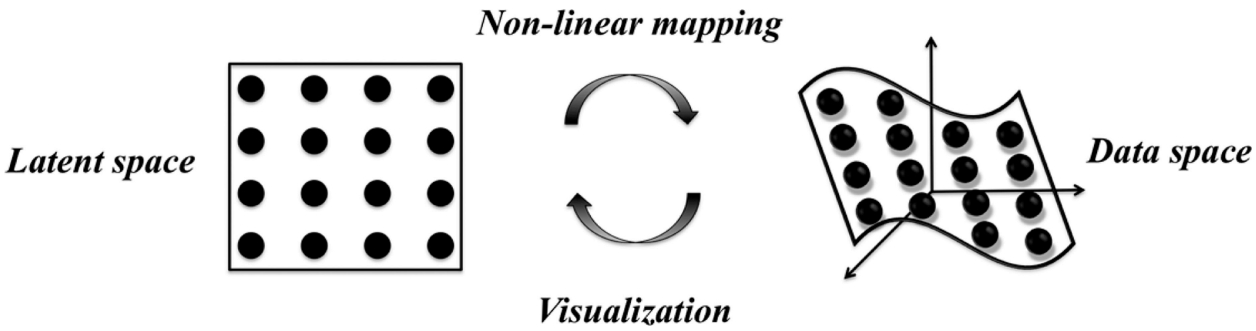


Figure 19. Schematic representation of the GTM algorithm: nonlinear mapping of grid nodes in 2D latent space onto a manifold in D-dimensional data space.

In GTM, each point in the low dimensional (usually 2D) latent space (LS) is mapped onto the manifold embedded in a high-dimensional data space (input space, IS), Figure 19.

The manifold is defined by a mapping function y (x; W) assessed with the help of *m* radial basis functions (RBFs) of width *w* regularly distributed in LS. The latent space is covered by a mesh containing *k* nodes each of which corresponding to a normal probability distribution (NPD) centered on the manifold in IS. Ensemble of NPDs is used to compute a posterior probability for a data point t_n in D-dimensional IS to be projected onto a node x_k:

$$p(x_k|t_n, \mathbf{W}, \sigma) = \frac{\mathrm{p}(t_n|x_k, \mathbf{W},\ \sigma)}{\sum_k\ \mathrm{p}\,(t_n|x_k, \mathbf{W},\ \sigma)}$$

where **W** is a parameter matrix and σ the variance of the distribution of $\boldsymbol{t}$:

$$p(\mathbf{t}|\mathbf{x}) = \left(\frac{1}{2\pi\sigma}\right)^{\frac{D}{2}} \exp\left\{-\frac{1}{2\sigma}\|\mathbf{y}(\mathbf{x};\mathbf{W}) - \mathbf{t}\|^2\right\}$$

The log likelihood of the whole data set is calculated according to equitation:

$$\mathcal{L}(\mathbf{W},\sigma) = \sum_n ln\left\{\frac{1}{K}\sum_k p(t_n|x_k,\mathbf{W},\sigma)\right\}$$

The GTM is optimized with an expectation-maximization (EM) algorithm using data likelihood ($\mathcal{L}$) as the objective function (the best GTM map corresponds to the highest $\mathcal{L}$). The mapping depends on four parameters: the number *m* of RBFs, the grid resolution *k*, the RBF width *w*, and the weight regularization coefficient *l*. The latter is used for re-estimating the **W** parameter matrix and influences the flexibility of the manifold. Notice that a too flexible manifold, although nicely approximating the training data, may lead to overfitting when training set objects ideally fit manifold but new data points are located too far from it. In this work GTM map parameters *w* and *l* were optimized by likelihood.

GTM approach is unsupervised machine learning methods, as it does not use target values while learning. Obtained map however can be colored by different properties to define zones of selectivity on the map, as it was done in our previous work [156]. This approach is used in this work for the analysis of S_N2 reactions chemical space.

3.1.4. Applicability domain

All QSAR models are strongly connected with the training set. Due to limitated number of trained instances prediction could be done more or less confidently for objects similar to training set ones. The set of objects that could be predicted confidently by a model is called its applicability domain (AD) [157], Figure 20. The test set compound 1 (in green) is inside the AD and its prediction is considered reliable while the test set compound 2 (in red) is outside and therefore, its prediction is considered unreliable. The concept of applicability domain is extensively studied in chemoinformatics [157–161] since datasets are usually rather limited and machine learning methods being good in interpolation are unstable in extrapolation.

One of the most simple applicability domains is bounding box [158]. It states that an instance is out of model's applicability domain if descriptor values for it are not within the min-max ranges valid for training set. The Bounding Box techniques by definition encompasses so-called Fragment Control: if the data set encoded in structural fragment descriptors, then any molecule of the test set possessing a new structural fragment considered to be out of AD.

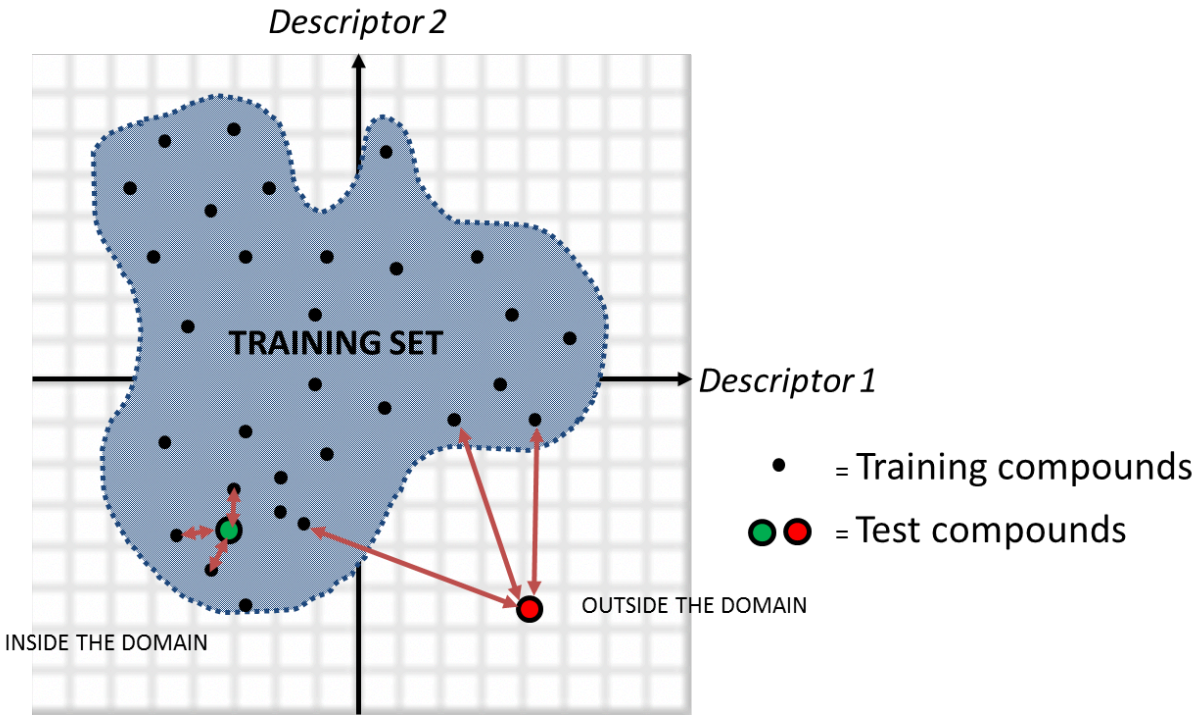


Figure 20. Representation of the Applicability Domain boundary in chemical space.

3.1.5. Genetic Algorithm driven parameters optimization

The fragment descriptor types (fragmentation scheme) and hyperparameters for machine-learning methods were selected within stochastic search approach. As an algorithm was selected program by Horvath [162] with genetic like algorithm of search. Predictive performance of models based on given vectors of parameters to be optimized is checked within cross-validation procedure. The first generation vectors are initiated unsystematically. Two sources of new vectors come from this time : randomly generated and generated by crossover using best-performing vectors. This strategy helps to avoid full iteration of all possible variants and gives reasonably good results in acceptable time.

3.1.6. Model publishing

Modeling procedure was made by in-house program called ChemoInformatics and Molecular Modeling Lab tools (CIMMtools). This program produce special compressed container with model. Such containers are stored at our server and are used by special dispatcher. All obtained models are available at http://cimm.kpfu.ru/predictions – visual interface for working with model dispatcher. Moreover, server have special module, that takes input reactions and makes all preprocessing before prediction, according to specification of model that will be used. Then model gives prediction for curated reaction and gives evaluation of prediction by bounding box applicability domain. The server can handle several users simultaneously or make a queue of queries. Usage of models is free of charge, but requires registration procedure on the site. The example of model application is given in chapter 4.4.

3.2. Matched Molecular Pairs

Matched Molecular Pairs (MMP) have generally been defined as "molecules that differ only by a particular, well-defined, structural transformation" [163]. The analysis of such pair is usually used by cheminformatics to see smoothness in change of the properties of two molecules that differ only by a single chemical transformation, such as the substitution of a hydrogen atom by a chlorine one (Figure 21). Since the structural difference between the two molecules is small, any experimentally observed change in a physical or biological property between the matched molecular pair is attributed to substituent effect.

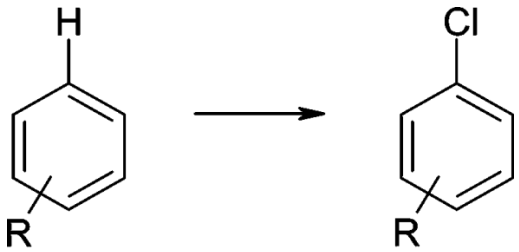


Figure 21. Example of MMP due to substitution of hydrogen by chlorine.

In this work the MMP approach was used for reactions for the first time. The concept was applied to CGRs of S_N2 type reactions to monitor the influence of substituents on reaction rate (Figure 22).

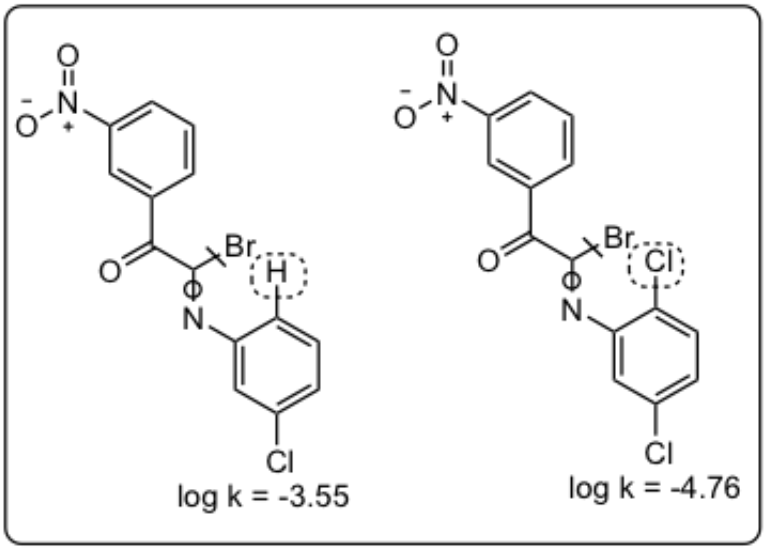


Figure 22. Example of MMP for reactions encoded by CGR.

Matched molecular pairs for CGR (called matched reaction pairs) was built by means of OCHEM.eu site that was modified by its owner I. Tetko to accept CGRs. Matched reaction pairs were used for verification of reaction mechanism (see chapter 4.2.1.4).

3.3. Quantum Chemistry calculations

Density functional theory (DFT) is a computational quantum mechanical modelling method used in physics, chemistry and materials science to investigate the electronic structure (principally the ground state) of many-body systems, in particular atoms, molecules, and the condensed phases [164]. DFT states that ground state energy of the system could be expressed as the functional of the electron density unlike classic quantum chemistry (Hartree-Fock and post-Hartree-Fock methods), which relates energy with the many-body wavefunction. Whereas the many-body electronic wavefunction is a function of 3N variables (the coordinates of all N particles in the system) the electron density is only a function of x, y, z -only three variables. To calculate energy of the molecular system and other properties one need to know exchange-correlation functional that approximates dependence of exchange and correlation energy on electronic density or so called Kohn-Sham orbitals.

In this work for DFT calculations Priroda [165] program was used, that is one of the fastest DFT code due to approximation of Coulomb and exchange-correlation terms

implemented by Laikov [166]. PBE exchange-correlation functional was used [126]. Built-in triple-zeta split valence basis set (called 3z within program itself, equivalent of Schäfer's TZVP basis [167]) was selected. This level of theory was used for geometry optimization, transition state localization, reaction path exploration and Intrisic Reaction Coordinate following along. Transition states were found with saddle point search algorithm of Priroda program.

Gaussian program [168] was used for solvation free energy calculation. In this case IEF-PCM model [4, 169] with SMD parameters for non-electrostatic terms [170] was utilized. Geometry optimization and Hessian calculation in solvent was performed in Gaussian program [171] using PBEPBE/6-311++G(d,p) level.

All structures under discussion were optimized using default Gaussian and Priroda program algorithms. Many different starting geometries were used in order to enhance chance to find global minima of energy. Frequency calculation supported that discussed structures have the right set of Hessian eigenvalues: all positive frequencies for molecules and intermediates, one imaginary frequency for transition states. The details are shown in the chapter devoted to 4.3.3.

3.4. Model implementation

For model publication, a client-server application was implemented. The description of interaction with client side is described in this chapter.

All QSRR models are available at cimm.kpfu.ru/predictor after registration procedure. The first page of predictor contains User manual and tools for interactive creation of molecules or reactions (Add task, Figure 23) and button for files upload. After creation of some dataset one should click the button "Validate" to check structure for correctness.

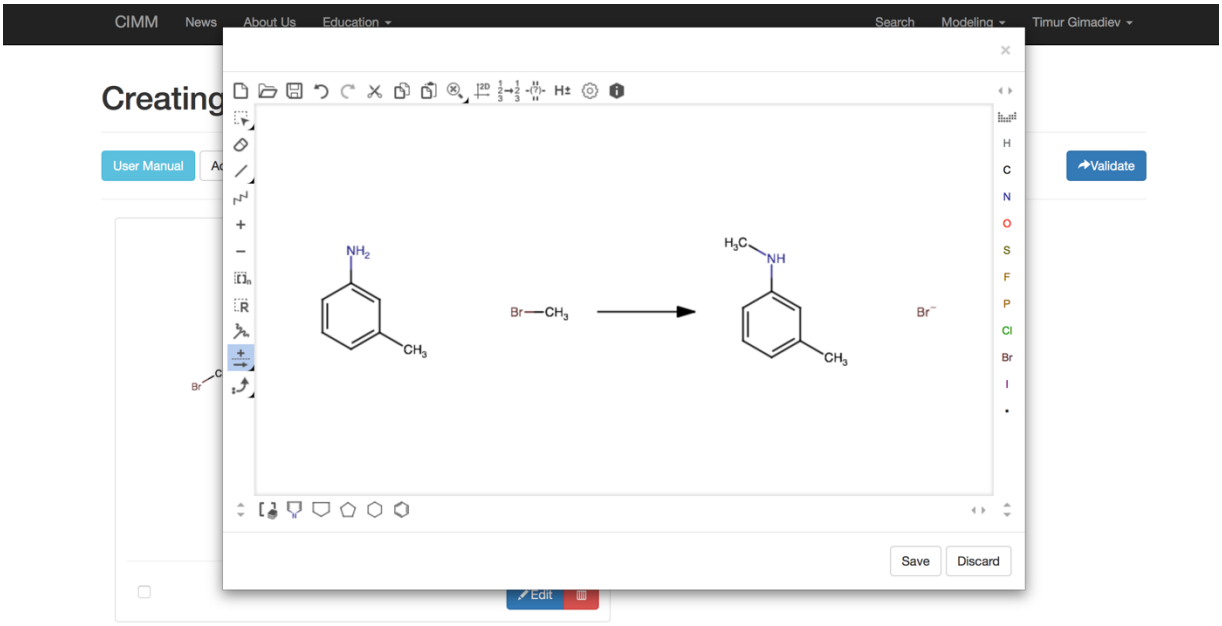


Figure 23. Interactive creation of reaction with Marvin Web application.

After validation procedure, new window with AAM established by embedded algorithms is given to user to check it and modify if needed. On the same page one can select the model to apply and choose conditions (Figure 24). If single molecule is entered, reaction models are not available.

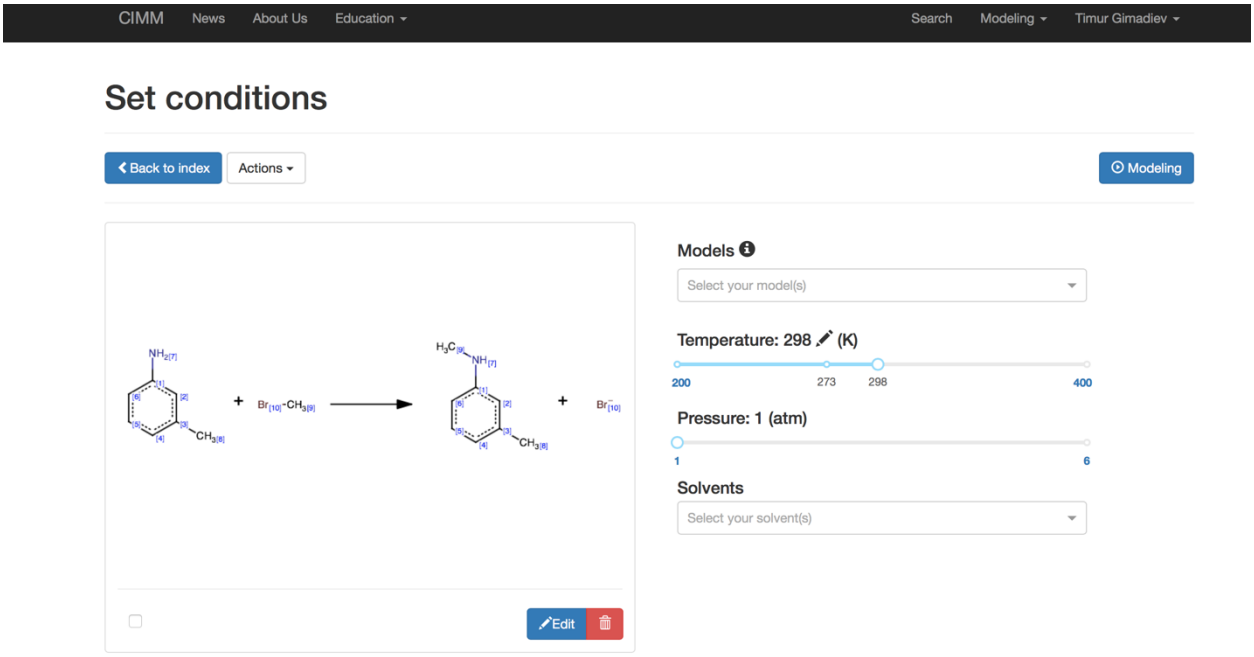


Figure 24. Window with selection of model and conditions for modeling.

After the selection of all parameters and clicking “Modeling” button, the server returns the result of modeling (Figure 25). At this window predicted value and variance of prediction within consensus is given. Also results of AD application are given in line “Trust of prediction”. The trustworthiness of prediction depends on fragment control applicability domain and variance of prediction. If prediction of test set appears to be out of fragment control AD for more than certain percent of models inside consensus or variance exceed certain value, trustworthiness will be lowered from “optimal” to “good”. If two conditions are fulfilled simultaneously it becomes “poor”. The reason of “Trust of prediction” lowering is given below in “Distrust reason” section.

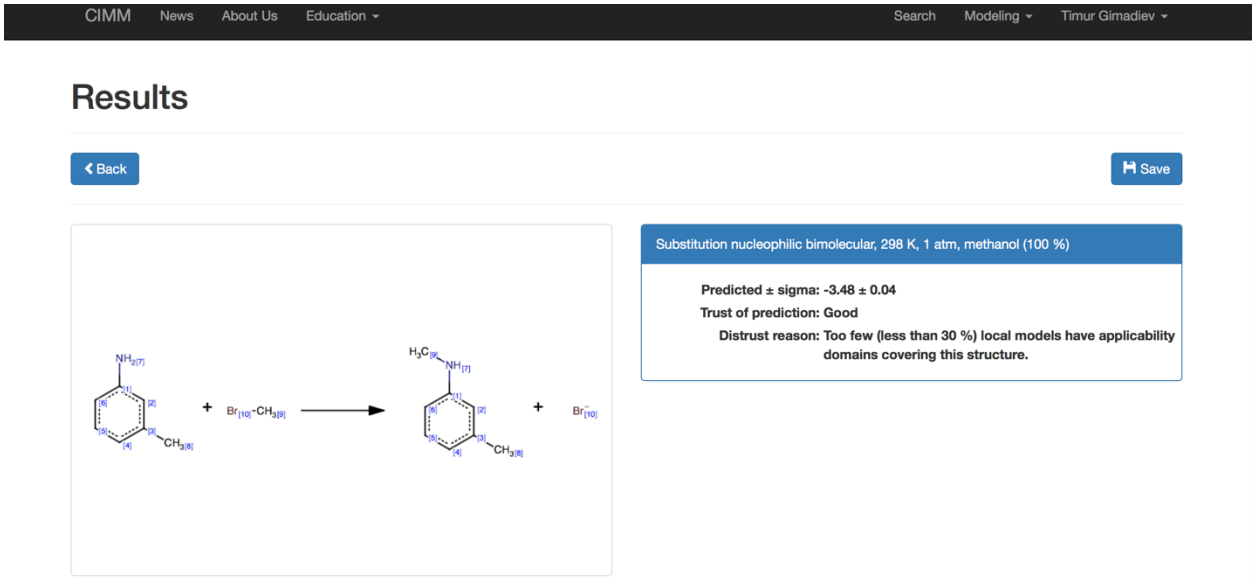


Figure 25. Results of modeling procedure.

Hense, modeling procedure is user-friendly and includes 3 simple steps: input – validation – results. All the preprocessing and descriptors generation is written in the model itself and hidden from the user. As a result fast and easy tool for reaction modeling was obtained.

3.5. In-house software tools

All the tools developed in the work are written in Python 3.5 language [172]. For handling graphs NetworkX library [173] is used. Sklearn [174] library is used for machine learning method (mainly SVM) application. Databases are implemented using Pony ORM [175] solution linked to PostgreSQL database management system [176]. Other dependencies include the following libraries as well: Pandas (data operation) [177],

NumPy (basic math and statistics) [178], ChemAxon Web Services (chemical data processing) [179]. Client side for model publishing server was written in JavaScript.

Using aforementioned tools several new tools and libraries were created. For databasing chemical reactions in-house chemical cartridge CGR DB was implemented. CGRtools library were created for reaction information management using CGR approach. For QSRR modeling CIMMtools library was implemented. All the developments were supervised by Dr. R. Nugmanov and Dr. T. Madzhidov from Chemoinformatics and Molecular Modelling Laboratory of KFU. The final code was implemented by Dr. R. Nugmanov on the base techniques that were developed during this work.

GTM method for visualization and modeling as well as the Fragmentor program for fragment descriptor calculation and Genetic Algorithm optimizer for model hyperparameter selection are the developments of the Laboratory of Chemoinformatics of the University of Strasbourg, supervised by Prof. A. Varnek.

Chapter 4.

Data collection, cleaning and representation

4.1. Development of a comprehensive reaction database

Nowadays, modern chemical databases (Reaxys, CASReact, InfoChem, etc) contain some 100 million reactions. They store every reaction as a separate record that consist of reagents, products and some additional information. Despite the fact that amount of known reactions is incredibly high, there are almost no data for modeling reaction rates, almost no data on other kinetics parameters annotated in databases. Usually only reaction time and yield of desired product are annotated according to the needs of synthetic chemists. Recent work shows that these parameters could hardly be modelled [150]. Thus the content of existing reaction databases is not satisfactory for reaction characteristics prediction.

From the other hand, standardization of reaction is tricky and not so well elaborated as for molecules. The reactions are usually represented in the way they were in original article, databases are just storage of information about successful reactions from the article with no moderation. Some preprocessing is done to merge reactions with the same structures of reagents-products. Notice that reactions with the same transformation and conditions are usually saved. Usually reactions in such databases are stochiometrically unbalanced (some reactants and products are lost), its additives and conditions are represented in text fields in non-standardized way and could be lost, for example, only half of reaction extracted from Reaxys had temperature in corresponding field [149].

To overcome the problem of lack of data, a new database of kinetic measurements of reactions was collected. Relational database was designed to deal with big amount of data on different reactions that can differ by transformation or/and conditions.

In this chapter the workflow of standardization, storage, representation and filtering of reactions will be described. This workflow appeared as the result of further development of CGR concept and its representation.

4.1.1. CGR technology development

The new challenges were come across during the realization of new database and standardization techniques that are related to the needs in further development of CGR technology, its adaptation to the challenges we faced upon reaction modeling and storage.

In this chapter the developments in CGR technology such as introduction of dynamic atom concept, more useful way of CGR storage and creation of reaction center signatures based on CGR are described. Such features are needed to increase the speed and easiness in processing and storage of reactions.

4.1.1.1. CGR technology extension

In this chapter the reasons for dynamic atom concept introduction into CGR approach is described.

Usually for representation of molecules and reactions reduced graph representation with implicit hydrogens is used. It assists not only to save data storage space, but reduce number of fragments descriptors that will be generated for the structure. Usually it does not cause problems, since the change of hydrogen position is encoded in change of bonds orders of between heavy atoms. But in some cases, e.g. for zwitter-ion formation reactions (type of tautomeric equilibria) or *SN2* reactions with OH- or NH-containing nucleophiles, implicit hydrogen migration is not captured in CGR since bonds between heavy atoms are not touched. Thus, only charge of some heavy atom is changed in reaction. There is several ways to encode hydrogen migration in CGR:

- Explicit hydrogen representation. Dynamic bond will appear between heavy atom and proton and thus migration of hydrogen will be encoded, Figure 26. From the other side, it will increase number of descriptors and the time required for its calculation.
- Usage of pseudoatoms representing atomic charge. The idea was to represent additional features of atom as connections from it to pseudo atoms standing for charge label (+1/-1/0), stereo label (R/S), etc. It was proposed by Jauffret [41]. Migration of hydrogen will be encoded as dynamic bonds between heavy atom and virtual atom that represent charge, Figure 26. However, it causes a problem with visualization and CGR could hardly be represented by widespread SDF format. Additional atoms will also drastically increase number of descriptors that decrease speed of model building.
- Dynamic atom. The third idea was to introduce analog of dynamic bond for one sole atom – dynamic atom. Dynamic atom encodes changes in some atomic property (charge, stereochemistry) in chemical reaction, Figure 26. This will solve a problem with encoding proton migration, as atomic formal charge is changed that could be captured by dynamic atom label. Generally the number of descriptors does not increase, as dynamic

charge is a simple atomic label. This label can be correctly represented in SDF format and visualized by common visualization tools (e.g. ChemAxon Marvin View). Thus, this approach was implemented in our SDF specification.

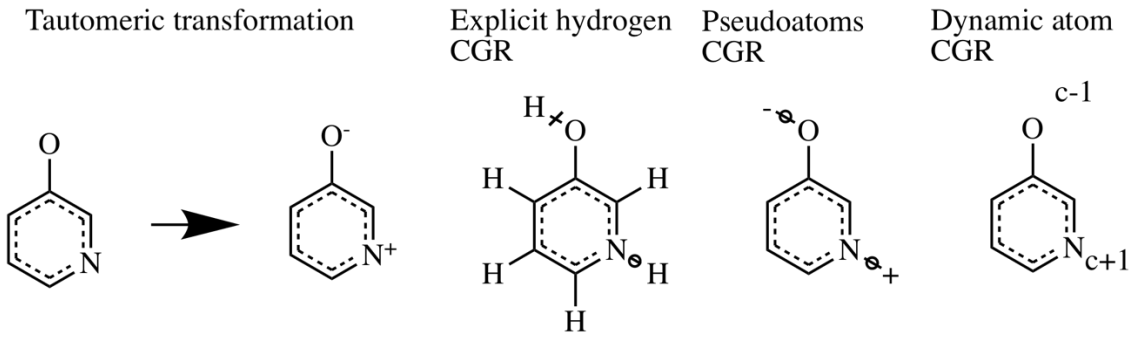


Figure 26. Dynamic atom representation in tautomeric equilibria. (ChemAxon Marvin Sketch representation [180]). Left side original tautomeric transformation, right side CGR representation with dynamic atoms.

Dynamic bond and dynamic atom labels could be encoded using standard fields in CTfile specification [181] and CGR could be stored in SDF format without non-standard fields. This representation gives opportunity to use ordinary programs for SDF file visualization without any changes.

4.1.1.2. CGR storage in SDF format

In this chapter reasons of SDF format usage for CGR storage will be discussed and features allowing to do it natively for SDF file format will be described.

MDL/SDF [25] is one of the most popular formats in the world for storage information about molecules. The CGR looks like pseudomolecule, this similarity was used to store CGRs in SDF file like an ordinary molecule. The solution is compatible with a standard MDL/SDF format. Common bonds are encoded according to MDL/SDF specification. Dynamic bonds are encoded additionally in bond section as 8 ("any bond" according to the CTfile specification [181]).

The type of dynamic property label is encoded in MOL file properties section using standard V2000 connection table specification of specific groups (Sgroups) in CTfile format [181], see Scheme 1 below. First line in properties section starting "M STY" in property section specifies number of dynamic properties. Then, each dynamic property is

encoded in 4 fields. It is shown from the first field (starting with "M SAL") which atom or bond CGR property corresponds to. Precisely first field encode atomic array that can contain 1 entry (for dynamic atom) or 2 entries (for dynamic bonds). The second line (starting with "M SDT") contains title of property: "dynatom" and "dynbond" for dynamic atom and bond correspondingly. The third line starting with "M SDD" contains information needed for visualization of CGR – coordinates on the screen where to place the label. Forth line starting with "M SED" contains value of dynamic atom and bond label. Dynamic bond labels representing changes in bond orders are given using the following scheme "R>P", where R and P bond orders of the bond in reagents and products (0 for absence of bond). Dynamic charge is represented using "c+1" and "c-1" labels standing for charge increase or decrease by 1 correspondingly. The example for SDF file with dynamic bonds and dynamic charge is shown on Scheme 1. More options are given in CGR Whitepaper describing other labels.

Scheme 1.

```
SDF
=================
Standard atoms block
=================
Standard bonds block
22 23  8 0  0  0  0
=================
M  STY  2   1 DAT   2 DAT                          # 2 CGR properties specified
M  SAL   1  2  22  23                              #prop1– for atoms 22,23 pair
M  SDT   1 dynbond                                 #prop1 – dynamic bond
M  SDD   1    6.2460    6.0953    DAU   ALL  0    0  #prop1–position on screen
M  SED   1 1>0                                     #prop1-single bond cleavage
M  SAL   1  1  12                                  #prop2 – for one atom - No 12
M  SDT   1 dynatom                                 #prop2 – valence state dynamic
atom
M  SDD   1   -18.8792    0.8913    DAU   ALL  0    0  #prop2–position on screen
M  SED   1 c-1                                     #prop2 – formal charge reduced
by 1
```

The dynamic labels were proved to be rendered by ChemAxon programs and the latter could be used for creating CGRs in proposed specification.

4.1.1.3. CGR signatures

At this chapter such newly developed features as canonical CGR and reaction center signatures will be described. These two techniques is used for fast duplicated transformation search and reaction classification without standard graph embedding procedure.

For canonical numbering Morgan [182] like algorithm was developed based on prime number usage as in paper of Ihlenfeldt and Gasteiger [183]. In it we use the property to that every number could be factorized uniquely. Unlike classic Morgan algorithm where extended connectivity index for a given atom is calculated summing up corresponding indices of surrounding atoms, we use multiplication of prime numbers representing atomic connectivity. In this case, there is no possibility that two atoms have the same extended connectivity index by chance.

Conceptually algorithm includes several steps. The first step includes ranking of tuples describing atomic properties (element, connectivity, charge, isotope, stereolabel, etc) of every atom in molecule as in CANGEN algorithm for SMILES canonization [20]. Based on the rank obtained prime number is selected from prime number table. Thus every atom assigned prime number that coincides for atoms having the same type. This number used to define amount of unique atoms. At the next step this number for a given atom is multiplied to the corresponding numbers of its neighbors and again used to define amount of unique atoms. Obtained numbers are sorted and again used for selecting prime number from table. This procedure is repeated several times, until the number of unique atoms will not change for 3 following cycles. This final numbers are sorted and the ranks are used as canonic atom numbers. The latter are applied for linear string generation using rules similar to canonic SMILES generation [20].

For CGR generation current implementation generates canonic numbers for all molecules in reagent and product side and SMIRKS-like string is generated using OpenSMILES compliant rules [184]. AAM is renumbered in a way that in reagent side AAM is sequentially numbered. Reactions where atoms are mapped in the same way (even if numbers for AAM encoding are different) will give the same text representation of graph, that we call reaction (or CGR) signature, CGRS (

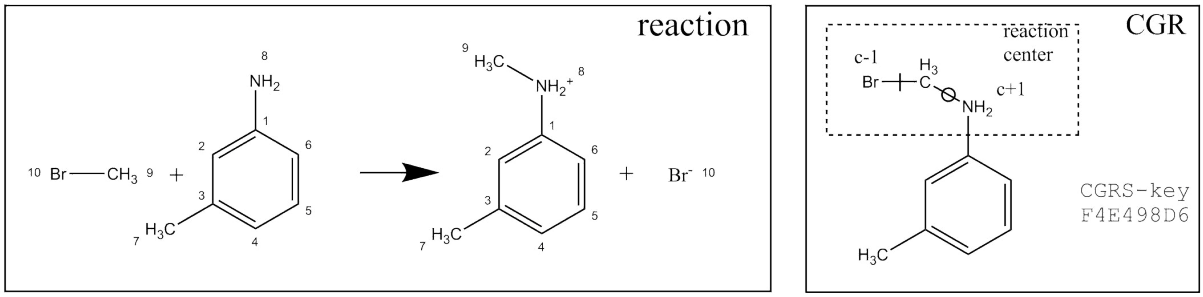
CGRS for reaction:	[C:1]:1(:[C:2]:[C:3](:[C:4]:[C:5]:[C:6]:1)-[C:7])(-[N:8].[C:9]-[Br:10])>>[C:1]:1(:[C:2]:[C:3](:[C:4]:[C:5]:[C:6]:1)-[C:7])(-[N:8]-

[C:9].[Br-1:10])

CSMIRKS for reaction: [C]:1(:[C]:[C](:[C]:[C]:[C]:1)-[C])-[N].[C]-[Br]>>[C]:1(:[C]:[C](:[C]:[C]:[C]:1)-[C])-[N]-[C].[Br-1]

CGRS for reaction center: [C:1](-[Br:2]).[N:3]>>[C:1](.[Br-1:2])-[N:3]

Figure 27). Long CGRS text strings could be hashed for shortness and represented as hashed hexadecimal number called CGRS-key (Figure 24).



CGRS for reaction: [C:1]:1(:[C:2]:[C:3](:[C:4]:[C:5]:[C:6]:1)-[C:7])(-[N:8].[C:9]-[Br:10])>>[C:1]:1(:[C:2]:[C:3](:[C:4]:[C:5]:[C:6]:1)-[C:7])(-[N:8]-[C:9].[Br-1:10])

CSMIRKS for reaction: [C]:1(:[C]:[C](:[C]:[C]:[C]:1)-[C])-[N].[C]-[Br]>>[C]:1(:[C]:[C](:[C]:[C]:[C]:1)-[C])-[N]-[C].[Br-1]

CGRS for reaction center: [C:1](-[Br:2]).[N:3]>>[C:1](.[Br-1:2])-[N:3]

Figure 27. Conventional reaction representation (left) and CGR (right). Cleaved bonds in CGR are crossed, formed bonds are denoted by circle. Dynamic atom option c+1 and c-1 mean that atom acquired positive and negative formal charge correspondingly. Numbers near atoms represent AAM.

We also developed canonical SMIRKS representation of reaction ignoring atom-to-atom mapping that we called CSMIRKS,

CGRS for [C:1]:1(:[C:2]:[C:3](:[C:4]:[C:5]:[C:6]:1)-[C:7])(-[N:8].[C:9]-

reaction:	[Br:10])>>[C:1]:1(:[C:2]:[C:3](:[C:4]:[C:5]:[C:6]:1)-[C:7])(-[N:8]-[C:9].[Br-1:10])
CSMIRKS for reaction:	[C]:1(:[C]:[C](:[C]:[C]:[C]:1)-[C])-[N].[C]-[Br]>>[C]:1(:[C]:[C](:[C]:[C]:[C]:1)-[C])-[N]-[C].[Br-1]
CGRS for reaction center:	[C:1](-[Br:2]).[N:3]>>[C:1](.[Br-1:2])-[N:3]

Figure 27. Mechanism of generation is exactly the same as for CGRS but AAM is not reflected in text string. Transformation with the same set of reagent and products will have same CSMIRKS whatever AAM is.

CGR signature could represent either the whole reaction (as described above) or only reaction center. To create the latter, atoms incident to dynamic bonds or possessing dynamic atom labels are left while others are deleted. For reaction classification and some other tasks one could need more detailed specification of reaction center, thus we implemented possibility to include into reaction center description its first, second, etc environments of dynamic atoms and bonds. CGRS is created with obtained reaction center graph. Resulted reaction center signatures could be used for similarity search in databases in a manner as ICClassify hash codes [185] implemented in SciFinder and Reaxys for reaction similarity search [186]. We used reaction center signatures for identification of AAM errors since signatures are different for precisely and imprecisely mapped reactions. The signature could be visualized and one can clearly understand that it corresponds to correct reaction center or appeared due to AAM error. In such a way one could easily identify reactions with wrong AAM.

All described signatures are quite long strings thus we compressed them using hash function and represented as hexadecimal number. Developed signatures help to match reactions superfast without expensive graph embedding procedure, cluster them and analyze influence of reaction center surroundings.

4.1.2. Data collection

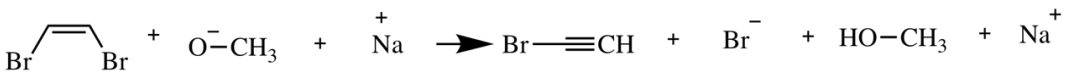
As we already mentioned there exists no database of well-structured data on chemical reactions, especially its kinetic characteristics that are ready for modeling. We decided to collect our own dataset of kinetic and thermodynamic properties of chemical reactions. Data were collected manually from the reference book by Palm [1] and PhD theses that were defended in Kazan Federal University, Russia. The information was extracted manually. In this chapter the rules that were used for selection and annotation of data will be described. These rules were used to minimize efforts needed for data standardization; some of them were introduced to make possible usage of the other modeling approaches rather than CGR used in this study.

Datasets of S_N2 and E2 reaction rate and equilibrium constants of tautomerization was collected from book edited by Palm [1], cycloaddition (CA) reactions rate constants, activation energy and pre-exponential factor were extracted from PhD theses of Kazan Federal University (totally about 20 theses). The databases were created using InstantJChem [187] database management system.

We posed some constraints on reaction to be extracted. For S_N2, E2 and CA reactions, only constants that were declared for bimolecular processes were added to database. Only reactions in pure solvent or water-organic solvent with known molar ratio of solvents were annotated. In all cases reactions with additives and catalysts were ignored, as well as reactions under not standard pressure (other than 1 atm).

The rules of reaction annotation were slightly different for different dataset. Order of molecules in reaction was fixed. Molecules in S_N2 reaction were drawn in following order: substrate, nucleophile, anion, cation -> major product, minor product, anion, cation. For E2 the order was following: substrate, base/reagent, counter ion of base/reagent -> product, cleaved group, changed base/reagent, counter ion of base/reagent (see Scheme 2). For cycloaddition reaction diene was drawn first, dienophile second. Counter cations or anions in all type of reactions are drawn after reagents and products.

Scheme 2.



Usually description of reaction in literature sources included only description of reagents. It is not problematic for most of reaction types to guess the product, however for CA reaction it caused unclear stereoconfiguration or even regioisomer of products, Figure 28. Consequently, CA reaction data was stored as a mixture of all possible reactions with special markers to identify them while dataset preparation.

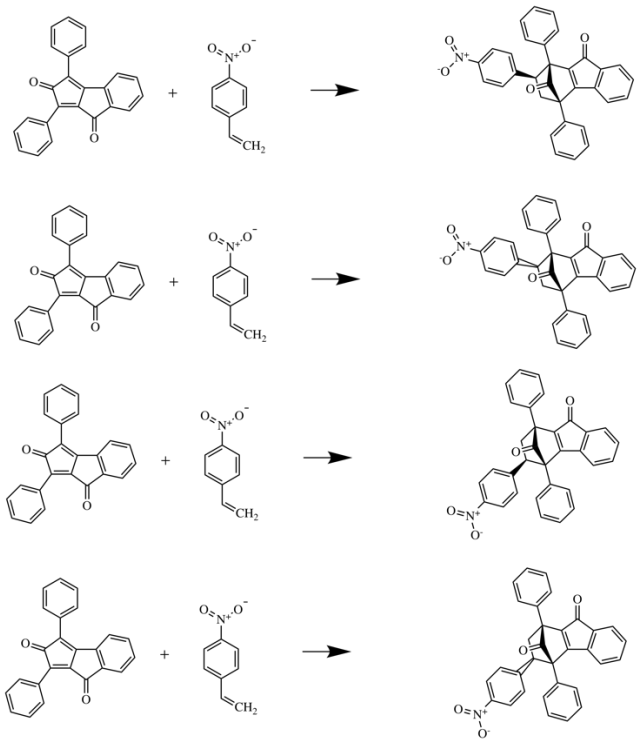


Figure 28. Unclear result of reaction in literature.

The tautomeric equilibria database has its own peculiarities. Only equilibria between two possible tautomers were collected. Since equilibrium is reversible reaction it is important to select which molecules is reagent and which is product. The roles of molecules were assigned according to the type of tautomerism, e.g. for keto-enol equilibria keton was reagent and enol was product.

For all databases ionic compounds were represented as two separate ions. Bonds of metals with heteroatoms – N, P, O, S, Se, Te, F, Cl, Br, I atoms – were considered ionic except of organometallic compounds. All molecules were stored as graphs with implicit hydrogen. All known information about reaction was extracted.

Mentioned rules helped to collect database that have unified representation of reaction. The amount of collected data for each reaction type is given in Chapter 4.1.5.

4.1.3. Data curation

Initial data set should be checked for possible errors in representation or in values of the property. This chapter will describe techniques of data curation and stages of this process. This workflow was used for preparation the dataset for modeling.

Quality of data is a very important issue for building predictive models using machine learning tools. While best practices for data curation in QSAR modeling [188, 189] and chemogenomics data cleaning [190, 191] were published there is no such kind of information for chemical reactions. Reaction data curation is more challenging task than molecular data curation mainly due to the fact that conditions should be taken into account. From the other side Arrhenius law could be used for finding doubtful data on reaction rate. Thus, we developed our own workflow for processing chemical reaction that takes into account the features of reactions as chemical objects.

Reaction cleaning process includes two stages: (1) structure and transformation cleaning, (2) experimental facts cleaning. First step produces standardized transformation representation with mechanistically correct atom-to-atom mapping. The second step is required for discarding duplicates, producing correct tuple transformation-condition-property for modeling. Both steps are based on some conventions. For structure cleaning one should decide how the product is specified, since the chemically correct representation is not implicitly best for the modeling purpose. So, it is not really important that structure should be chemically correct. It is more important, that similar transformations have similar representations. Since different people extracted reactions manually, one have to take into account possibility of different representations of molecules participating in reaction, existence AAM errors. These issues should be fixed.

For the second step one has to decide what condition parameters influence rate constant. Our convention was that concentration of reagents and additives should not influence the reaction rate constant. And thus reactions proceeding under the same temperature and solvent but with different initial concentration of reagents should be considered duplicates.

Usually the same transformation could proceed under different conditions and thus with different rate constant. For further usage in modeling it is more useful to group all reactions having the same transformation. Then dataset could be stored as relational table.

4.1.3.1. Structure standardization

The first step is standardization of reaction representation is unification. In order to find group of reactions with same transformation that differ in conditions, initial database structures was standardized in order to remove different representations of the same molecules using ChemAxon Standardizer [192]. The main problem is representation of aromatic bonds in aromatic or Kekule form. It causes a lot of erroneous dynamic bonds in CGR structure or makes same reactions looks different due to different representations. To fix this problem standardization procedure was included following steps: de-aromatization, functional group standardization, re-aromatization and atom-to-atom mapping (AAM) of reactions with keeping existing manual mapping (if some atoms were not mapped by mistake).

4.1.3.2. Atom-to-atom mapping

Second step, that will be described in this chapter, is devoted to search and correction of AAM by means of new CGR feature as linear representation of CGR and canonical SMIRKS.

CGRs could be used to identify incorrect AAM. CGR corresponding to incorrect AAM usually have more dynamic atoms and bonds than the correct one and thus is more complicated that is consistent with minimum chemical distance principle [193], Figure 29.

For every transformation we created two alphanumerical strings: canonical SMIRKS representation of reaction ignoring atom-to-atom mapping (CSMIRKS) and CGR signature (CGRS). The latter represents a unique identifier of reaction transformation like SMILES or InChI for molecules. Both line notation strings were created by in-house program that uses aforementioned Morgan-like algorithm for canonical numbering. Then canonical name is created according to defined numbering using SMILES-like rules. For human readability CSMIRKS and CGRS long text strings of variable length are compressed using MD5 hash function [194] and represented as hexadecimal number that are called CSMIRKS-key and CGRS-key correspondingly. Transformation with the same set of reagent and products will have same CSMIRKS whichever AAM is. CGRS will depend on AAM: two formally different but correct AAM

(reactions ***A*** and ***B*** on Figure 29) have the same CGRS while they do not coincide for reactions with correct and wrong AAM (reaction ***A*** and ***C*** on Figure 29). As there were no reliable and numerous data on rates of reaction involving stereoisomeric reagents or with formation of isomeric products it was decided to ignore stereochemistry of reaction and to use 2D descriptors. Thus upon CSMIRKS and CGRS generation stereochemistry of reaction was ignored and transformations with different absolute stereolabels of reagents/products were considered the same.

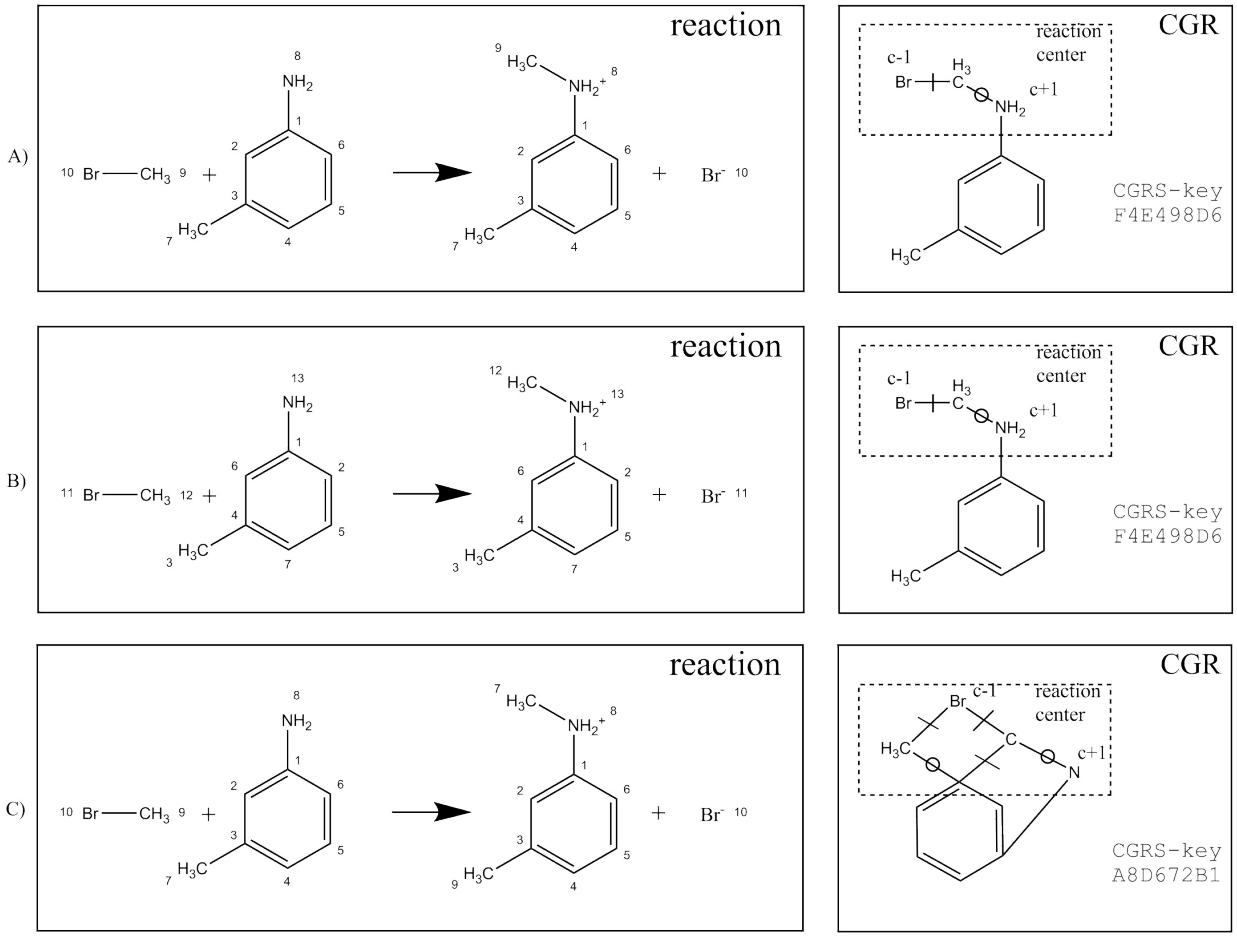


Figure 29. Conventional reaction representation (left) and CGR (right). Cleaved bonds are crossed, formed bonds are denoted by circle, dynamic atom option c+1 and c-1 mean that charge of atom increased and decreased by one respectively. Numbers near atoms represent AAM. Examples of reaction with formally different but correct AAM (***A*** and ***B***) and reaction with wrong AAM (***C***) are shown. CGRS and reaction center signatures (RCS) are shown below reactions. For human readability CGRS are compressed using MD5 hash function [194] and represented as hexadecimal number CGRS-key.

Reactions having the same CSMIRKS are grouped together. Different CGRS for reactions having the same CSMIRKS could be caused only by AAM error. Such reactions were identified and their AAMs were manually fixed.

The described approach identifies AAM errors for duplicated transformations however it does not work if the transformation is performed in one sole condition or the error was made for all transformation of given type. Thus, we used approach for checking validity of AAM by verifying signatures of reaction centers.

Reaction center is a set of atoms that change their environment in reaction. Generally, complex reactions could have more than one reaction center and thus deletion of conventional atoms and bonds will lead to disconnected graph. Thus, reaction center is a one connected component of such graph. Reaction center signatures for a given transformations were generated using approach described earlier in chapter 4.1.1.3. For a given reaction type the number of signatures is limited, e.g. for S_N2 reactions cleaned dataset it is equal to only 29. Error in AAM lead to appearance of wrong reaction center (Figure 29). They could be easily manually distinguished and AAM of corresponding reactions were fixed.

4.1.3.3. Curation of temperature data

The next task is curation of reaction conditions. The most important conditions for us is solvent and temperature. Error in solvent specification could be corrected only by human and could hardly be identified automatically. Temperature data being the most important characteristics could be automatically verified.

Since we mainly interested in reactions between dissolved molecules, solvent or mixture have to be in liquid phase. All datapoints with temperatures below freezing point or higher than boiling point of pure solvent or component in the solvent mixture were discarded as doubtful.

As a second filtering rule we required that temperature dependence should follow chemically meaningful behavior. Since the activation energy (E_a) is unknown for reaction of dataset Arrhenius equation could not be applied and Van't Hoff rule was utilized. The latter states that reaction rates usually increase 2-4 times every 10 degrees Celsius [195]. This law can be used as a filter criterion for anomaly change of rate constants with change of temperature.

As the second check, we applied Arrhenius equation to compare temperature and reaction rate constants for different reaction conditions description corresponding to the same transformation. We mark condition description of reactions as suspicious if two same transformations differ only in temperature, the difference in temperature is less than 10°C but the difference in log*k* is greater than 1 (according to Arrhenius equation it should be not more than 0.6). Suspicious conditions were manually examined. If there were examples supporting that the value is correct (similar measurements from other references), this datum was considered as valid. If there were not such examples, datum was discarded from modeling set.

4.1.3.4. Duplicate detection

This chapter is devoted to decisions that should be taken in final dataset preparation and how duplicated data are removed.

In this work measurements of rate constant corresponding to the same transformation considered as duplicates if they have the same conditions that are considered valuable for reaction rate constant: temperature, solvent and concentration of organic solvent in mixture. Thus, measurements taken from different literature source and records differed only by concentration of reactants were considered as duplicated reactions. The latter was done, because the second-order rate constant that was collected should not depend on concentration of components (there are some exceptional cases that are discussed below). Since reaction stereochemistry were neglected, reactions involving diastereoisomers and enantiomeric molecules running in same conditions were considered as duplicates as well.

Descriptions of conditions of duplicated reactions were combined and its rate constant logarithm (log*k*) were averaged if the difference in the property is lesser than 0.5 log units. Otherwise, they are marked as suspicious. The threshold 0.5 was taken from comparison of rate constant logarithm for duplicated reactions. The difference in log*k* for S_N2 reaction set (the biggest collected set) was mainly within 0.5 log units (Figure 30). We consider this value as estimation of reproducibility of reaction rate value.

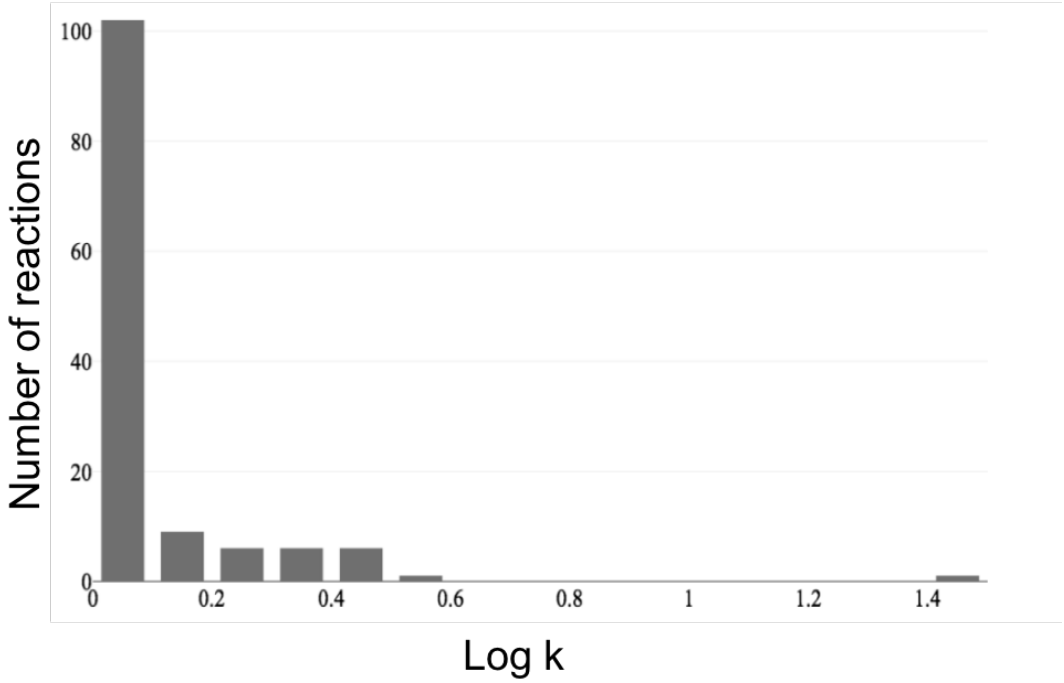


Figure 30. Difference in log*k* values for duplicated reactions.

Suspicious reactions were curated manually. If there were some evidences that the log*k* value is correct (for example, close values reaction constant of other duplicates) the value is accepted and averaged with other duplicates if any. Otherwise, values were discarded from the final modeling set. From the final dataset we excluded reactions that were performed in solvents, for which solvent descriptors described below have not been known.

The described workflow for data curation results in the curated dataset that was used for modeling.

4.1.4. Relational tables

Preprocessed data need to be stored in a database specifically designed for the reaction storage. This database will be briefly described in this section.

Relational database was found to be optimal for modeling dataset storage. It is logical, because database contain a lot of supporting tables about person that added record, time, molecules that contained in database, condition, properties, etc. Reaction with the same structure of reagents and products (we call it transformation) could be performed in different conditions. Thus, one-to-many relations between structure and

conditions is straightforward solution for data storage which is useful for modeling. The main two tables used for reaction storage are, Figure 31:

- Transformation table contain structural representation of reaction transformation, its CGRS and CSMIRKS signatures. As additional fields MD5 hashed representation [194] of CSMIRKS and CGRS were used.
- Conditions table contain information about conditions used for measurement and corresponding reaction property.

The transformation table has one-to-many relation to conditions table. Hashed CGRS and CSMIRKS representation are used for fast reaction search and finding transformation for a given reaction. The latter is needed to establish relations in database. All approaches developing for data curation procedure become essential part of the database and showed their effectiveness in storage of all data that were collected manually and have some errors. The more database is flooded with data the more efficient, useful and the same time robust automatic procedures become. Each new entity that goes into database is checked by CGRS and CSMIRKS. If there is no such CGRS in database, but CSMIRKS already exist, new reaction with high probability has mapping error that needs to be fixed.

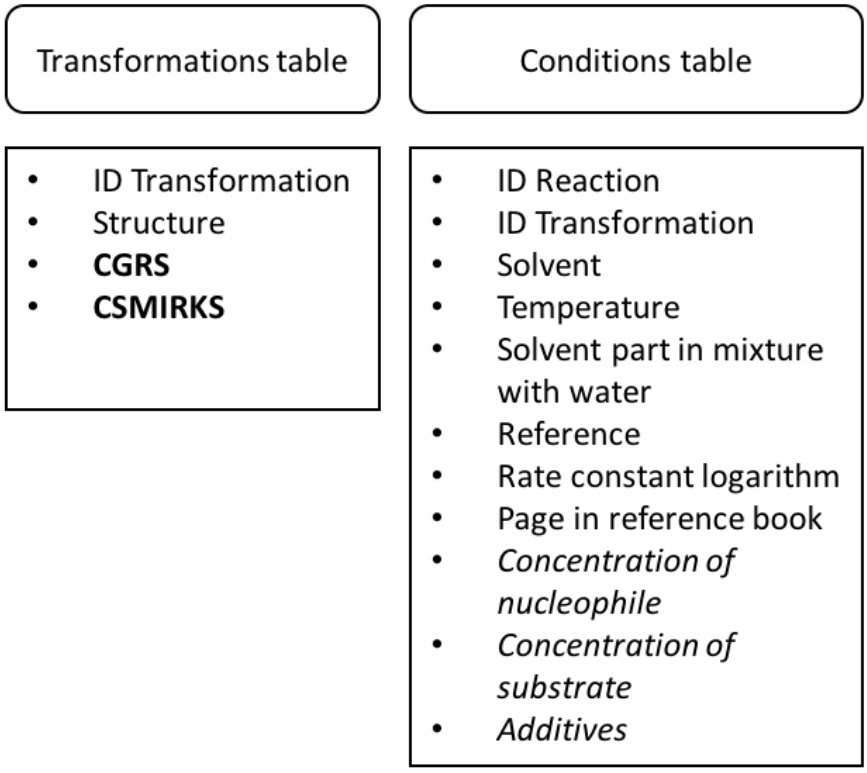


Figure 31. Structure of the reaction database. Calculated fields are shown in bold, required fields by plain text and optional fields in italics.

Thus, we created a tool that incorporates our experience in reaction data processing and allows fast and feasible reaction search and data curation.

4.1.5. Content of the database

Summary of results of the data collection and curation is given in Table 5. Data annotation was performed by several people and took about 4 years.

In total, about 10998 records of raw data of bimolecular nucleophilic substitution (S_N2), bimolecular elimination (E2), Diels-Alder (DA) reactions and tautomeric equilibria (TAU) were collected (Table 5). "Data collected" column in Table 5represents total number of reactions collected. "Curated data" shows the number of reactions after structural curation, AAM correction and deletion of full duplicates, i.e. entries where all fields coincide. The latter reason, being the major one, is caused by the fact that the dataset was collected simultaneously by several people and sometimes by error the same reaction was extracted twice. Errors in structure or absence of mandatory field were the second reasons for the deletion of datapoint. "Model set" column contains reaction selected for modeling. Data curation procedure in this case includes condition curation, duplicate identification (as described in chapter 4.1.3.4) and averaging properties of duplicates. Total number of transformation types (combinations of reagents and products) is given in field "Transformations in model set". Difference between model set size and number of transformations reflects the fact that the property for some transformation were measured in several conditions.

Table 5. Data that were collected and curated for this research.

Dataset	Data collected	Curated data	Model set	Transformations in model set	Source of data
S_N2	7848	7544	4830	1382	[1]
E2	1431	1389	1043	843	[1]
CA	1178	1130	880	679	PhD thesis defended in KFU
TAU	905	840	782	367	[1]

Every reaction/equilibria has four layers of associated information:

- Transformation – description of structural changes in reaction/ equilibria:
  - structures of reagents and products in MDL RDF format(two forms of tautomers were stored as reagent and product),
  - atom-to-atom mapping, based on reaction type reported in the source.
- Conditions – description of media and physical conditions at which reaction rate constant was measured
  - temperature (in Celsius)
  - solvent name (only organic solvent for solvent mixtures with water)
  - molar percent of organic solvent in mixture with water (100% for pure solvent). Except of DA reactions providing in pure solvents.
- Reference – the link to the source of information:
  - table in reference book [1] or thesis from which data were taken from,
  - page in reference book [1] or thesis.
- Property:
  - Reaction rate or equilibria constant.

The data were collected using InstantJChem tool [187]. Atom-to-atom mapping was done manually according to the mechanism or reaction reported at the reference book [1]. Sets of reactions kinetics data and tautomeric equilibria were stored separately,

since workflows for standardizations are different (molecules in tautomeric equilibria dataset should not be tautomerized during standardization).

The initial dataset was cleaned and converted into our own relational database format. As the result 10 903 curated reactions are stored for CA, S_N2, E2 reaction and tautomeric equilibria. The developed tools and database was shown to be extremely useful for curation and storing the data.

4.2. Modeling procedure

All QSRR models that are mentioned in this chapter have the same modeling procedure. The workflow is shown on the Figure 32.

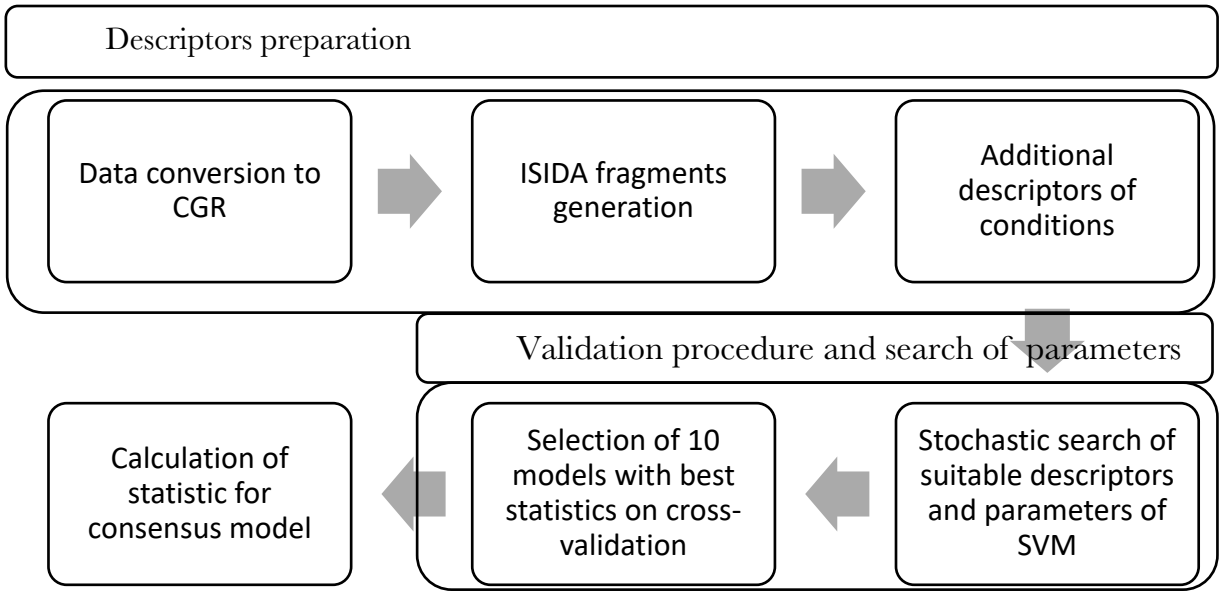


Figure 32. General modeling workflow

Modeling procedure includes several steps: descriptors preparation, selection of best type fragmentation and model parameters, model performance evaluation.

- Descriptors preparation

Initial reaction data with implicit hydrogen representation was converted to CGR using in-house CGRtools library. Resulting SDF files with CGRs were used for fragment descriptors generation by ISIDA Fragmentor program [196]. Augmented atom and sequence type descriptors were selected. Minimal and maximal length of fragments varied from 2 to 8 in all possible combinations. Descriptors sets containing all possible fragments, fragments that having at least one dynamic bond or atom or only dynamic

bonds were generated. Additional options like formal charge specification and all possible path explorations were optionally used as well. In total 616 descriptors sets were generated for each dataset. Fragment descriptors vector was concatenated with descriptors of solvent, temperature, percentage of solvent in mixture with water. All resulted descriptors sets were used for model building.

- Selection of best model parameters

Evaluation of each model was made within standard 5 fold cross-validation procedure that was repeated 10 times. Repetitions were made to exclude influence of random fluctuations of dataset composition on estimation of model performance. It leads to increased time for model preparation that is why checking all sets of fragment descriptors with all possible hyper parameters for SVM is impractically long. To avoid this time consuming calculation, genetic algorithm [162] was used for the best model selection. 10 best models returned after certain number of epochs that differ in kernel or/and descriptor set were selected for consensus model. The training set datasets in cross validation were used in consensus. Thus, consensus includes 500 individual models: 10 best combinations of fragment descriptor type with SVM kernel times 50 training sets used in cross validation (5 folds times 10 repetitions).

- Model performance evaluation

Test set predictions of individual models in consensus within cross-validation procedure were averaged and used for statistics calculations. The performance of models was defined by standard evaluation metrics R^2 and RMSE. Outliers are defined as points that have error in prediction that is more than 3*RMSE of model.

Chapter 5.

Models for rate constants of bimolecular nucleophilic substitution reactions

This chapter is devoted to the modeling of the logarithm of reaction rate (log*k*) of bimolecular nucleophilic substitution reaction. Bimolecular nucleophilic substitution reaction is very common and experimentally well-studied type of reaction. 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 [197] (see Figure 33.). 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. It is important to note that synchronicity is essential feature of S_N2 reactions. Nucleophilic substitution reactions that proceed monomolecularly through formation of carbocation with following ion recombination are usually denoted as S_N1.

Nucleophiles could be either neutral (usually amine or alcohol) or negatively charged species (alcoholates, thiolates, halogen or other inorganic salt anions, neutral amines). Usually, only reactions with aliphatic carbon in reaction center are called as S_N2 reactions (see example on Figure 33.). Reactions that involve substitution at aromatic or unsaturated carbon of substrate atom are usually asynchronous and proceed through addition-elimination (S_NAr) or elimination-addition (S_N1 or benzyne) mechanism and thus are not called S_N2. One should notice that S_N1 (two-stage mechanism with monomolecular kinetic equation) and S_N2 (one-stage, synchronous with bimolecular kinetic equation) mechanism are indeed two extremes that rarely take place in pure state. Usually mechanism of reaction is rather complex, having features of both extreme reaction types [198].

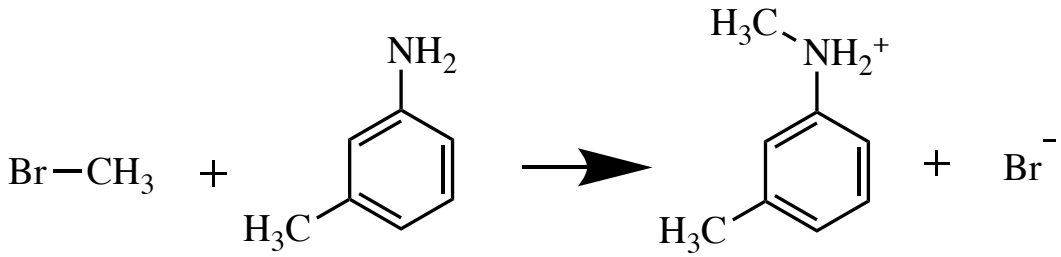


Figure 33. Example of S_N2 reaction.

Several attempts were made to build models for S_N2 reaction rate constant using QSRR approach [2, 96, 139, 141], see Chapter 2.4.3.1. However none of them could

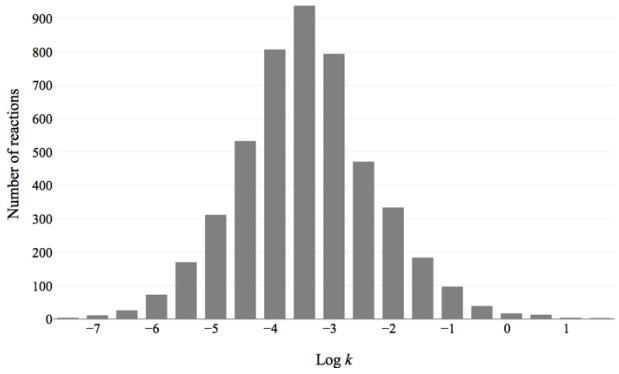
predict reactions in different solvents including water-organic solvent mixture and with different types of reagents.

In this chapter we report the information about the dataset containing some 8000 S_N2 reactions proceeding in 43 different solvents and water-organic mixtures at different temperatures. The data analysis and visualization were performed with the help of Generative Topographic Mapping [199]. For the first time, the Matched Molecular Pairs approach [163] was applied to the analysis of substituent effect. A new technique of for unbiased validation of the structure-reactivity model was suggested. Finally, results of external validation of the proposed model will be discussed.

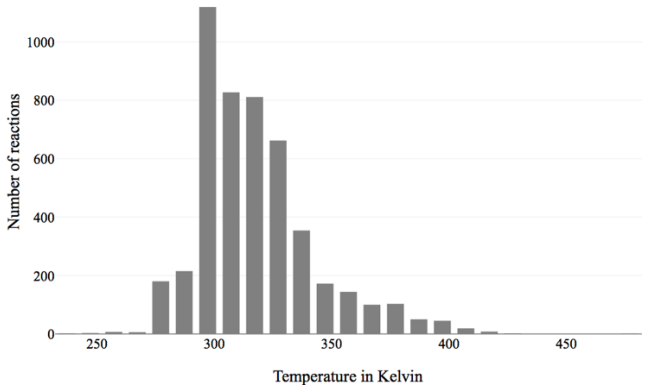
5.1. Data description

Dataset was curated by means of strategy described in section 4.1.3. The curated modeling set contains 4830 log*k* data points for 1382 transformations with the logarithm of rate constant varying from -7.68 to 1.65 in 43 different solvents and their mixtures with water. Distribution of log*k*, temperature and solvent is shown on Figure 34. One could notice that log*k* distribution is almost a perfect Gaussian curve whilst temperature distribution is highly skewed with expected cliff at 25°C. The most popular solvents were ethanol, methanol, acetone often used in mixtures with water and nitrobenzene. Most rate constants were measured at several conditions, only 551 reactions have only one reported condition; One transformation was measured at more than 100 conditions, however for vast majority of reactions less than 10 measurements of rate constant were reported. The data set contained 2882 reactions involving neutral nucleophile and 1948 reactions involving anionic nucleophile.

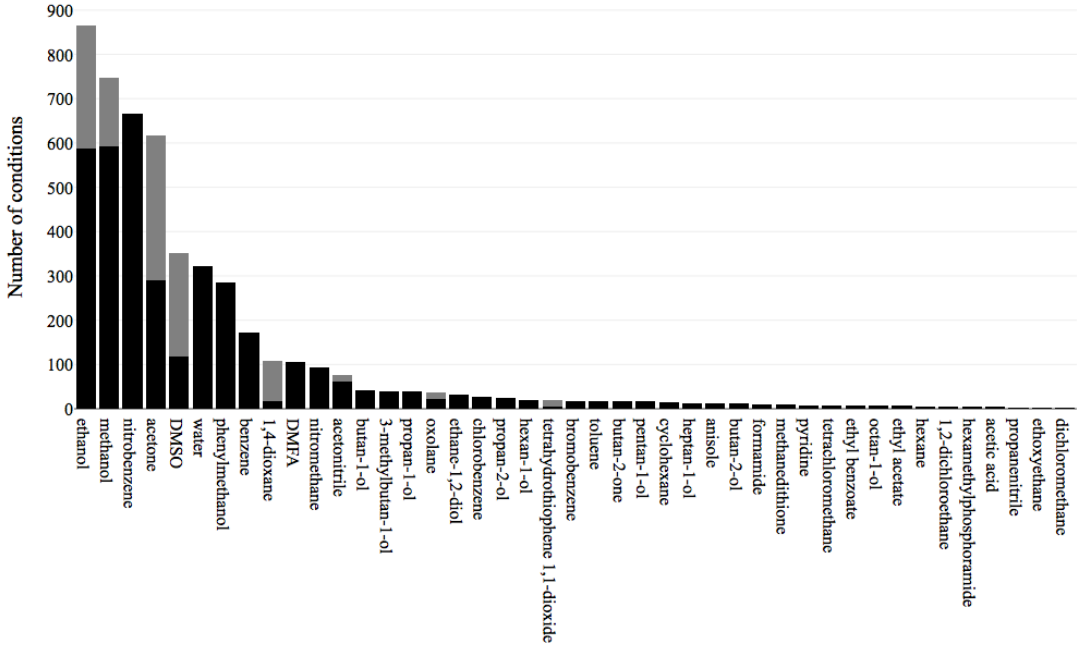
A)


Number of reactions
900
800
700
600
500
400
300
200
100
0
−7
−6
−5
−4
−3
−2
−1
0
1
Log k

B)


Number of reactions
1000
800
600
400
200
0
250
300
350
400
450
Temperature in Kelvin

C)


Number of conditions
900
800
700
600
500
400
300
200
100
0
ethanol
methanol
nitrobenzene
acetone
DMSO
water
phenylmethanol
benzene
1,4-dioxane
DMFA
nitromethane
acetonitrile
butan-1-ol
3-methylbutan-1-ol
propan-1-ol
oxolane
ethane-1,2-diol
chlorobenzene
propan-2-ol
hexan-1-ol
tetrahydrothiophene 1,1-dioxide
bromobenzene
toluene
butan-2-one
pentan-1-ol
cyclohexane
heptan-1-ol
anisole
butan-2-ol
formamide
methanedithione
pyridine
tetrachloromethane
ethyl benzoate
octan-1-ol
ethyl acetate
hexane
1,2-dichloroethane
hexamethylphosphoramide
acetic acid
propanenitrile
ethoxyethane
dichloromethane

D)

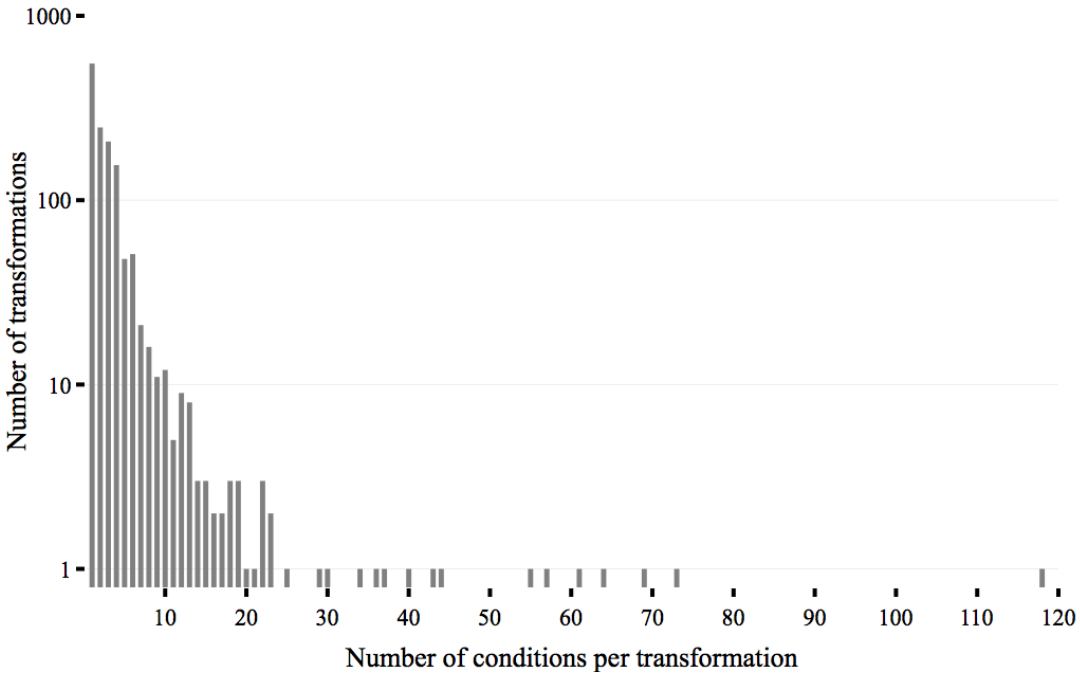


Figure 34. Data distribution with respect to (A) rate constant, (B) temperature, (C) solvent (dark part is for pure solvent one and light part is for mixture with water), (D) experimental conditions per transformation.

5.2. Data visualization and analysis

Visualization of chemical space is a powerful tool for data analysis of dataset. However, for chemical reactions this technique was extremely rarely been used since the work of Gasteiger who used Kohonen self-organized maps (SOM) for reaction classification [200]. Visualization by SOM dataset of chemical reactions of different types was used also for selection of reaction signatures [104].

Generative Topographic Mapping (GTM) [201] implemented in our laboratory [202] have shown its superiority over other visualization methods [199] and have never been used for the analysis of reaction space. In this method, the data points originally located in *D*-dimensional space (where *D* is equal to number of descriptors) are projected on 2D latent space (called manifold). The main difference from the other visualization method is that objects are projected probabilistically (with different probabilities, called responsibilities) to grid nodes of manifold, thus every projected object is indeed represented by a distribution on the map. In this case, a position of an object on the map are calculated as gravity centers of its probability distributions. Thus, although manifold is represented by a set of the nodes (grid of points, like in SOM), the positions of object on the GTM map is continuous and not tight to the node positions contrary to SOM.

The power of maps in data analysis relies on the possibility to color objects according to different criteria. Here, we colored data points according to reaction signatures, substrate, and nucleophile nature (see Figure 35). GTM was built using 25x25 grid, with other parameters set to defaults (12 RBF with width 2.8 and regularization set to1) that according to our experience is the best choice. Since we wanted to analyze only structural diversity in reaction space, condition descriptors was omitted, sequences of length from 2 to 4 containing at least one dynamic atom or bond were used as structural descriptors.

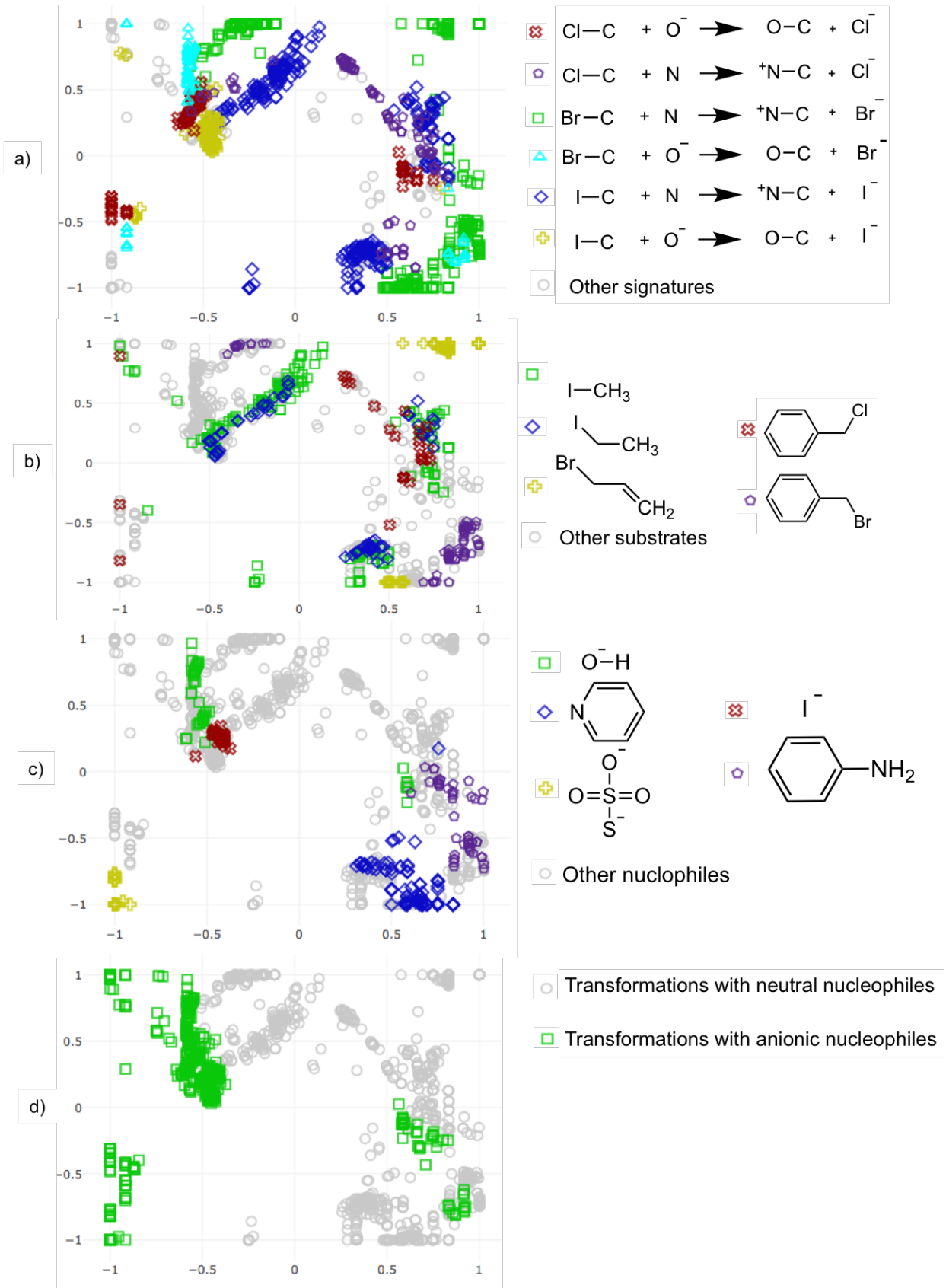


Figure 35. GTM map on 1394 transformations encoded by ISIDA fragments. Objects are colored according to a) reaction center signature (only reaction center atoms included), b) substrates, c) nucleophile structure, d) nucleophile type. The most popular signatures or molecules are explicitly shown.

One can see on Figure 35 that GTM map quite effectively separates reactions, substrates and nucleophiles chemotypes and nucleophiles classes (anionic and neutral).

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

MMP is well known and widely used approach in medicinal chemistry [163]. MMP is defined for a pair of molecules, which are different with a respect of a single group. The extension of MMPs to chemical reactions encoded by CGRs is straightforward since CGR represents a molecular graph with additional atom and bond labels. 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 36. It allows to understand how a variation in structure of reactants influences a speed or other property of chemical reaction.

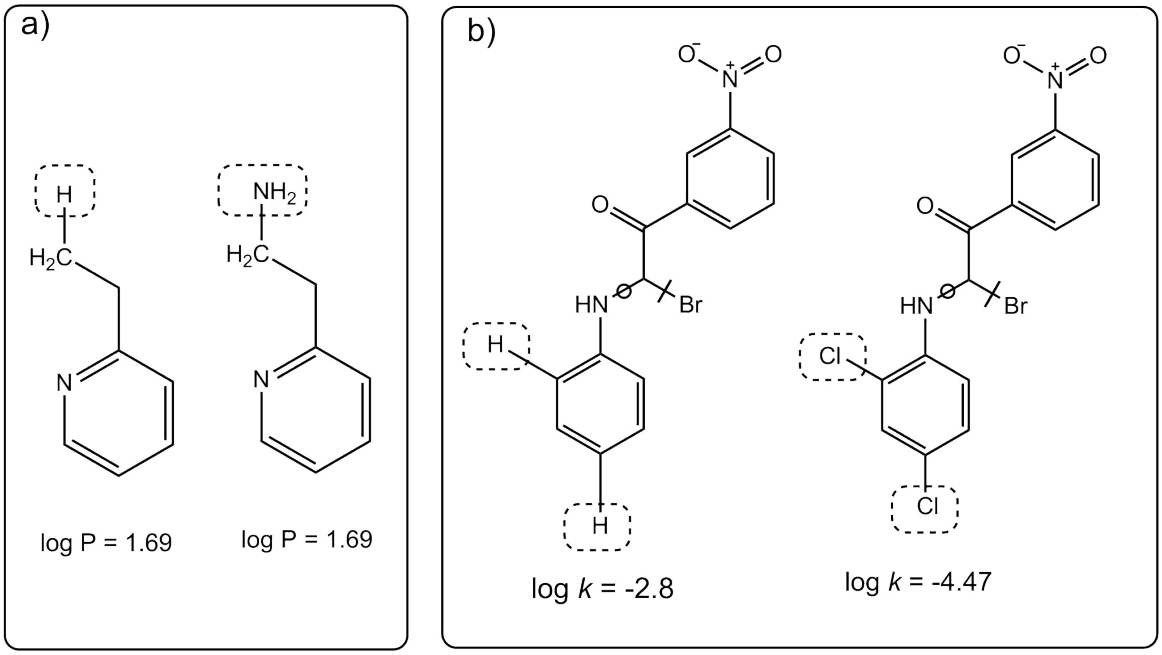


Figure 36. MMP built with the OCHEM software for molecules (left) and reactions encoded by CGR (right).

It should be noted that the reaction rate depends not only on reagents structure but also on experimental conditions. It is clear that MRP reflects only structural factor. Therefore, only reactions in almost same conditions could be analyzed with MMR. Here,

reactions in pure methanol running under ambient conditions (25-35°C) have been selected for this analysis.

For S_N2 reaction, substituents effects could be interpreted 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.

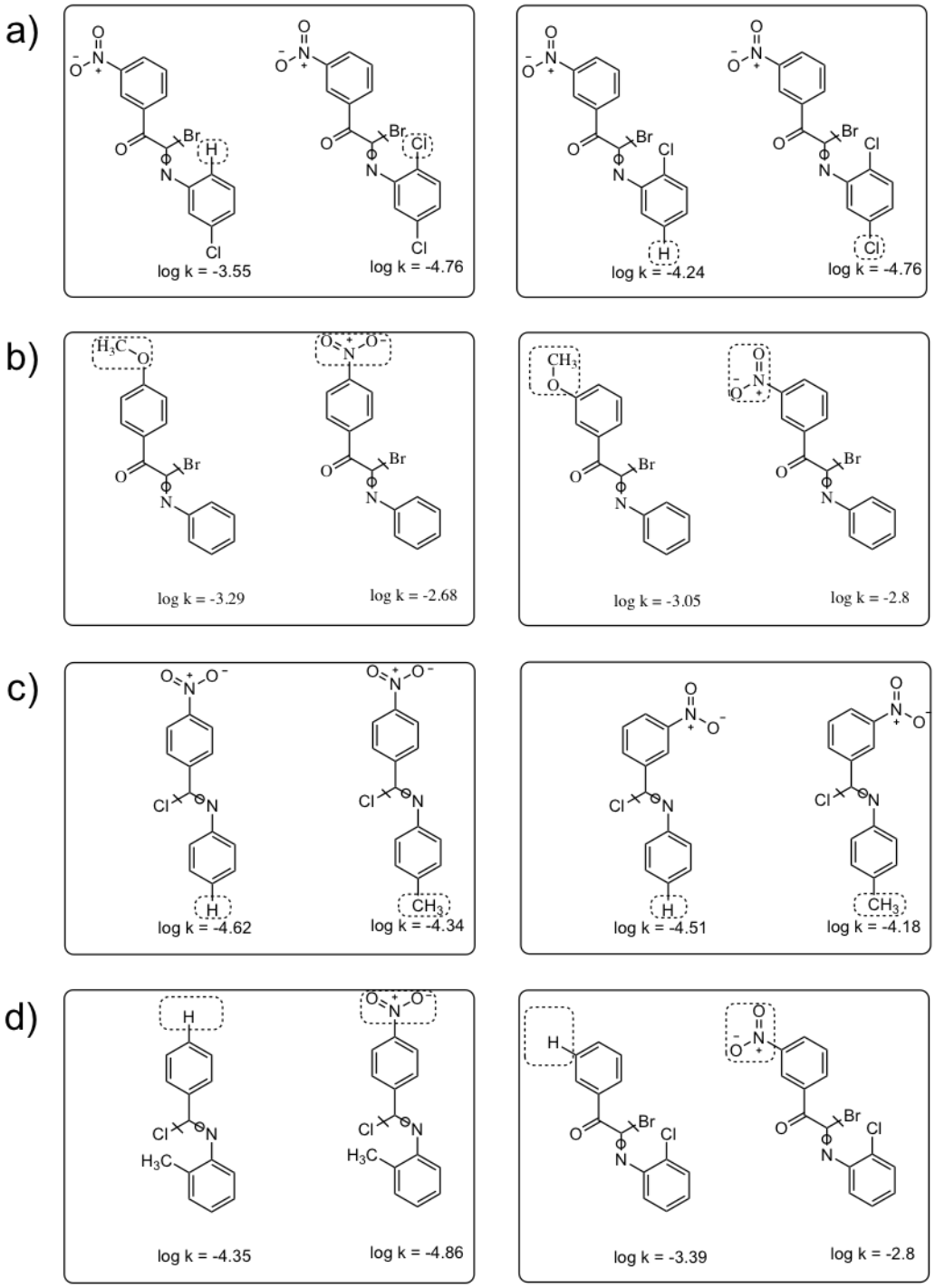


Figure 37. Examples of Molecular Reaction Pairs for S_N2 reactions in methanol at ambient temperature. In CGRs, formed and broken bond are shown as crossed and circled respectively. a) Reaction rate decreases due to substitution of hydrogen by chloride groups in nucleophile. b) Reaction rate constant increases due to replacement of metoxy to nitro group in substrate. c) Replacement of hydrogen to methyl group leads to small changes in log*k* depending on the position of the group in aromatic ring of amino-aromatic nucleophile. d) Substitution by nitro-group in substrate leading to decrease (left) or increase (right) of the rate constant.

Figure 37 provides several examples of MRP that fully supports the known mechanism and effect of substituents. As expected, substitution of hydrogen in

nucleophile by electron acceptor chlorine atom slows down reaction (Figure 37a), replacement of electron donating methoxy- to acceptor nitro-group in substrate molecule increases its speed (Figure 37b). However, some replacements, e.g., hydrogen to methyl group (Figure 37c) lead to small alteration of rate constant in different directions – the sign could be either negative or positive depending on position (ortho- or para-). Considering that reaction rate is measured with error of almost 0.5 log units (section 4.1.3.4.), these small log*k* variations could be attributed to data noise.

MRP could also be very helpful for analysis of the data quality. If change of substituent leads to different in sign changes of reaction rate constant that are greater than experimental errors as shown on Figure 37d it could be an indication of plausible error in data. For example, strange trend in log*k* variation for the MRP describing substitution of hydrogen by nitro-group in substrate (Figure 37d) was observed. One can see that in this case reaction rate constant could either increase or decrease. Indeed, right reaction pair on Figure 37d shows acceleration due to nitro-substituent in substrate molecule and this fact is fully in line with theoretical concepts of substituent effect for S_N2 reactions. Moreover, these two reactions have halo-acetophenone substrate for which S_N2 reaction mechanism is virtually the most probable. However, in 8 reaction pairs, one of which is shown in left part of Figure 37d, substitution of hydrogen by nitro-group leads to decrease in reaction rate constant. It is common for S_N1 reactions which form carbo-cationic intermediate that is destabilized by electron-withdrawing substituents. However 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 down reaction. The primary sources from which these reactions were taken from were carefully examined and we came to conclusion that almost all measurements correspond to S_N2 mechanism. However one paper [203] was very old (1925), that time mechanism of nucleophilic substitution reactions were unknown and it seems that unimolecular reaction have been considered bimolecular and thus data were incorrect (strong dependence of rate constant on reagent concentration was found for neutral nucleophile that should not take place if reaction rate constant of S_N2 reaction correctly determined). These data points were excluded from the dataset.

5.4. Model building and validation

The collected reactions were encoded by CGRs for which ISIDA fragment descriptors were generated. Each solvent was represented by 14 physico-chemical parameters accounted for polarity, polarizability, H-acidity and H-donor ability as well as molar percent of organic solvent in water to model solvent mixtures (it equal to 100% if solvent is pure). Inverse temperature was also used as descriptor. Variety of fragment descriptors of different size and topology were generated, then concatenated with conditions descriptors and use in the building of Support Vector Regression models. Optimal SVR hyperparameters and the best fragmentation schemes were selected by genetic algorithm. Ten best fragmentations that allow creation of models with highest predictive performance were selected. For each of them 10 repetitions of 5-fold cross validation have been performed, then all models were saved and used in consensus predictions. The performance of consensus model in cross-validation is very competitive RMSE=0.34 log*k* units and R^2=0.92. Plot of predicted vs experimental values is given in Figure 38.

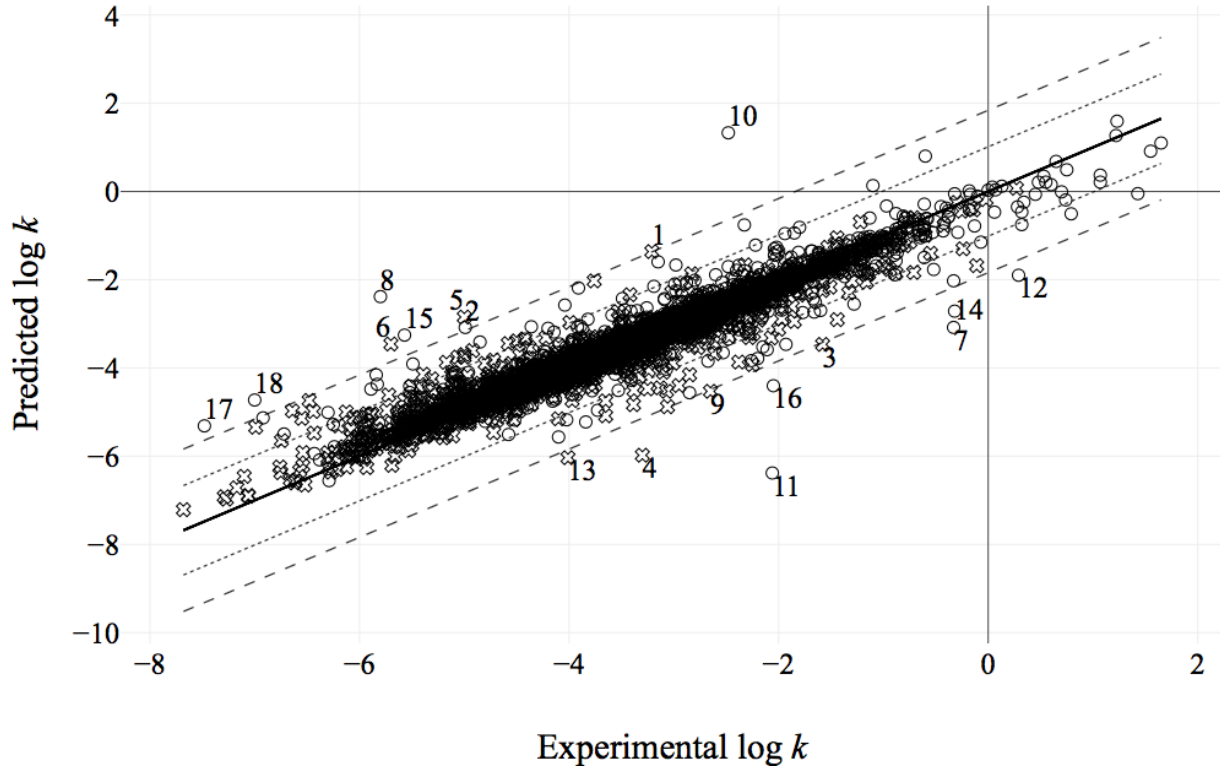


Figure 38. Predicted (with global model) vs experimental log*k* values. Solid line correspond to perfect predictions, doted lines specify margin with values predicted within

3·RMSE, dashed lines specify margin with values predicted within $3 \cdot RMSE_{UDP}$ (see below), crosses and circles represent reactions with neutral and anionic nucleophile reactions correspondingly. Numbers correspond to outliers, that have prediction error > $3RMSE_{UDP}$.

Analysis of outliers shows that 105 data points are predicted with error more than 3·RMSE. Among those 53 data points correspond to reactions involving anionic nucleophile and 52 to neutral nucleophile.

Further inspection of cross-validation procedure showed that model performance estimation is too optimistic. This situation arises from "naïve" cross validation procedure. The problem is very straightforward: if two data points correspond to the same reaction proceeding under slightly different conditions, the difference in log*k* value is small. If one of these reactions is selected to test set the other to training set, the object from test set will be predicted very close to the true value. Hence more similar conditions per transformation are reported, more chance to observe too optimistic estimation of the model performance in cross-validation.

In order to avoid this problem we decided to assess model performance only on data points which were measured under one condition only (we called them unique data points, UDP). As an unbiased estimation of predictive performance one could use prediction of UDP in cross-validation. In this case overestimation of predictive performance is impossible, since a given reaction in the test set can never occur in training set. Among selected 551 UDP, 202 and 349 belong to reactions with anion and neutral nucleophiles, correspondingly. Statistical performance of parameters for UDP reactions in cross-validation procedure is much closer to experimental error: $RMSE_{UDP}$=0.61 and R^2_{UDP}=0.75 (see Figure 39 for predicted vs experimental plot on UDPs).

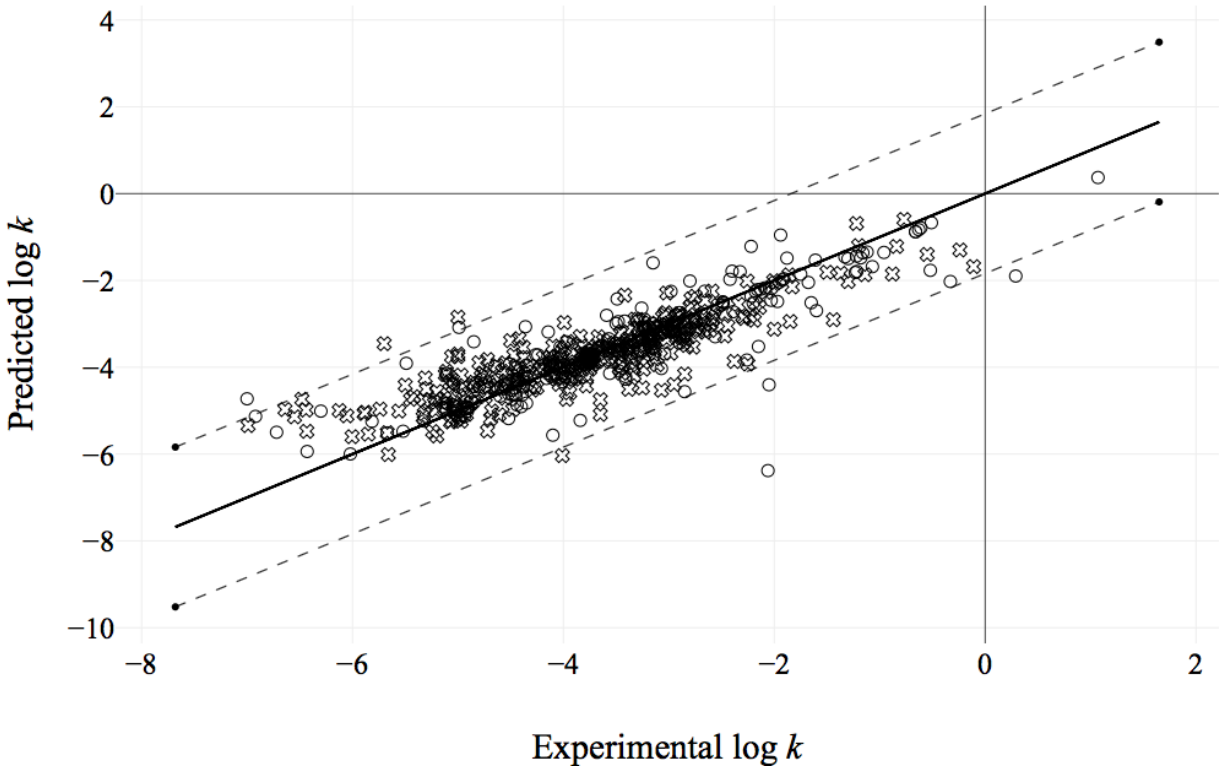


Figure 39. Predicted vs experimental values of log*k* for unique data points. Solid line corresponds to perfect predictions, dashed lines specify margin with values predicted within 3RMSE, crosses and circles are neutral and anionic nucleophile reactions respectively.

5.5. Outliers analysis

Examination of data points for which difference between predicted and experimental values exceeds $3 \cdot RMSE_{UDP}$ reveals 18 outliers, among which only 7 reactions with neutral nucleophile and 11 with anionic nucleophile.

Analysis of outliers shows that most of them result from dataset imperfectness and modeling procedure used. The errors were caused by the following reasons (reactions are drawn in Table 6).

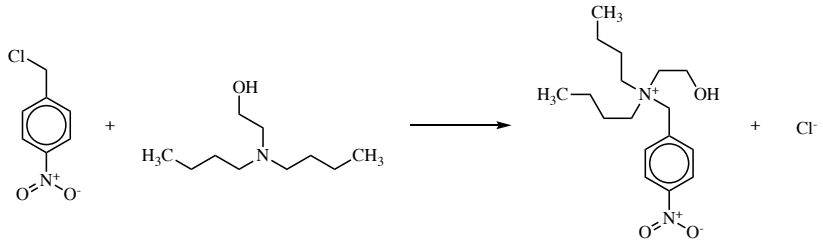
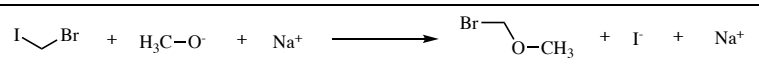
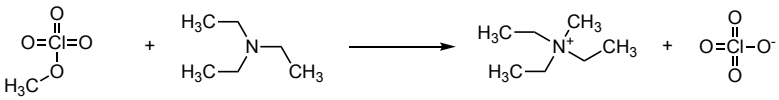
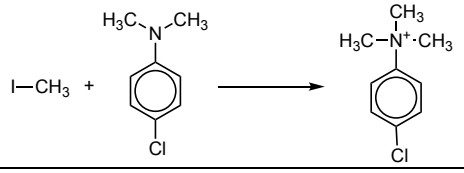
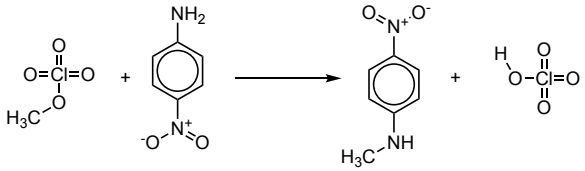
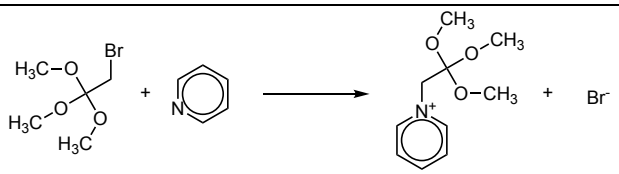
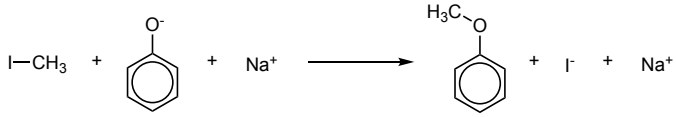
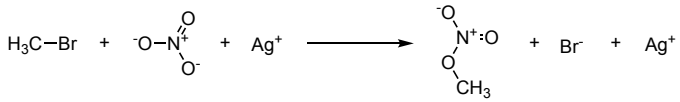
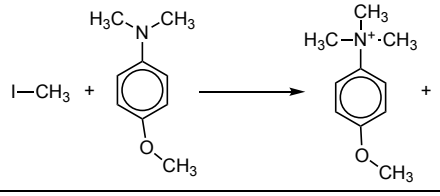
- Non-continuous dependency of the rate constant on temperature, reaction **1**. E.g. log*k* for a given temperature -4.09 (90 °C), -3.81 (100 °C), -3.68 (105 °C), -3.58 (110 °C), -3.42 (115 °C). Such a small rate constant -3.21 (201 °C) is out of trend and model logically predict it as -1.36.
- Complex structural effects that were not learned by model due to lack of representatives. Reaction **5** represents effect of direct polar conjugation for

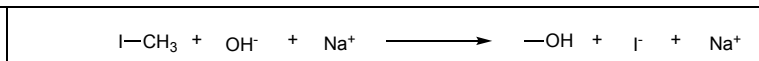
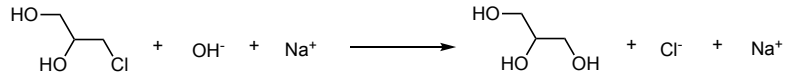
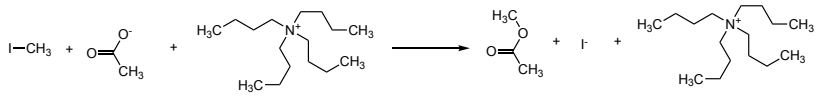
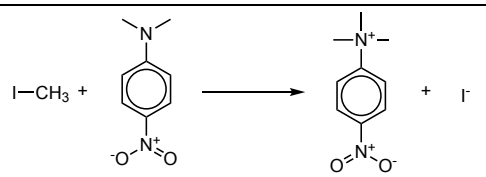
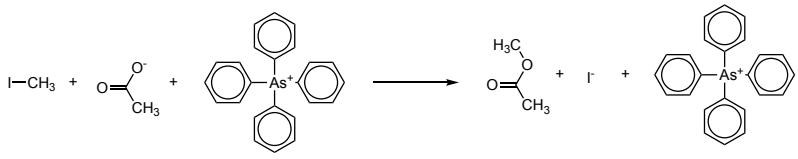
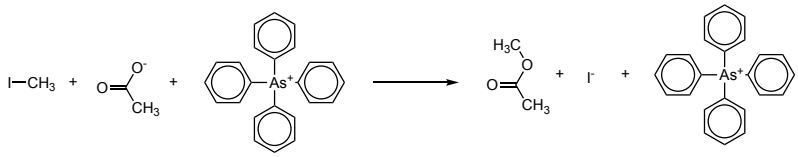
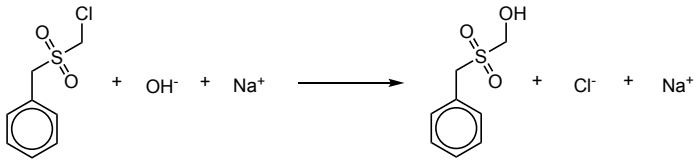
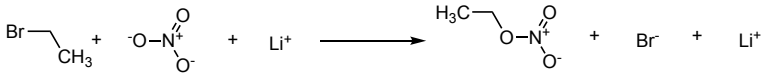
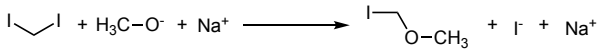
nucleophile, which strongly influences reaction rate however could not be captured by a model since it is the only reaction with para-nitroaniline.

- Transformations possessing rarely occurred structural patterns in dataset are outliers (items **16**, **11**, **6**, **3**), since the number of instances possessing given fragment in the training set is not sufficient to learn the contribution of these patterns. E.g. there was only 1 substrate possessing -SO_2CH_2Cl group (item **16**) thus it was mispredicted. Such a large rate constant for reaction 11 could be explained by anchimeric assistance, due to lack of data this effect is fully neglected by the model. Reaction **3** has rare transformation of tertiary amine to quaternary. The opposite effect could happen for small size reactants constituted of common small fragments (reactions **2**, **8**, **12**, **17**, **18**). These fragments can hardly distinguish the difference between the reactant structure.
- Solvent could have unexpected influence on the rate constant that is not captured by the model. For example, transformation **7** has 3 magnitudes higher log*k* in aprotonic polar DMFA solvent than in protonic methanol and ethanol solvents present on training set. Non-continuous dependency of the rate constant on solvent mixture composition leads to outlying prediction for reaction **10**. According to training set data rate speed increases with rise of DMSO percentage in mixture with water. For instance,. log*k* for a given percentage of DMSO in water (given in brackets) is the following: 1.07 (81% DMSO), 0.06 (65% DMSO), -1.01 (46% DMSO), -2.08 (30% DMSO), -2.93 (18% DMSO), -3.66 (8% DMSO), -4.09 (2% DMSO). Thus, such a small rate constant for pure DMSO -2.48 (100% DMSO) is out of the trend.
- Mistakes in data source. In primary source (paper [204]) rate constants for items **4**, **9**, **13** correspond to reactions of benzylamines while in reference book [1] from which data were taken reactions were annotated as phenylamines. Thus in our database wrong substrates were annotated. Alkylamine group is stronger nucleophile than amine conjugated with phenyl ring and the reaction rate constants for the reaction with the latter about 3 orders of magnitude smaller.
- Another problem revealed from the outlier analysis is based on reactions with similar descriptors but different properties. If two very similar reactions with drastically different log*k* are present in the dataset, the both can be mispredicted.

This is a case of reactions **14** and **15**, which differ only in solvent. Their rate constants differ by two powers of magnitude. When one of these reactions is selected to test set, the model predicts log*k* shifted towards reaction rate of the training set reaction.

Table 6. Experimental ("exp") and predicted ("pred") rate constant logarithms for nucleophilic substitution reactions involving anionic nucleophiles.

N	Reaction	Conditions	Exp	Pred
1		Phenyl-ethanol 100 %, 201 °C	-3.21	-1.36
2		methanol 100 %, 50 °C	-4.99	-3.08
3		methanol 100 %, 0 °C	-1.58	-3.47
4		toluene 100 %, 30 °C	-3.30	-5.98
5		methanol 100 %,0 °C	-5.00	-2.85
6		methanol 100 %,55 °C	-5.70	-3.46
7		DMFA 100 %, -20 °C	-0.33	-3.08
8		water 100 %, 25 °C	-5.80	-2.38
9		toluene 100 %, 30 °C	-2.65	-4.53

10		DMSO 100 %, 25 °C	-2.48	1.33
11		water 100 %, 0 °C	-2.06	-6.38
12		DMFA 100 %, 0 °C	0.29	-1.90
13		toluene 100 %, 30 °C	-4.01	-6.04
14		acetonitrile 100 %, 25 °C	-0.32	-2.71
15		metanol 100 %, 25 °C	-5.57	-3.26
16		1,4- dioxane 100 %, 50 °C	-2.05	-4.40
17		ethanol 100 %, 25 °C	-7.48	-5.31
18		methanol 100 %, 20 °C	-7.00	-4.73

5.6. Local and global models

Results reported previous section reveal that model's performance assessed on the reactions with neutral and anionic nucleophiles is similar. Thus, the following parameters were obtained in cross-validation: R^2=0.91 and 0.91, RMSE=0.39 and 0.3 for anionic and neutral nucleophiles, respectively. Similar situation is observed for a subset of unique datapoints: R^2_{UDP} = 0.75 and 0.74, $RMSE_{UDP}$ = 0.68 and 0.57 for anionic and neutral nucleophiles, respectively. So, it was interesting to build models for datasets, containing only particular types of reactions (here, with neutral or anionic nucleophiles), we call them local models to distinguish from global models built on the entire set. According to UDP-based validation local model performance is very close to the global model one shown above, $RMSE_{UDP}$ (anion, local model) = 0.72 and $RMSE_{UDP}$(neutral, local

model)=0.59. So, local models behave like global one and splitting of the dataset in two separate does not provide any rise in accuracy.

The question arises: whether solvent descriptors correctly capture solvent effects ? From the entire set, we selected 6 subsets of reasonable size for the reactions proceeding in particular solvents (nitrobenzene, methanol, ethanol, acetone, water, benzene). The models were built on these subsets using only fragment descriptors and unique data points. In most of cases, RMSE obtained in cross-validation for particular datasets is similar for the global and local models (Figure 40). This means that solvent descriptors are rather good to account for solvent effect in log*k* modeling . Notice that accuracy of predictions is not similar for different solvents: prediction error observed for nitrobenzene and ethanol subsets is smaller than for other solvents

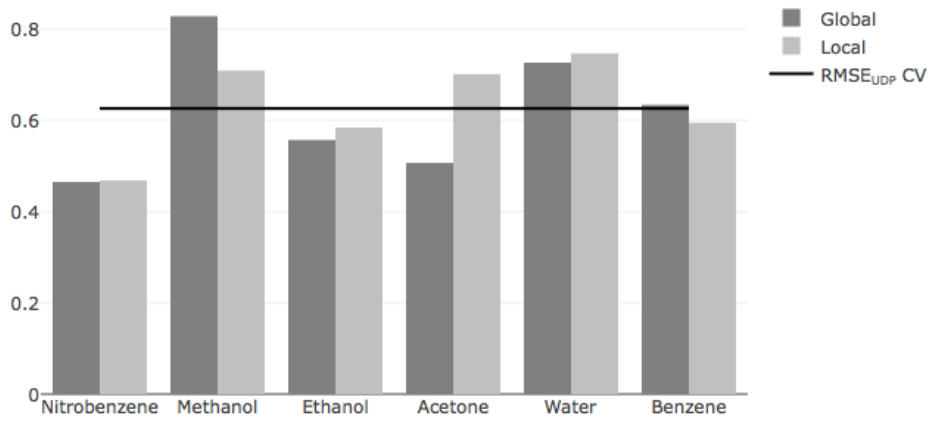


Figure 40. Cross-validated RMSE of global and local models on the subsets corresponding to particular solvents.

5.8. Validation on the external set

For validation of the model, external data set containing 104 Menshutkin reactions was collected from the papers published in 1990-2010s. Since the reference book [1] serving the data source was dated 1978, external set data don't overlap with the training set ones. Predicton performance was slightly worse than that observed in cross validation for UDP (see Figure 41) RMSE=0.8 and R^2=0.64.

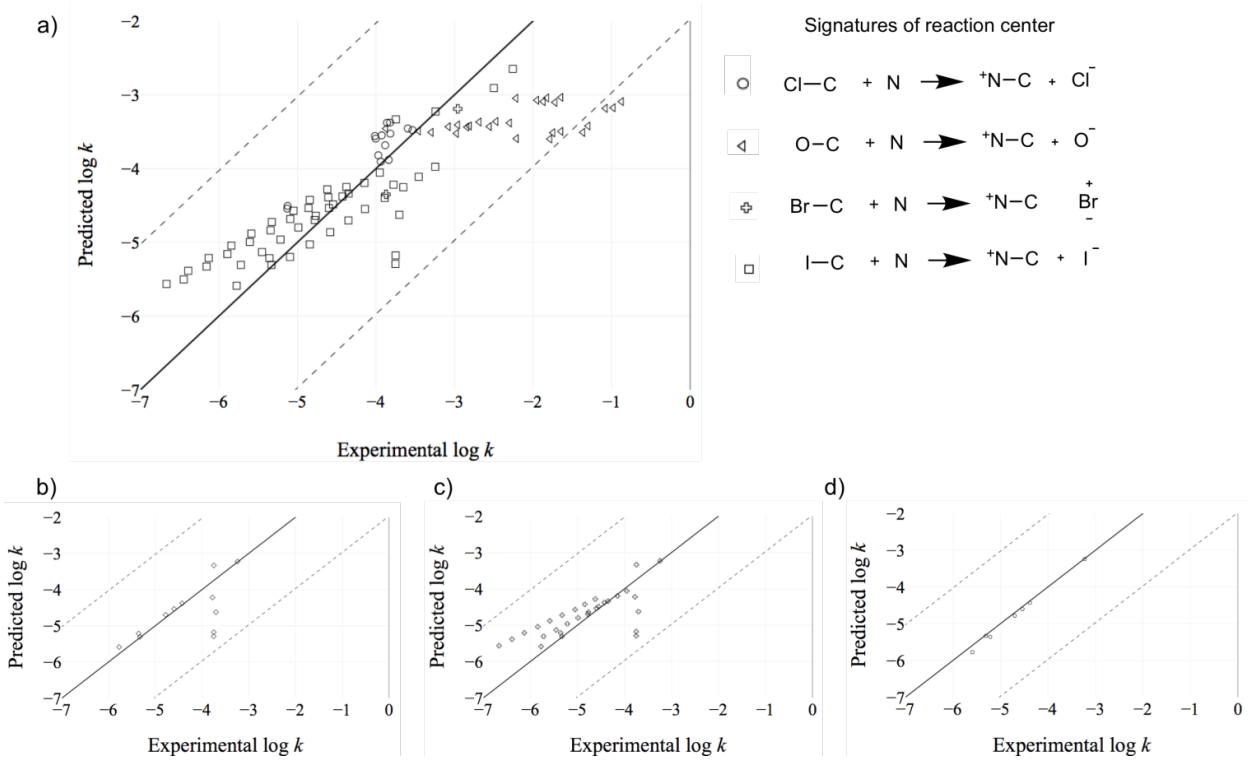
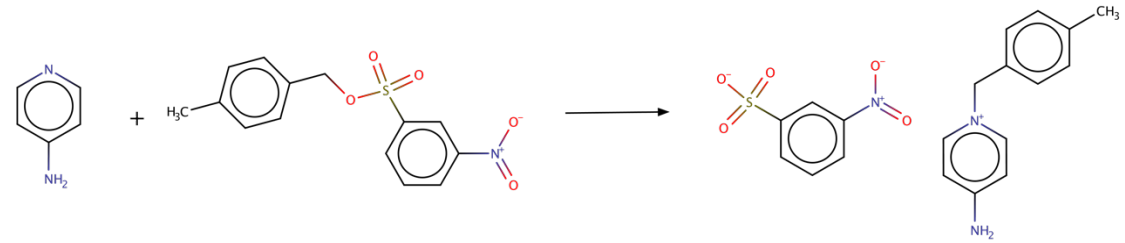


Figure 41. External set prediction by the global model. Solid line corresponds to perfect prediction. Dotted line are 3*RMSE$_{UDP}$ away from perfect prediction. (a) reactions are labelled according to signature types. (b) Only reactions having reaction signatures similar to training set ones are shown. (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 phenilsulphonate leaving group (Scheme 3). Generally reaction rate constant for them was predicted from -3 to -4 log units while experimental value (triangles in Figure 41) varied from -3.5 to -0.8. For 34 reactions of similar type occurring in the training set, log*k* never exceeded -2.5 which explains the model behavior.



Scheme 3.

We tried several applicability domain definitions but none of them was perfect. Thus, selecting reactions for which reaction center with its second environment equivalent to that in training set reactions (similar to ICClassify narrow signatures [205]), Figure 41b, retains only 13% reactions with high RMSE = 0.69 and low R^2 = 0.22. At the same time reaction center signature with first environment considers all reactions lying within AD.

The "consensus control" applicability domain [160] was also considered. According to its definition, consensus prediction is considered unreliable if a given reaction is outside of bounding box AD for a certain percentage (50% by default) of individual models. This AD was too restrictive although it efficiently discards the outliers. If we smooth the consensus control requirements and accept even one individual model considering a given reaction within its AD, this leads to RMSE = 0.61 and R^2 = 0.5 with 32.3% coverage (Figure 41c). Raising threshold to 50% leads to retaining very few similar to training set reactions for which, however, log*k* was perfectly predicted (RMSE = 0.1, R^2 = 0.98, coverage = 7.5%), Figure 41d.

Conclusions

The consensus model for the rate constant of S_N2 reaction proceeding under different reaction conditions has been built using fragment descriptors generated for Condensed Graph of Reactions and special descriptors accounting for experimental conditions. The model displays a reasonable performance both in cross-validation and on the external test set. The global model obtained on the entire set performs similarly to local models built on the subsets corresponding to particular solvents or nucleophile types. The models are available for the users on our server (cimm.kpfu.ru).

We have demonstrated that Matched Reaction Pairs approach could efficiently be applied for the analysis of substituent effect. It was found that mostly it is fully in line with theoretical concepts issued from the reaction mechanism. Detection of unusual substituent effect is in MRP analysis, could help to identify either unusual reaction mechanism or error in data annotation. Thus, both data visualization and MRP analysis could be used as tools facilitating data cleaning process.

Chapter 6.

Modeling of rate constants of bimolecular elimination (E2) reactions modeling

6.1. Models built on CGR-based reaction descriptors

Bimolecular elimination reaction (E2) is base-assisted simultaneous reaction of cleavage of bonds with electron-withdrawing group and hydrogen near single bond resulting in formation of a double bond. As for S_N2 reaction, it involves a one-step mechanism in which carbon-hydrogen and carbon-leaving group bonds break simultaneously and kinetic equation has second-order (first order with respect to substrate and base), Figure 42. Regioselectivity of reaction with asymmetric substrates follows Zaitsev's rule [206] which states that in the case of possibility of several alkene formation the one with the least number of hydrogens on double bond is formed. Moreover, this reaction is stereoselective: leaving group and proton should be located in antiperiplanar position for effective elimination. The latter explains why different diastereomers form products with opposite orientation of substituents of the double bond.

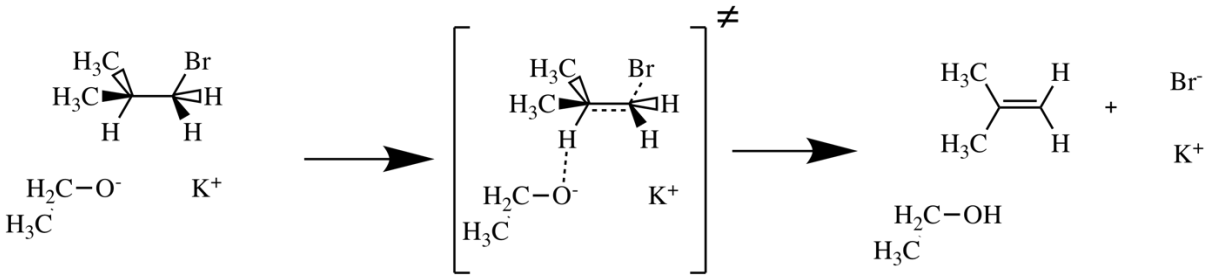


Figure 42. Mechanism of E2 reaction.

In comparison to previous modeling, E2 reaction have issues that connected with stereochemistry if reaction. Unfortunately, amount of data available for different stereoisomers was very low. Moreover, analysis of dataset shows that difference in reaction rates of products formation with opposite configuration at double bond was similar to interlaboratory experimental error. Therefore, values of reaction rate constants related to the same reaction in same condition that differ only in stereochemistry were averaged. The initial dataset of 1389 reactions after curation included averaging rate constants of reactions with stereoisomers had 1043 entries. The dataset prepared in such a way was used for modeling using standard workflow described earlier.

The results of modeling, as well as all procedures were published in the article in Russian Journal of Structural Chemistry (see below).

Journal of Structural Chemistry. Vol. 56, No. 7, pp. 1227-1234, 2015.

STRUCTURE–REACTIVITY RELATIONSHIP IN BIMOLECULAR ELIMINATION REACTIONS BASED ON THE CONDENSED GRAPH OF A REACTION

T. I. Madzhidov[1], A. V. Bodrov[2], T. R. Gimadiev[1,3], R. I. Nugmanov[1], I. S. Antipin[1], and A. A. Varnek[1,3]

UDC 544.169:544.412.2

By means of a structural representation of the chemical reactivity as a condensed graph a model predicting rate constants of the bimolecular elimination reaction is derived for the first time. The model developed enables the prediction of rate constants of reactions proceeding in different solvents or water-organic mixtures at different temperatures. It demonstrates a good predictive performance: a mean square deviation of predicted values from experimental ones is less than 0.7 logarithmic units. An outlier analysis shows that prediction errors are mainly due to the imperfection of the training data containing unique reactions. The model is available for users at arsole.u-strasbg.fr.

DOI: 10.1134/S002247661507001X

Keywords: bimolecular elimination, reaction rate constant, condensed graph of a reaction, cheminformatics, reaction descriptors, solvent descriptors.

INTRODUCTION

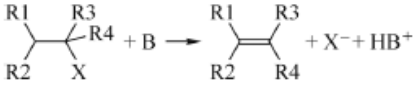
Rate reaction constants are extremely important characteristics that make it possible not only to evaluate the dynamics of chemical processes but also to calculate product yields, to estimate the selectivity of competing processes, and so on. At the same time, modern chemists use mostly phenomenological rules to qualitatively estimate the effects of certain factors on the reaction rate and selectivity.

Despite that the simulation apparatus has been sufficiently well developed in current computational chemistry, the prediction of rate constants in the condensed phase is a highly difficult problem even for one-stage processes such as bimolecular nucleophilic substitution and elimination reactions. The solution of this problem by means of rigorous quantum chemical approaches is absolutely impractical because the calculation speed is rather low and the estimation of the solvent effect greatly complicates this problem. The achievement of chemical accuracy in the prediction requires the performance of very sophisticated and resource-intensive calculations. Results of a blind competition performed within SAMPL2 Challenge [1], in which various, mainly quantum chemical, approaches competed in the prediction of solvation energies and energy differences of tautomers, have shown that the computational accuracy they reach now (the mean square error of the energy calculation was ~2.5-3 kcal/mol) is insufficient for the adequate description of the energy and is much lower than the experimental one. Almost all methods used had fitting parameters adjusted to a training set proposed in the competition. In the calculation of reaction rate constants even less accurate predictions should be expected.

[1]Kazan Federal University, Russia; Timur.Madzhidov@kpfu.ru. [2]Kazan State Medical University, Russia. [3]University of Strasburg, France. Translated from *Zhurnal Strukturnoi Khimii*, Vol. 56, No. 7, pp. 1293-1300, November-December, 2015. Original article submitted October 7, 2015.

Methods based on the use of simple correlations and constants of substituents and solvents have been relatively successfully applied to the predictions of rate constants [2]. This means that it is practically simpler to derive regularities from the available data rather than to employ deductive approaches as quantum chemical methods. Their problem is the limited applicability of models: they can mainly make predictions for one class of compounds with different substituents or for one compound in different solvents.

Recently cheminformatics methods have been applied to the solution of this problem [3]. By coding a chemical reaction in the form of a vector of numerical descriptors it is possible to search for complex nonlinear regularities in the data. The main problem of the coding consists in that a reaction is a complicated object hierarchically uniting several molecules. The condensed graph of a reaction approach [4, 5] is a very promising way to solve the problem of predicting the characteristics of chemical reactions because it enables the successful application of the standard cheminformatics methods for modeling the relationship between the structure and the reactivity. Only models providing the prediction of rate constants of nucleophilic substitution reactions have been published so far [6-10]. In this work, the rate constants of the bimolecular elimination reaction (E2 reaction) have been modeled for the first time. The general scheme of the studied elimination reactions has the following form:



where X is the leaving group (most often the halogen atom); B is the base (N-, O-, or S-containing base: anion or neutral molecule); R1-R4 are the substituents.

The central element of the approach proposed is the application of the condensed graph of a reaction. The latter is a usual pseudo-molecular graph (structural diagram) in which the forming and breaking chemical bonds are marked (Fig. 1). In order to obtain the condensed graph of a reaction it is necessary to establish the atom-to-atom mapping, i.e. to establish the relationship between the atoms of reagents and products (Fig. 1), after which the dynamic bonds are identified by superimposing the reagent and product atoms with the identical numbers.

EXPERIMENTAL

The Instant JChem program package (ChemAxon) [13] was used as a database management system for storing reactions. The structures of chemical compounds involved in the reaction were standardized using the Standardizer plugin of the JChem package [14]. The standardization procedure included: the aromatization of structures, the removal of isotopes, the standardization of nitroso groups, aromatic N-oxides, azides, nitro groups, isocyanates, sulfones, tert-N-oxides, and the removal of explicitly specified hydrogen atoms.

In order to perform the atom-to-atom mapping the Standardizer tool was also used. Errors of the atom-to-atom mapping were identified and corrected manually. Condensed graphs of the reaction were generated using the own CGR Condenser program.

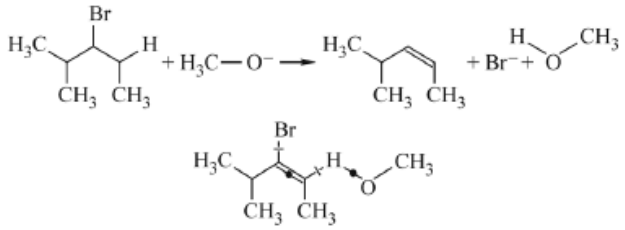


Fig. 1. Example of a condensed graph (bottom) corresponding to the elimination reaction (top). A circle denotes the forming double bond C=C and O–H; the crossed bonds denote the breaking single bonds C–H and C–Br.

ISIDA descriptors for the condensed graphs were generated using the Fragmenter program [15]. As solvent descriptors we used: SPP [16], SA [17], SB [18] Catalan constants, α [19], β [20], π* [21] Kamlet–Taft constants. The following descriptors characterizing the solvent polarity and polarizability effects were also taken: the Born function $f_B = \frac{\varepsilon - 1}{\varepsilon}$, the Kirkwood function $f_K = \frac{\varepsilon - 1}{2\varepsilon + 1}$, functions $f_1 = \frac{\varepsilon - 1}{\varepsilon + 1}$, and $f_2 = \frac{\varepsilon - 1}{\varepsilon + 2}$ (ε is the solvent permittivity), $g_1 = \frac{n^2 - 1}{n^2 + 2}$, $g_2 = \frac{n^2 - 1}{2n^2 + 1}$, $h = \frac{(n^2 - 1)(\varepsilon - 1)}{(2n^2 + 1)(2\varepsilon + 1)}$ (the refractive index n_D^{20} of the solvent was designated as n). A molar fraction of the organic solvent in the water-organic phase was added as a descriptor of mixtures. It was 100% for the pure solvent.

The descriptors and optimal hyperparameters of the SVR method were selected using the SVM Optimizer program [12].

Statistical characteristics of the quality of predictions were calculated by the following formulas:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\chi_i^{\mathrm{pred}} - \chi_i^{\mathrm{obs}}\right)^2}{N}}, \quad Q^2 = 1 - \frac{\sum_{i=1}^{N}\left(\chi_i^{\mathrm{pred}} - \chi_i^{\mathrm{obs}}\right)^2}{\sum_{i=1}^{N}\left(\chi_i^{\mathrm{obs}} - \overline{\chi_i^{\mathrm{obs}}}\right)^2},$$

where χ_i^{pred}, χ_i^{obs} are the predicted and observed (experimental) $\log k_2$ values for the i-th reaction; $\overline{\chi_i^{\mathrm{obs}}}$ is the average value of the logarithm of the rate constant; N is the number of objects in the sample.

RESULTS AND DISCUSSION

In order to derive the model, chemical transformations corresponding to E2 reactions, their conditions, and rate constant logarithms ($\log k_2$) were manually extracted from [2]. In total, 1043 reactions were extracted, out of which 97 were performed in water, 213 in water-organic mixtures, and 733 in organic solvents. The rate constants of E2 reactions range from –7.22 logarithmic units to 2.16 logarithmic units. The frequency histogram of rate constant distributions of E2 reactions is presented in Fig. 2. The almost normal character of the distribution of the predicted characteristic with the center near –3 units is worthy of attention.

As the descriptors characterizing the reaction transformation we used ISIDA fragment descriptors [4]. The value of each descriptor is equal to the occurrence frequency of this fragment in the molecule. ISIDA descriptors count the number of all possible fragments of certain topology (atomic chains of a certain length or augmented atoms – atoms with the nearest environment, Fig. 3). As options only the fragments with a dynamic bond (only the shortest paths) can be left, to include or not information on atomic types, bonds, the occurrence or changes in formal charges on atoms in the description of the fragment. With the use of all possible combinations of options 260 sets of descriptors consisting of several tens up to several thousands of descriptors responsible for the occurrence of one or another fragment was generated.

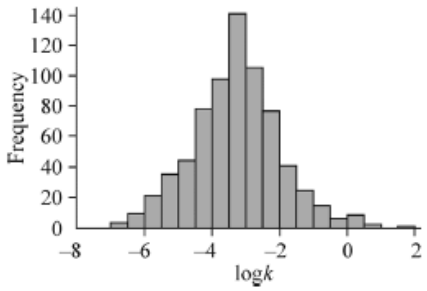


Fig. 2. Frequency histogram of the E2 reaction rate constant distribution.

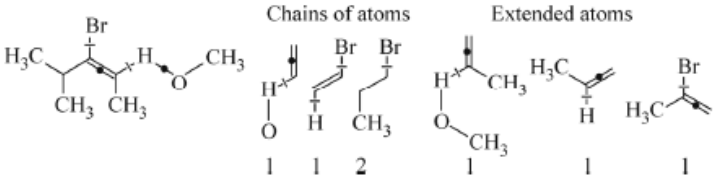


Fig. 3. Examples of ISIDA descriptors based on chains of atoms and augmented atoms, below the fragment occurrence is given, which forms a descriptor vector.

Since the reaction rate also depends on the temperature and solvent, the reciprocal temperature of the reaction course and 14 characteristics of the solvent were included in the vector of descriptors for modeling. Solvent polarity and polarizability parameters, H acceptor and donor abilities were used as such characteristics (see EXPERIMENTAL). For the description of water-organic mixtures the molar fraction of the organic solvent in the mixture was added as a descriptor.

From the obtained sets of descriptors it is necessary to choose the best ones allowing the generation of a model with the highest predictive ability. The predictive ability of the model was controlled by the fivefold cross-validation procedure. To this end, the whole sampling was divided into five parts, one of which was the test set and the other parts were used to generate the model. The obtained model was applied to the test set. The procedure was repeated for each of the five parts in turn, so each part was once the test set. In the end, the predictions were combined and the error was calculated. The quality of models was evaluated using the determination coefficient (Q^2) and the mean square deviation of the predicted values from the experimental ones (the root-mean square error, RMSE). The whole fivefold cross-validation procedure was repeated 30 times. The so-called consensus model in which all intermediate models generated in the course of the cross-validation are retained and can be used to predict $\log k_2$ for new reactions was used as the final one. Consequently, 150 predictions are obtained for each object and they are averaged which allows a decrease in the prediction error and an increase in the model stability due to the reduction of fluctuations caused by that different objects fall into the training set. The quality of the consensus model was evaluated by averaging the predictions for objects from the test set followed by the calculation of $RMSE_{cons}$ and Q^2_{cons} by the usual formulas. Since the characteristics predicted by the model in the test set are always taken into account, then $RMSE_{cons}$ and Q^2_{cons} evaluate the quality of the prediction of new data.

As a machine learning method the support vector regression (SVR) method was employed [11]. It tries to draw a flexible multidimensional tube with a certain radius in the descriptor space so that to provide the maximum occurrence of the objects inside the tube. The method has three hyperparameters (coefficient C characterizing the cost for not falling into the tube, the kernel parameter γ, and the parameter ε characterizing the tube radius) whose values are chosen so that to guarantee the maximum predictive ability of the model. Thus, there are several parameters that must be chosen in modeling: a set of descriptors and SVR hyperparameters. The optimal parameters were selected by the evolution algorithm [12] so that to provide the maximum predictive ability of the model using the fivefold cross-validation.

The highest predictive ability of the model was provided by the descriptors based on atomic chains with a length from two to six atoms with regard to the information on charges on the atoms. The mean square deviation of the predicted values from the experimental ones for the consensus model ($RMSE_{cons}$) was 0.69 logarithmic units, which is somewhat worse than that for the previously described model for S_{N^2} (about 0.5) [6, 7]. The determination coefficient (Q^2_{cons}) for the consensus model was 0.75. The correspondence between the predicted and experimental values for the consensus models is depicted in Fig. 4. From the data presented and Fig. 4 it is seen that $\lg k_2$ is predicted well. The database contains rate constants measured by different methods and in different laboratories. The occurrence of differences in the measurements of rate constants causes the appearance of noise in the data. Since the methods used extract the dependences from the experimental data, the uncertainty of the prediction because of the data noise arises in the model itself. Unfortunately, we failed to find in the database the identical reactions performed under the same conditions but in different laboratories, therefore it was impossible to analyze the interlaboratory measurement errors. However, we assume the measurement errors

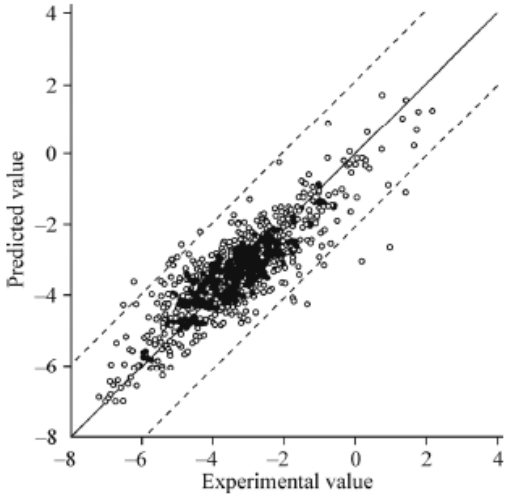


Fig. 4. Predicted E2 reaction rate constants in comparison with the experimental ones. A solid line corresponds to the ideal coincidence of the predicted and experimental values. Dashed lines correspond to the deviation of the predicted values from those observed in $3RMSE_{cons}$.

to be no less than the similar value for the S_{N^2} reaction that reached 0.5 and more logarithmic units [6]. Thus, we can consider the prediction quality of the model proposed to be comparable with the measurement accuracy of rate constants.

For the analysis of the quality of model we have sought for the objects for which there was a significant deviation of the predicted values from the observed ones. If the difference between the predicted and experimental values exceeded three mean square deviations for the consensus prediction, then it was considered outlier. In total, 14 reactions of this type were found (Fig. 5). An analysis of the outliers made it possible to reveal the following main reasons for their appearance:

- When there is an insufficient number of certain type fragments the model cannot generalize its contribution to the total reaction rate constant. Apparently, if in the reaction in the test set there was a fragment that substantially affected the constant and if there was no the reaction with this fragment in the training set, then we could not expect the qualitative prediction of the sought characteristic. Therefore the rate constant was predicted incorrectly for seven reactions (*c*, *d*, *e*, *h*, *j*, *m*, *n*). Several reasons for the appearance of unique fragments in the descriptor representation of the reactions were found. For example, in reaction *m* an incorrect atom-to-atom mapping led to the derivation of an incorrect condensed graph of thereaction, as a result of which the descriptor vector dramatically differed from that for the structurally similar reactions. Despite that the atom-to-atom mapping was checked, there was an error in the manual analysis. Yet another error in the prediction (reaction *n*) was caused by the problem in the automatic standardization of the representation of the nitro group, as a result of which the nitro group turned out to be represented uniquely. In reactions *a*, *c*, *d*, *e*, *h*, *j*, *k* there were unique fragments because of the specificity of the database itself. For instance, reaction *d* was a single example of the use of the iodide anion as a base. Reaction *h* is a unique example of diphenyltrochloromethylmethane with acceptor substituents (note that similar reaction *n* was represented incorrectly). The situation is additionally complicated by the use of a rare thiophenolate anion as a base. In reaction *j* a dichloro- substituted derivative is used whereas in all other cases, there are only trifluoro- or monochloro-substituted derivatives. Similar problems can be noted for other reactions *c* (there was deuterium) and *e* (the leaving group is nitrate).
- If the reactions with very large or very small (with respect to the training set) $\log k_2$ values fall in the test set, the predictions are potentially subject to an error due extrapolation problems. The prediction becomes especially difficult for small reagents containing a small number of molecular fragments used as descriptors. For reaction *a* there is such an error.

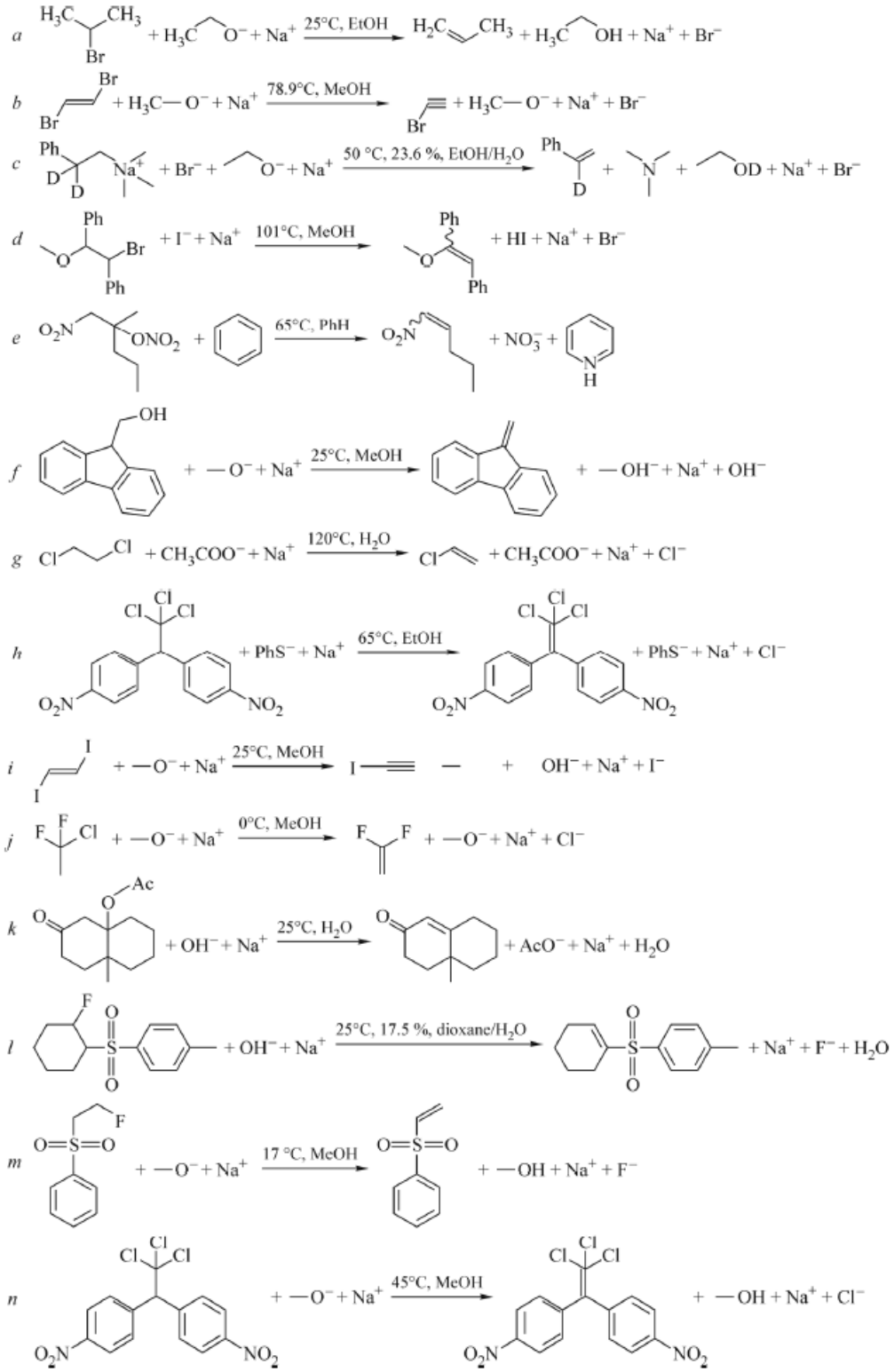


Fig. 5. Reactions for which a strong deviation of the predicted rate constants from the experimental values was obtained.

- Reaction *f* seems to be erroneously assigned by the author of the handbook to the E2 type and then extracted to the database. This dehydration reaction is unlikely to pass in one stage as the E2 reaction. All in all, there are five such reactions but since the four remained have the rates close to –3, then for them the model returns predictions close to true values. For very new objects SVR predicts the desired characteristic by average value of the property in training set, which is close to –3.
- Yet another example given by reaction *g* fell into the number of poorly predictable objects, which seems to be due to the specific reaction conditions. It appears that this reaction was conducted under autoclave conditions whereas all the other reactions in the database were performed at a normal pressure.

Imperfectness of the simulation technique used was also one of the reasons for the appearance of appreciable errors in the prediction. In the model proposed stereochemistry was not taken into account because it is extremely difficult and seems to require the development of new descriptors. Most often the stereochemistry of the reaction center insignificantly affected the rate constant; its effect was appreciably smaller than the prediction error. However, the rate constant of reactions *b*, *i*, which were dehydrohalogenation at a double bond with the *trans*-configuration, was substantially higher than that for analogues with the *cis*-configuration. Hence, the rate constant was considerably underestimated. Another problem was the effect of very small fragments. Since as a result of the selection of optimal sets of descriptors, chains with a length of two atoms and more were chosen, then if the reaction equation contains structures consisting of one heavy atom (hydrogen atoms were obviously not taken into account), information on this molecule is lost. Thus, in reactions *l* and *k* a hydroxide anion that was not taken into account in the simulation served as a base. The majority of monoatomic structures is represented by chloride and bromide ions that are close in activity, therefore the absence of features corresponding to a base in the descriptors was interpreted by the model as the presence of Br^- and Cl^-. A higher activity of the hydroxide ion caused errors. This effect is also likely to affect the appearance of reaction *d* in the outlier list.

The model is available for users in the online predictor at http://arsole.u-strasbg.fr. In the editor a user draws the reaction of interest and the server automatically creates the atom-to-atom mapping, after which the user points out the solvent and temperature of interest (Fig. 6). The program yields the predicted value of the logarithm of the reaction rate constant with the estimate of the quality of the prediction (optimal, moderate, or bad).

CONCLUSIONS

The predictive model for the reaction rate ($\log k_2$) of bimolecular elimination proceeding under different conditions has been built. The model involves fragment descriptors generated from the Condensed Graphs encoding ensemble of reactants and products of a given reaction. Unlike reported in the literature approaches, it allows one to predict $\log k_2$ for the E2 reactions proceeding as a function of temperature or reaction medium (solvent of mixture of solvents). Although predictive performance of the model is high enough, it fails to predict the rate for reactions in which substrates contain unique fragments with respect to any other compound in the training set. The model is available for users at http://arsole.u-strasbg.fr.

The authors are grateful to the Russian Scientific Foundation (contract No. 14-43-00024) for the support.

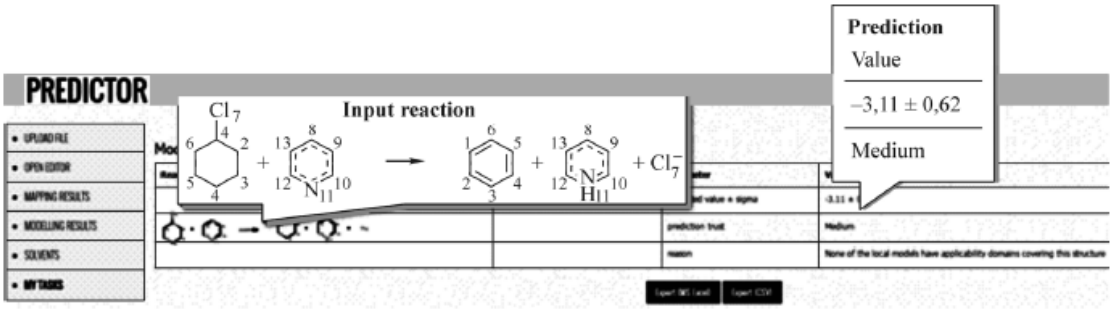


Fig. 6. Example of the output of prediction results at the web-server. Not only the predicted value but also the quality of the prediction are given as the results.

REFERENCES

1. M. Geballe, A. G. Skillman, A. Nicholls, et al., *J. Comput.-Aided Mol. Des*., **24**, No. 4, 259 (2010).
2. V. A. Pal`m, *Usp. Khim.*, **30**, No. 9, 1069 (1961).
3. A. Varnek and I. I. Baskin, *Mol. Inf.*, **30**, No. 1, 20 (2011).
4. A. Varnek, D. Fourches, F. Hoonakker, et al., *J. Comput.-Aided Mol. Des*., **19**, Nos. 9/10, 693 (2005).
5. S. Fujita, *J. Chem. Inf. Model.*, **26**, No. 4, 205 (1986).
6. T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, et al., *Russ. J. Org. Chem*., **50**, No. 4, 459 (2014).
7. R. I. Nugmanov, T. I. Madzhidov, G. R. Khaliullina, et al., *J. Struct. Chem*., **55**, No. 6, 1026 (2014).
8. F. Hoonakker, N. Lachiche, A. Varnek, et al., *Int. J. Artif. Intell. Tools*, **20**, No. 2, 253 (2011).
9. A. A. Kravtsov, P. V. Karpov, I. I. Baskin, et al., *Dokl. Chem*., **440**, No. 2, 299 (2011).
10. A. A. Kravtsov, P. V. Karpov, I. I. Baskin, et al., *Dokl. Chem*., **441**, No. 1, 314 (2011).
11. H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Support Vector Regression Machines*, in: *Advances in Neural Information Processing Systems*, M. C. Mozer, J. I. Jordan, and J. I. Petsche (eds.), vol. 9, MIT Press (1997), p. 155.
12. D. Horvath, J. Brown, G. Marcou, et al., *Challenges*, **5**, No. 2, 450 (2014).
13. *InstantJChem 15.7.27.0*, ChemAxon (2015); http://www.chemaxon.com.
14. *Standardizer, JChem 15.8.3.0*, ChemAxon (2015); http://www.chemaxon.com.
15. G. Marcou, V. Solov'ev, D. Horvath, and A. Varnek, *ISIDA Fragmentor2011-User Manual* (2012).
16. J. Catalán, V. López, P. Pérez, et al., *Liebigs Ann.*, **1995**, No. 2, 241 (1995).
17. J. Catalán and C. Díaz, *Liebigs Ann.*, **1997**, No. 9, 1941 (1997).
18. J. Catalán, C. Díaz, V. López, et al., *Liebigs Ann.*, **1996**, No. 11, 1785 (1996).
19. R. W. Taft and M. J. Kamlet, *J. Am. Chem. Soc*., **98**, No. 10, 2886 (1976).
20. M. J. Kamlet and R. W. Taft, *J. Am. Chem. Soc*., **98**, No. 2, 377 (1976).
21. M. J. Kamlet, J. L. Abboud, and R. W. Taft, *J. Am. Chem. Soc*., **99**, No. 18, 6027 (1977).

1234

Conclusive remarks

The first model predicting the rate constant of E2 reaction with different reagents and in wide variety of conditions was prepared using fragment descriptors generated for Condensed Graph of Reaction and some special descriptors accounting for experimental conditions. The predictive performance of the consensus model in cross-validation was high enough: RMSE = 0.69 log units, R^2 = 0.75. Analysis of outlier was shown that model fails to predict log*k* for reactions involving substrates with unique fragments . The model was published online. The results supported the universality of the approach.

6.2. Models built on mixture-based reaction descriptors

The question rises how well the CGR-based descriptors perform in comparison with other known descriptor types for reaction. Thus benchmarking study of different descriptor types was made. CGR-based fragment descriptors were compared with several difference fingerprints calculated by RDKit and proposed in the work [100]. Moreover the novel type of descriptors suitable for reaction modeling was introduced. They are based on the SiRMS descriptors for compounds mixture [18]. Formally left- and right-hand side of reaction could be considered as mixture of reactants and products respectively and corresponding SiRMS mixture descriptors for them could be calculated. Resulting descriptor vector could be combined by concatenation or subtraction. Hence three completely orthogonal descriptor generation strategies were compared: difference fingerprints, CGR-based fragment descriptors and mixture descriptors for chemical reaction.

Random Forest [153] (RF) machine learning algorithm for model building. For the study we needed non-linear regressor efficiently working with very large descriptor space (when number of descriptors is substantially larger than number of reactions to model), having as few hyperparameters as possible to adjust. RF perfectly suited our need due to efficient tackling non-linearities, non-sensitivity to descriptor vector size, and only one hyperparameter that is required to be adjusted – ratio of descriptors that is randomly selected for tree branching (number of trees in the forest should be as large as possible, we used 500, and the other parameters of RF influence poorly and default values could be accepted).

E2 reactions were used in the benchmarking study. This dataset is not as large as S_N2 one and reliable results could be obtained in reasonable time. Some of descriptor spaces (for example, for based on SiRMS descriptors of mixtures) were extremely large and descriptor storage was also an important issue. Moreover, it was decided to reduce number of reactions in data set by careful manual examination of the E2 dataset and exclude all doubtful data, data where stereoisomery was important, data containing structural errors. This was done to ensure that the descriptors indeed reflect the relevant information and outliers due to experimental errors does not influence the performance metrics. Additionally, fragment control for assessment of model applicability domain was used.

And finally for unbiased estimation of model performance stratified cross-validation was used. It includes the following procedure: reactions were united by products formed, and in cross-validation for test sets reactions whose products do not coincide with the ones in training set were selected. Monte-Carlo algorithm was used for controlling that the number of reaction constituting test sets in 5-fold cross-validation was about 20% of whole dataset. The paper describing modeling procedure and results of benchmarking was published in Journal of Computer-Aided Molecular Design and attached below.

J Comput Aided Mol Des (2017) 31:829–839
DOI 10.1007/s10822-017-0044-3

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

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

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

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 ($\mathbf{B^-}$) 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,

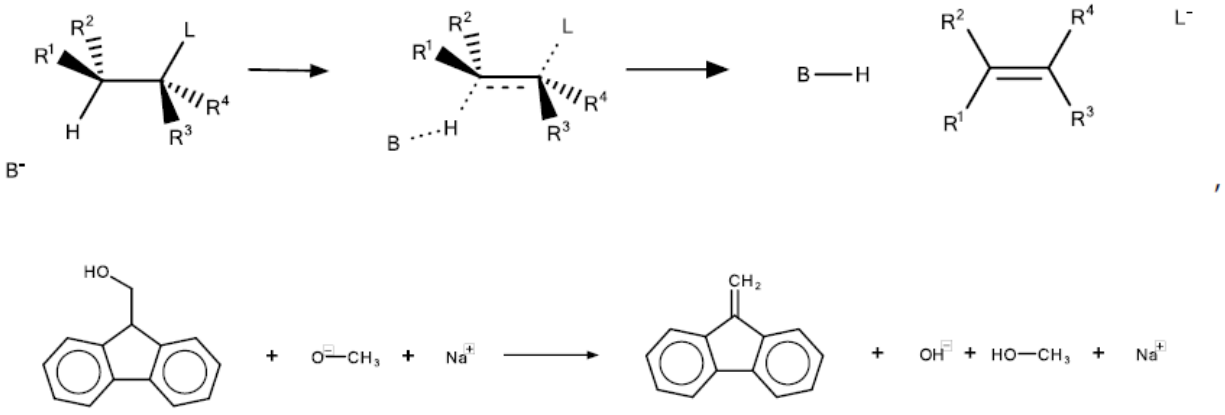
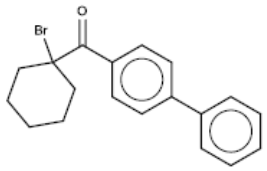
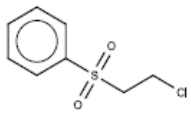
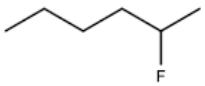
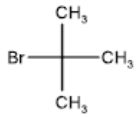
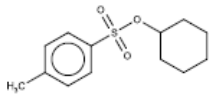
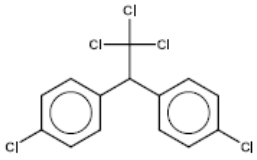
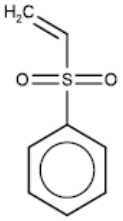
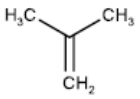
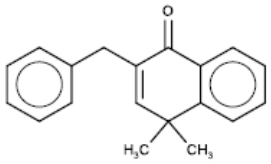
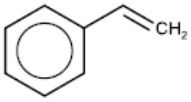
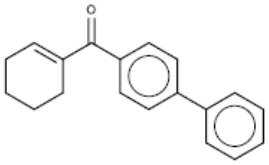
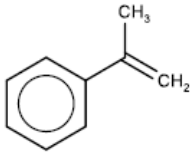
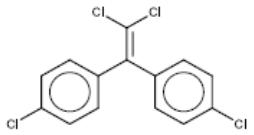


Fig. 1 A bimolecular elimination reaction. (*top*) Schematic representation of the E2 reaction mechanism, where B^- is a base and L is a leaving group. (*bottom*) An example of an E2 transformation of (9*H*-fluoren-9-yl)methanol into 9-methylene-9*H*-fluoren, where CH_3O^- (from sodium methylate) is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

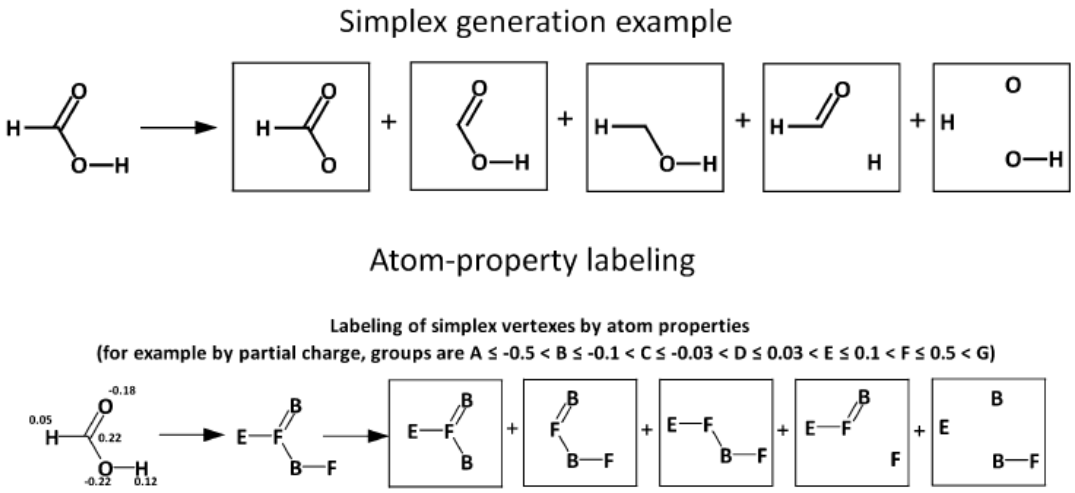


Fig. 2 An example of simplex descriptor generation for individual compounds

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.

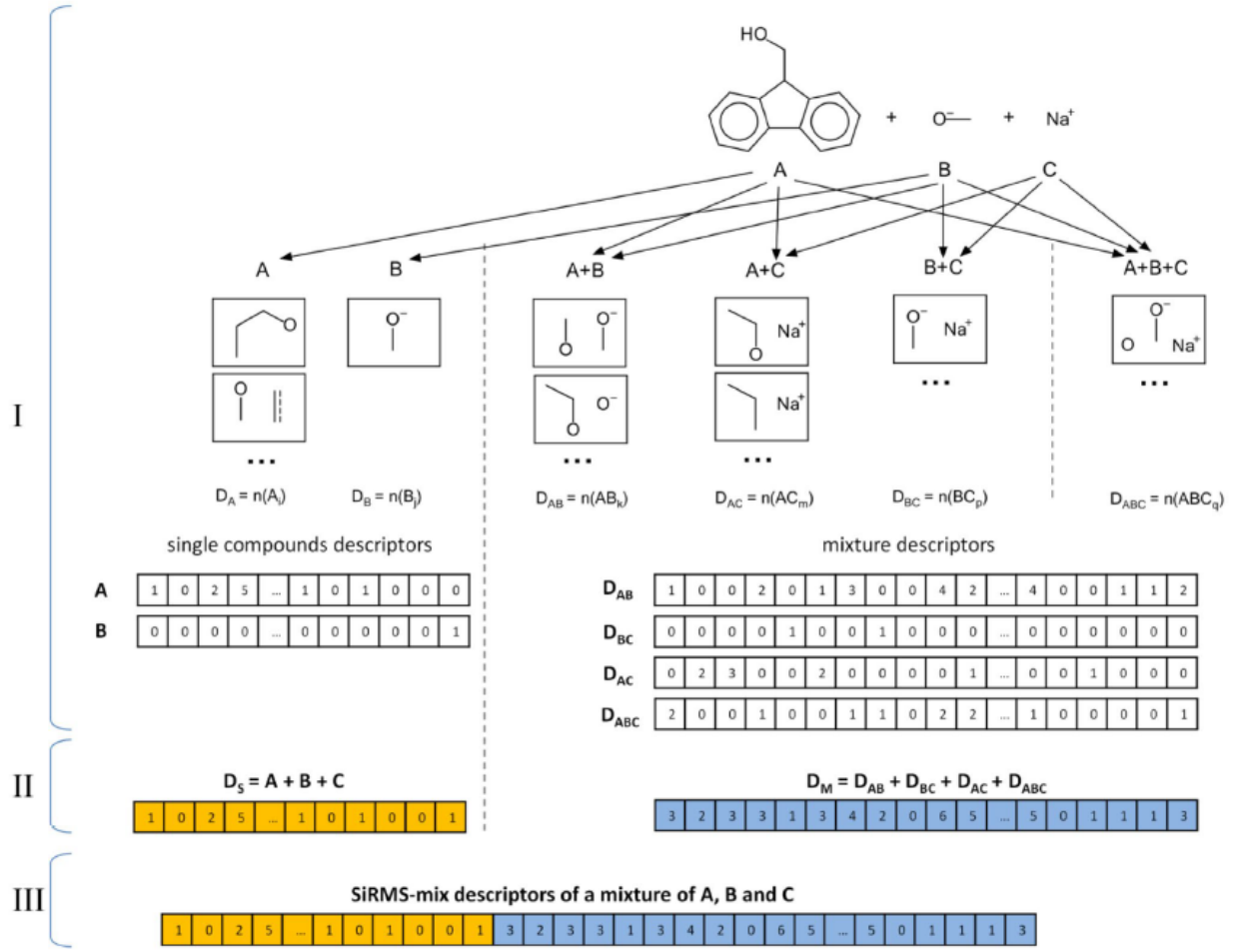


Fig. 3 Generation of simplex descriptors for a mixture of three components

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

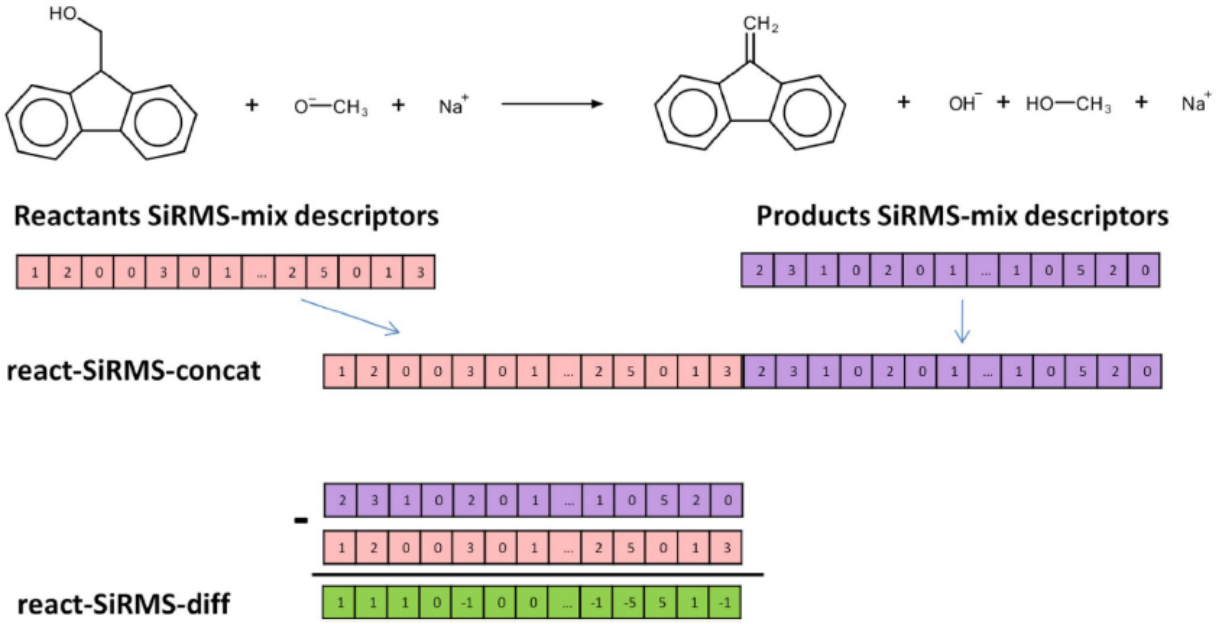


Fig. 4 Reaction descriptor vectors based on the concatenated product and reactant mixture descriptors (*react-SiRMS-concat*) and on their difference (*react-SiRMS-diff*)

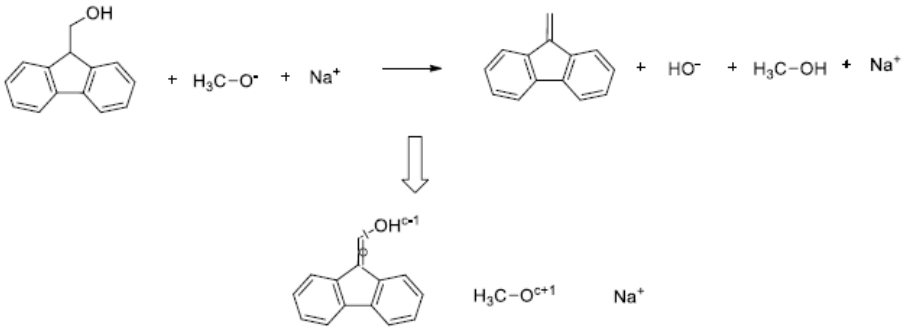


Fig. 5 An example of encoding of an E2 reaction into a Condensed Graph of Reaction. The broken and formed bonds are labeled by a crossing and a circle, respectively. Oxygen atoms changing their formal charges are denoted by symbols "c + 1" (negative-to-neutral) and "c − 1" (neutral-to-negative)

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).

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

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

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*).

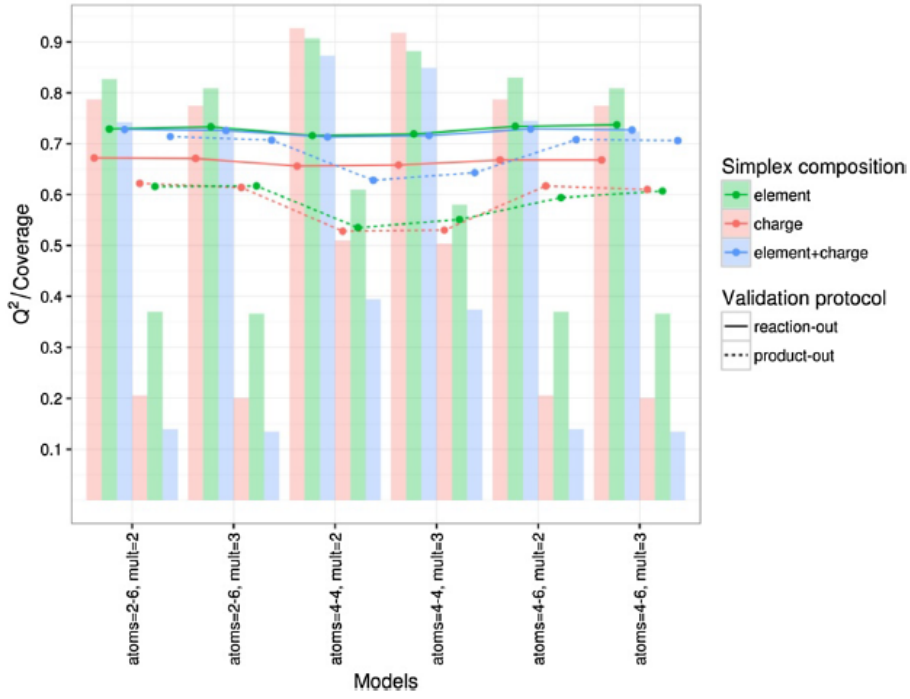


Fig. 6 The cross-validation determination coefficient Q^2 and the data coverage of the *react-SiRMS-concat* models as a function of size, complexity and composition of simplex descriptors. The model applicability domain was taken into account. *Solid and dashed lines* represent the Q^2 values, respectively, 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

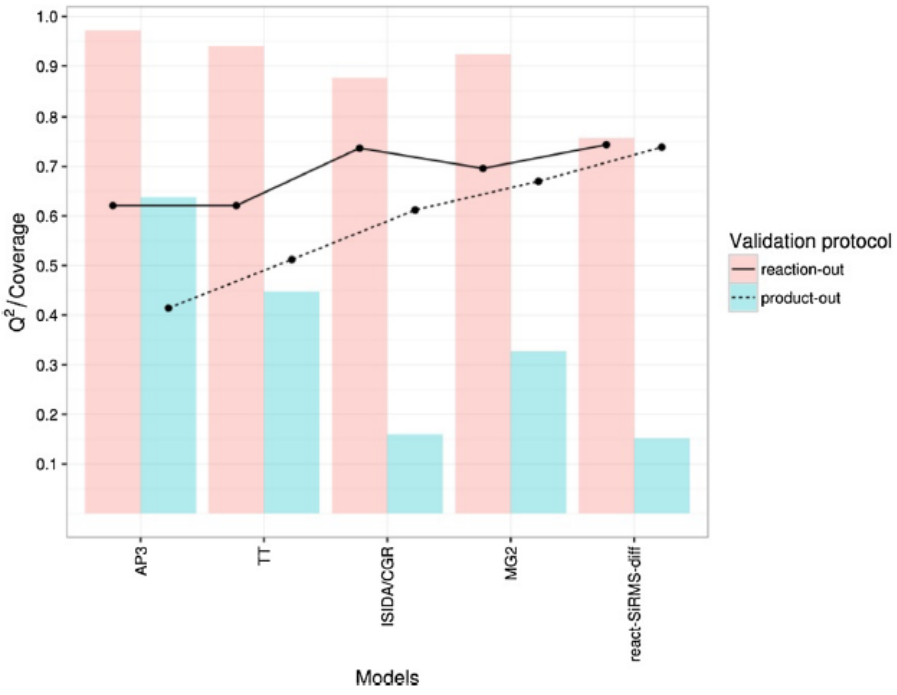


Fig. 7 Benchmarking of the models for E2 reaction rate constants involving different descriptors. The *dots* connected by *solid and dashed lines* represent Q^2_{DA} values calculated considering applicability domain for "reaction-out" and "product-out" validation strategies correspondingly. The *bars* 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.

Conclusive remarks

In this section, we described new type of descriptors for chemical reaction based on mixture representation of reagents and products. Unlike CGR-based descriptors this approach does not require Atom-to –Atom Mapping as a data pre-processing step. One could notice that CGR-based and SiRMS mixture descriptors are pretty competitive with respect to the model's quality. The Morgan based difference fingerprints were slightly better than CGR-based ISIDA fragments but worse than SiRMS on this particular dataset (if AD is taken into account). Other strategies for descriptor generation were worse. Accounting for bounding box applicability domain improves the model performance for all descriptors, but dramatically reduced the coverage, especially for CGR-based ISIDA and SiRMS mixture descriptors. It should be noted that ISIDA descriptors were optimized in this study. Fair large-scale benchmarking on an extended number of reaction datasets is still needed.

Reaction-out cross-validation and product-out cross-validation lead to quite different ranking of descriptors type according to the models performances. The bias introduced in reaction-out cross-validation could drastically affect the conclusions drawn on models performance.

Chapter 7.

Modeling of rate constants of Diels-Alder reactions

Diels-Alder reactions are one of the most well-known and important reactions in chemistry. It belongs to cycloaddition reaction class when two bonds are simultaneously formed with ring closure. These reactions are widely used as an efficient tool for cyclic system synthesis and especially popular in steroid chemistry. Diels-Alder reactions, called also [4+2]π cycloadditions, involve diene and alkene (called dienophile) which synchronously break three double bonds with formation of 2 single and unsaturated ring, exemplary reaction is shown on Scheme 4. Aromatic compounds with low aromaticity energy (furan, anthracene etc.) could also participate in reaction as dienes. Diels-Alder reactions proceed in one step without intermediate formation and characterized by second-order kinetics. Reactions proceed at room temperature or upon heating, no catalyst or irradiation are needed. Reaction has complex regio- and stereoselectivity driven by molecular orbital interactions. Reaction between asymmetric diene and dienophile leads to two possible products: *endo*-product when bulky substituents upon interaction are close to each other, and *exo*-product when they are distant, Figure 43.

Scheme 4.

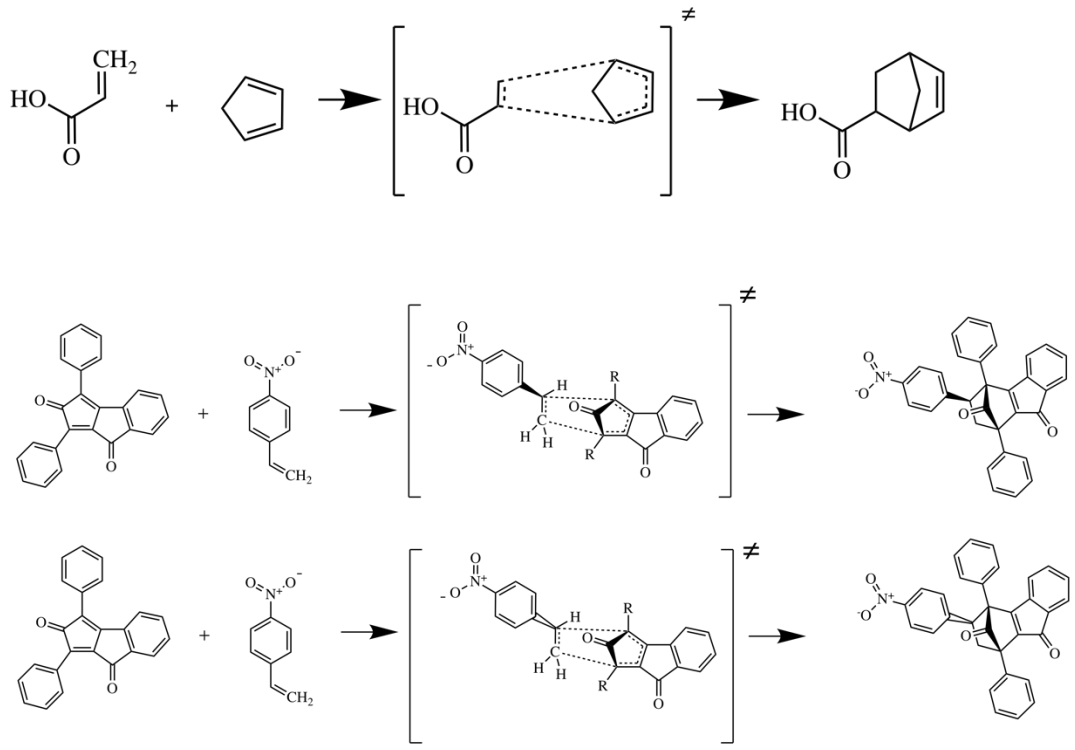


Figure 43. Example of endo (top) and exo (bottom) cycloaddition. Notice that the products of the reactions are diastereomers.

The modeling set contained 880 Diels-Alder reactions measured only in pure solvents. In cycloaddition not only initial stereo configuration of reagent molecules, but also endo and exo cycloaddition are characterized by different rate constants. So, even one reaction with specified stereo configuration of reagents can give two products. Typically this information was neglected in the initial sources, but for some reaction stereoconfiguration of reagents and products was known. For the latter we found the difference between rate constants values, which, fortunately, didn't exceed interlaboratory error of some 0.5 log*k* units. That's why, the data on stereoisomers were merged and rate constants were averaged. The modeling procedure and results are described in the article published in Russian Journal of Structural Chemistry, see below.

Journal of Structural Chemistry. Vol. 58, No. 4, pp. 650-656, 2017.

STRUCTURE–REACTIVITY RELATIONSHIP IN DIELS–ALDER REACTIONS OBTAINED USING THE CONDENSED REACTION GRAPH APPROACH

T. I. Madzhidov[1], T. R. Gimadiev[1,2], D. A. Malakhova[1], R. I. Nugmanov[1], I. I. Baskin[3], I. S. Antipin[1], and A. A. Varnek[1,2]

UDC 544.169:544.412.2

By the structural representation of a chemical reaction in the form of a condensed graph a model allowing the prediction of rate constants (log*k*) of Diels–Alder reactions performed in different solvents and at different temperatures is constructed for the first time. The model demonstrates good agreement between the predicted and experimental log*k* values: the mean squared error is less than 0.75 log units. Erroneous predictions correspond to reactions in which reagents contain rarely occurring structural fragments. The model is available for users at https://cimm.kpfu.ru/predictor/.

DOI: 10.1134/S0022476617040023

Keywords: [4+2]π-cycloaddition, Diels–Alder reaction, rate constant, condensed graph of the reaction, chemical reactions, chemoinformatics.

INTRODUCTION

Cycloaddition is one of the most used and important reactions in synthetic chemistry. They are especially interesting because these reactions lead to the formation of aromatic and unsaturated rings, which is important for medical chemistry [1]. Recently increased interest in them is explained by that many click chemistry reactions [2], especially those used in bioorthogonal chemistry [3], are cycloaddition reactions, e.g., azide–alkyne [4], alkyne–nitron [5], tetrazine–alkene [6], and so on. Perspectives to use a cycloaddition reaction in bioorthogonal chemistry are, as a rule, determined by its rate with regard to very low concentrations of reagents needed to provide biocompatibility [7]. Moreover, taking into account the possible formation of regioisomers in the cycloaddition reaction, their ratio can be estimated knowing the reaction rate constants. Thus, it is extremely important to predict cycloaddition reaction rates. In general, the reaction rate constants enable the estimation of not only the dynamics of chemical processes but also the calculation of product yields and their ratio.

Still, there are no well-established approaches for the reliable estimation of cycloaddition reaction rates in a wide range of solvents. The application of quantum chemistry methods for predicting the rates and conditions of cycloaddition reactions is not efficient because of difficulties to handle solvent effects and the conformational flexibility of molecules, as well as due to high computational costs. Methods based on the use of simple correlations, substituent and solvent constants are more successful in predicting rate constants [8, 9]. However, they can be applied only to relatively small congeneric data

[1]Kazan Federal University, Russia; Timur.Madzhidov@kpfu.ru. [2]University of Strasbourg, France. [3]Moscow State University, Russia. Translated from *Zhurnal Strukturnoi Khimii*, Vol. 58, No. 4, pp. 685-691, July-August, 2017. Original article submitted November 27, 2016.

sets of structurally homogeneous compounds or to one selected reaction proceeding in different solvents, which strongly restricts their applicability.

The employment of chemoinformatics methods makes it possible to avoid the restrictions mentioned above: the chemical reaction and its performance conditions are encoded by a set of descriptors, and the application of machine learning methods enables the construction of a model able to predict the reaction characteristics in a wide range of structures and conditions [10-15]. The main problem of coding is that the reaction is a complex object hierarchically relating several molecules. The approach of the condensed graph of the reaction, which was proposed by G. Vladutz [16] and developed by A. Varnek [17, 18], is a very convenient method to present reactions because it allows the coding of the ensemble of all reactants and reaction products of the into one sole pseudo-molecular (condensed) graph. In the condensed graph of the reaction (CGR) the transformation is represented in a reduced form where along with the usual chemical bonds remained unchanged in the reaction, the so-called dynamic bonds characterizing chemical transformations are specified (Fig. 1). To obtain CGR it is needed to perform the atom-to-atom mapping, i.e. to identify the correspondence between reagent and product atoms. Further it is possible to apply the standard approaches of chemoinformatics to calculate descriptors and then construct regression [10-12, 18] or classification models [19, 20] of chemical reactions. Moreover, CGR-based algorithms help identify errors of the atom-to-atom mapping [21] or refine the mapping using the consensus of several programs [22].

The [4+2]π cycloaddition reactions, so called the Diels–Alder reactions, are among the most interesting and widely used cycloaddition reactions. In this work for the first time the rate constants of Diels–Alder reactions are modeled using modern machine learning technologies. An example of the reaction from the database is given in Fig. 1.

At the Kazan Federal University the cycloaddition reactions have long been studied in A. I. Konovalov's group that collected a large set of different kinetic parameters (reaction rates, barriers, entropies, reaction activation volumes, etc.) measured under various conditions. These data served as a base to develop a model for the prediction of rates of the Diels–Alder reactions of interest.

EXPERIMENTAL

The Instant JChem program package [27] of the ChemAxon Company was used as a database management system of reactions. The structures of chemical compounds involved in reactions were standardized using the Standardizer tool of the

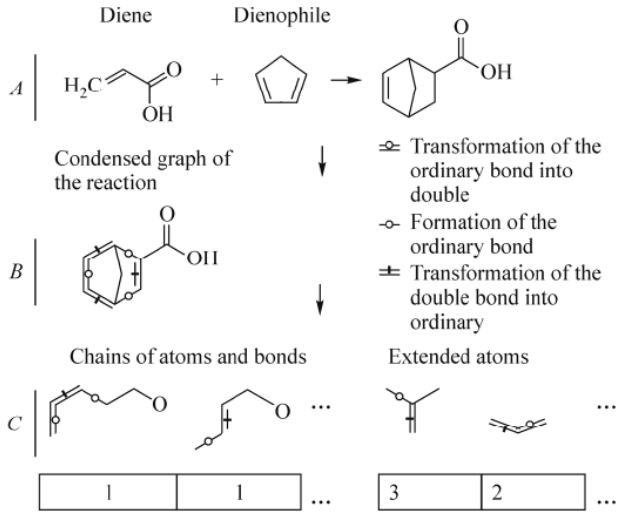


Fig. 1. Example of the reaction (*A*), the corresponding condensed graph (*B*), and ISIDA descriptors based on atomic chains and extended atoms (*C*). Digits correspond to the occurrence of the given fragment.

JChem program package [28]. The standardization procedure included: aromatization of structures, removal of isotopes, standardization of nitroso groups, aromatic N-oxides, azides, nitro groups, isocyanates, sulfones, tertiary N-oxides, removal of explicit hydrogen atoms, and the atom-to-atom mapping. The errors of the atom-to-atom mapping were identified and corrected manually. CGRs were generated using the in-house CGR Condenser program.

ISIDA fragment descriptors for the condensed graphs were generated using the Fragmentor program [17]. Catalan constants (SPP [29], SA [30], SB [31]), Kamlet–Taft constants (α [32], β [33], π* [34]) were used as descriptors of the solvent as well as the descriptors, characterizing the effect of solvent polarity and polarizability: the Born function $f_B = \frac{\varepsilon - 1}{\varepsilon}$, the Kirkwood function $f_K = \frac{\varepsilon - 1}{2\varepsilon + 1}$, $f_1 = \frac{\varepsilon - 1}{\varepsilon + 1}$, and $f_2 = \frac{\varepsilon - 1}{\varepsilon + 2}$ (ε is the solvent permittivity), $g_1 = \frac{n^2 - 1}{n^2 + 2}$, $g_2 = \frac{n^2 - 1}{2n^2 + 1}$, $h = \frac{(n^2 - 1)(\varepsilon - 1)}{(2n^2 + 1)(2\varepsilon + 1)}$ (n is the solvent refractive index n_D^{20}).

Support vector regression (SVR) [24] was used as a machine learning method. It requires the choice of some hyperparameters: the kernel type, the tube width parameter ε, and the penalty parameter C. Note that the model performance varies as a function of fragment descriptor types. The optimal types of descriptors and hyperparameters of the SVR method were selected using the SVMOptimizer program [25].

The root-mean-square error (RMSE) and the determination coefficient (Q^2) on the test sets were used to describe the model performance

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (x_i^{\mathrm{pred.}} - x_i^{\mathrm{exp.}})^2}{N}},$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{N} (x_i^{\mathrm{pred.}} - x_i^{\mathrm{exp.}})^2}{\sum_{i=1}^{N} (x_i^{\mathrm{pred.}} - \langle x_i^{\mathrm{exp.}} \rangle)^2},$$

where $x_i^{\mathrm{pred.}}$, $x_i^{\mathrm{exp.}}$ are the predicted and experimental log$k$ values for the i-th reaction; $\langle x_i^{\mathrm{exp.}} \rangle$ is the average logarithm of the rate constant; N is the number of objects in the sample. The test and training sets were selected using a ten times repeated fivefold cross-validation procedure.

The prediction was obtained by averaging the predictions of individual models obtained with the use of various data subsets obtained during the fivefold cross-validation procedure.

RESULTS AND DISCUSSION

The models were constructed using a dataset on reaction rates and their performance conditions extracted manually from candidate and doctor theses of Academician A. I. Konovalov's group. In total, more than 880 reactions performed in various solvents were extracted. Toluene (347 reactions), chlorobenzene (172), and dichloroethane (170) were most abundant. In the extracted reactions the most abundant dienophiles were maleic anhydride, tetracyanoethylene, and *para*-benzoquinone; among dienes substituted phencyclones, pentacenes, and tetracenes along with 2,5-dicarbo-methoxy-3,4-diphenylcyclopentadiene occurred most often. The rate constants of cycloaddition reactions ranged from –7.7 to +8.5 log units. The occurrence frequency distribution histogram of rate constants of cycloaddition reactions is given in Fig. 2.

We used ISIDA fragments as the descriptors characterizing the reaction transformation [17]. Each descriptor value is equal to the occurrence frequency of the given fragment in the molecule. ISIDA descriptors calculate the number of all possible fragments with a certain topology. There are two main topologies of ISIDA fragments: (i) atomic chains with a given length or (ii) atoms with the nearest environment (augmented atoms) – the fragments contain all atoms remote from the central atom at a given topological distance (Fig. 1). In the fragmentation using different options only fragments containing at

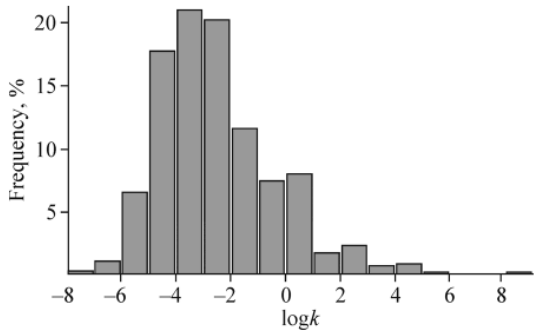


Fig. 2. Occurrence frequency distribution histogram of rate constants of Diels–Alder reactions.

least one dynamic bond can be left, only shortest path fragments can be kept, and a description of the fragment can be controlled: information on atom types, bonds, occurrence or change in formal atomic charges can be added or ignored.

Thus, for the used dataset there were 616 types of descriptor descriptions different only in the fragmentation method of Diels–Alder CGR.

From the obtained sets of descriptors it is necessary to choose the best ones that enable the construction of the model with the highest predictive ability. The predictive ability of the model was assessed by the fivefold cross-validation procedure. In this procedure the whole data set is divided into 5 equal parts, one of which is selected into the so-called test set and the other 80% objects compose the so-called training set. The model is based on the test set. By means of the model obtained the characteristics of test set objects are predicted. The procedure is repeated for each of the five parts in turn so that each object is present only once in the test set. At the end, the predictions are combined and the prediction error is calculated. The model performance was estimated using the determination coefficient (Q^2) RMSE of the predicted values from the experimental ones. In order to remove the effect of the order of compounds in the test set on the performance index, the dataset was randomly shuffled before the cross-validation procedure. This procedure was repeated 10 times and the log$k$ values predicted for a given reaction were averaged.

In order to predict rate constants for new reactions all models generated during the cross-validation procedure were maintained and used to predict log$k$. Finally, 50 predictions were obtained for each object and then averaged. This approach similar to the bagging technique [23] makes it possible to decrease the prediction error and to increase the stability of the model because the results of calculations are no more biased with the data set composition. The model performance the can be evaluated by averaging the predictions for objects from test sets. By comparing the obtained predictions with the experimental data RMSE and Q^2 were calculated by the formulas given in the EXPERIMENTAL section.

SVR was applied as the machine learning method [24]. Within this method in the descriptor space a multidimensional tube with a radius set by the ε parameter is constructed so that objects fall into it. If the object is outside the tube, the model is fined proportionally to the remoteness of the object from the tube surface and the penalty parameter C. By means of the similarity kernels it is possible to construct nonlinear dependences whose complexity is determined by the kernel type and its parameters. In the work three types of kernels are used: linear, polynomial, and Gaussian. Thus, there are two hyperparameters (the coefficient C and the parameter ε) and the kernel type which are selected so that to provide the maximum predictive ability of the model. For each type of the descriptors the optimal hyperparameters and kernels can differ. Since the number of combinations of descriptors and hyperparameters of the SVR model is very high, they were selected using a genetic algorithm [25] so that to guarantee the optimal predictive ability of the model during the fivefold cross-validation procedure.

The highest predictive ability is demonstrated by the models constructed with the use of fragment descriptors based on chains with a length varying from one to six atoms with regard to information about charges on them. RMSE of the

predicted values in comparison with the experimental ones for the model using a consensus approach was 0.75 log units. The determination coefficient Q^2 for the model was 0.87.

The plot of the correspondence between the predicted and experimental values is depicted in Fig. 3. It seems to be impossible to construct a better model due to the occurrence of noise in the data, which appears because of several reasons: different measurement procedures, neglection of several aspects of the structure (e.g., in this work we did not take into account the stereochemistry of attachment), random errors. Apparently, a deviation of the predicted data from the experimental ones on average cannot be less than the noise in the data. It is difficult to estimate this noise, however, our prediction error agrees in the order of magnitude with those obtained previously from the models for bimolecular nucleophilic substitution reactions [11]. In it, some examples of deviations between rate constants obtained in different works are presented, which in general confirm the correspondence of noise in these prediction errors. Moreover, a similar prediction error was obtained in the model predicting the bimolecular elimination rate [12]. Thus, we consider the prediction quality of the proposed model to be comparable with the reproducibility of the rate constant in different experiments.

In order to evaluate the model performance and its limitations the objects were analyzed for which significant (by more than 3·RMSE) deviations between the experimental and predicted values were found. The analysis of these prediction errors has shown that they are due to the occurrence of specific, rare structural fragments (Fig. 4). If a reaction containing these rare fragments occurs in the test set, the training set, as a rule, involves a statistically insignificant number of reactions or does not involve the reactions of this type at all. Therefore the effect of these rare fragments on the reaction rate cannot be

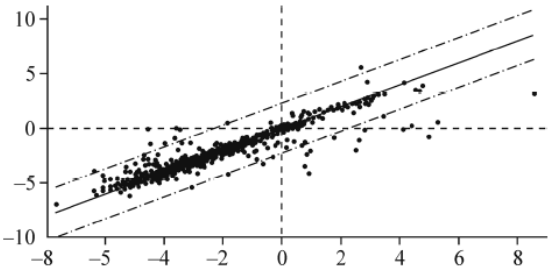


Fig. 3. Predicted rate constants of [4+2]π cycloaddition reactions in comparison with the experimental ones. Solid line corresponds to the ideal correlation between the predicted and experimental values. Dashed line corresponds to deviations of the predicted values from the observed ones on 3·RMSE.

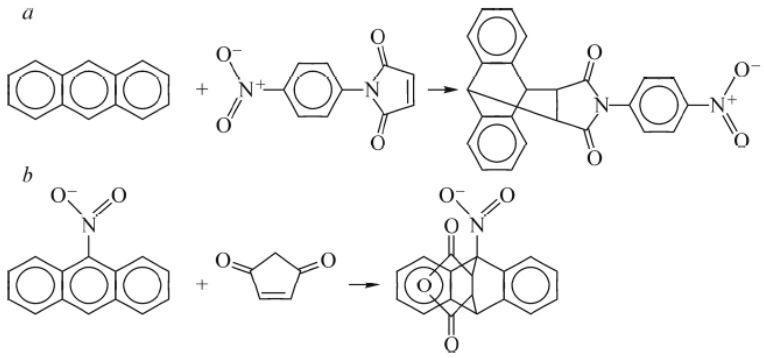


Fig. 4. Examples of prediction errors. Reaction conditions (solvent, temperature), predicted and experimental values: toluene, 60 °C, pred.–3.16, exp. 0.76 (*a*); 1,4-dioxane, 130 °C, pred. –1.92, exp. –4.23 (*b*). Both reactions contain a nitro substituent absent in other reactions of the training set.

predicted, and the reaction rate obtained contains a potentially large error. Note that the prediction errors of this type can be successfully identified by different approaches estimating the applicability field of the model, e.g., the fragment control method [26]. Within this approach, if a condensed graph of some test reaction contains a fragment that did not occur in graphs of objects in the training set, then the prediction of the model obtained based on this training set is considered to be unreliable. Since in the final model the prediction is made based on the consensus of many models obtained on different sets of objects, then the applicability domain is found separately for each individual model. If the prediction for more than 70% models involved in the consensus turns out to be unreliable, then the prediction of the consensus model is also considered to be unreliable. This threshold value was found by considering various possible values (from 0% to 100%) so that to minimize the prediction error during the cross-validation procedure.

The obtained model is available for users via the on-line predictor hosted on the server of the Laboratory of Chemoinformatics and Molecular Modeling at the Kazan Federal University at https://cimm.kpfu.ru/predictor/. The user can draw or download a reaction of interest and to set solvent and temperature. Predicted log*k* value is returned along with the information reporting whether a given reaction belongs to the model's applicability domain.

CONCLUSIONS

By means of the condensed graph approach and generated descriptors of the chemical transformation in combination with solvent and temperature descriptors we have obtained for the first time the model predicting the rate constant of Diels–Alder reactions involving various reagents, which occur in many organic solvents. It is shown that the prediction accuracy is comparable with the level of experimental noise. The analysis of prediction errors also shows that the model performance is sufficiently high for the identification of data errors and objects with the unique structure with respect to this set of reactions. By means of the data obtained a method of evaluating whether the reaction of interest belongs to the applicability domain of the model or not was implemented. The model obtained is available at the server https://cimm.kpfu.ru/predictor/.

The work was supported by the Russian Scientific Foundation (project No. 14-43-00024).

The authors are grateful to the ChemAxon Company for providing software tools.

REFERENCES

1. M. Hartenfeller, M. Eberle, P. Meier, C. Nieto-Oberhuber, K.-H. Altmann, G. Schneider, E. Jacoby, and S. Renner, *J. Chem. Inf. Model.*, **51**, No. 12, 3093 (2011).
2. H. C. Kolb, M. G. Finn, and K. B. Sharpless, *Angew. Chem. Int. Ed. Engl.*, **40**, No. 11, 2004 (2001).
3. E. M. Sletten and C. R. Bertozzi, *Acc. Chem. Res.*, **44**, No. 9, 666 (2011).
4. H. C. Kolb and K. B. Sharpless, *Drug Discov. Today.*, **8**, No. 24, 1128 (2003).
5. D. A. MacKenzie, A. R. Sherratt, M. Chigrinova, L. L. Cheung, and J. P. Pezacki, *Curr. Opin. Chem. Biol.*, **21**, 81 (2014).
6. M. L. Blackman, M. Royzen, and J. M. Fox, *J. Am. Chem. Soc.*, **130**, No. 41, 13518 (2008).
7. Y. Gong and L. Pan, *Tetrahedron Lett.*, **56**, No. 17, 2123 (2015).
8. V. A. Pal`m, *Principles of Quantitative Theory of Organic Reactions* [in Russian], Khimiya, Leningrad (1977).
9. V. A. Pal`m, *Uspekhi Khimii*, **30**, No. 9, 1069 (1961).
10. R. I. Nugmanov, T. I. Madzhidov, G. R. Khaliullina, et al., *J. Struct. Chem.*, **55**, No. 6, 1026 (2014).
11. T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, et al., *Russ. J. Org. Chem.*, **50**, No. 4, 459 (2014).
12. T. I. Madzhidov, A. V. Bodrov, T. R. Gimadiev, et al., *J. Struct. Chem.*, **56**, No. 7, 1227-1234 (2015).
13. A. A. Kravtsov, P. V. Karpov, I. I. Baskin, et al., *Dokl. Chem.*, **441**, No. 1, 314 (2011).
14. A. A. Kravtsov, P. V. Karpov, I. I. Baskin, et al., *Dokl. Chem.*, **440**, No. 2, 299 (2011).
15. N. M. Halberstam, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Mendeleev Commun.*, **12**, No. 5, 185 (2002).
16. G. E. Vladutz, *Inf. Storage Retr.*, **1**, Nos. 2/3, 117 (1963).

17. A. Varnek, D. Fourches, F. Hoonakker, and V. P. Solov'ev, *J. Comput. Aided. Mol. Des.*, **19**, Nos. 9/10, 693 (2005).
18. F. Hoonakker, N. Lachiche, A. Varnek, et al., *Int. J. Artif. Intell. Tools.*, **20**, No. 2, 253 (2011).
19. A. De. Luca, D. Horvath, G. Marcou, et al., *J. Chem. Inf. Model.*, **52**, No. 9, 2325 (2012).
20. G. Marcou, J. Aires de Sousa, D. A. R. S. Latino, et al., *J. Chem. Inf. Model*, **55**, No. 2, 239 (2015).
21. C. Muller, G. Marcou, D. Horvath, et al., *J. Chem. Inf. Model*, **52**, No. 12, 3116 (2012).
22. T. I. Madzhidov, R. I. Nugmanov, T. R. Gimadiev, A. I. Lin, I. S. Antipin, and A. A. Varnek, *Butlerovskie Soobshcheniya*, **44**, No. 12, 170 (2015).
23. L. Breiman, *Mach. Learn.*, **24**, No. 2, 123 (1996).
24. H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Support vector regression machines. Advances in Neural Information Processing Systems*, M. C. Mozer, J. I. Jordan, and J. I. Petsche (ed.); MIT Press, Vol. 9, 155 (1997).
25. D. Horvath, J. Brown, G. Marcou, et al., *Challenges.*, **5**, No. 2, 450 (2014).
26. I. V. Tetko, I. Sushko, A. K. Pandey, et al., *J. Chem. Inf. Model*, **48**, No. 9, 1733 (2008).
27. *InstantJChem 15.7.27.0. ChemAxon*; http://www.chemaxon.com (2015).
28. *Standardizer, JChem 15.8.3.0. ChemAxon*; http://www.chemaxon.com (2015).
29. J. Catalán, V. López, P. Pérez, et al., *Liebigs Ann.*, **1995**, No. 2, 241 (1995).
30. J. Catalán and C. Díaz, *Liebigs Ann.*, **1997**, No. 9, 1941 (1997).
31. J. Catalán, C. Díaz, V. López, et al., *Liebigs Ann.*, **1996**, No. 11, 1785 (1996).
32. R. W. Taft and M. J. Kamlet, *J. Am. Chem. Soc.*, **98**, No. 10, 2886 (1976).
33. M. J. Kamlet and R. W. Taft, *J. Am. Chem. Soc.*, **98**, No. 2, 377 (1976).
34. R. W. Taft and M. J. Kamlet, *J. Am. Chem. Soc.*, **98**, No. 10, 2886 (1976).

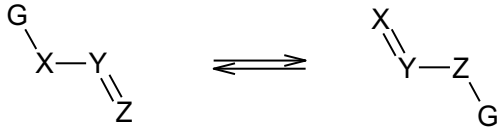
Conclusions remarks

The CGR-based approach and developed workflow were used to build consensus model for Diels-Alder reactions rate constant prediction. The model displays a reasonable predictive performance: R^2 =0.87 and RMSE = 0.75 log units in cross-validation. The model was published on line on http://cimm.kpfu.ru/predictor.

Chapter 8.

Modeling of tautomeric equilibrium constants

Tautomerism represents one of the most common phenomenon in organic chemistry. According to IUPAC definition, the tautomerism is the easily undergoing isomerism of the following general type



The most common form of tautomerism is prototropy (prototropic tautomerism), where G is hydrogen. The importance of tautomerism in drug design was pointed out in numerous publications [129, 207–212]. The main problem with tautomers is related to the fact that each tautomer behaves as an individual chemical compound with its own spectral, chemical and physical properties and, at the same time, all tautomers of a given compound coexist in dynamic equilibrium. For this reason, identification of molecules and its properties that are able to tautomerize can be challenging.

Since a molecule can exist in several forms with different descriptor vectors and structure fingerprints, one have to decide which form should be taken into account in the course of modeling or storage in chemical database. It has been reported several times that several tautomeric forms of the same compound can be found in the same database, and even different prices are given for them [213, 214]. In QSAR studies and chemical database management only the most stable (most populated) tautomeric form in water is considered mainly. It should however be taken into account that guessing the most stable form may sometimes be rather difficult problem, because tautomer equilibrium is known to be solvent and temperature dependent. Selecting one of several alternative tautomers is usually performed using rules of tautomer canonicalization [215–217]. In structure-based drug design several tautomeric forms with energy lying within the given energetic window are considered. In this case relative stabilities of tautomers can be estimated using quantum- chemical [218, 219] or force-field modeling [220, 221].

The dataset of tautomeric equilibria was collected from reference book [1]. The tautomeric equilibria were represented as reactions that turn one compound to another, in order to fit equilibria to the developed approach. In this representation equilibria constants belong to one exact direction of process. This required to check additionally that left and right parts of the reaction equation are represented in a unified way. E.g. keto-enol tautomerism have to be represented only as transformation of ketone into enol. Additional problem we came across is, that using dynamic bond concept in initial

formulation of CGR [2] with implicit hydrogens, encoding of neutral - zwitter-ion tautomeric equilibria produces CGR without any dynamic bond. Therefore, dynamic atoms were introduced for description of such reactions.

Initial curated set of 840 records was curated for modeling, duplicated reactions' rates were averaged and resulted modeling set was divided into training and external set. External test set was selected for evaluation of the model and comparison with quantum chemistry calculations. As the result 739 equilibria in different conditions were selected for training set and 46 equilibria were selected to external test set. Additional comparison of the model with commercial Tautomerizer tool (ChemAxon) for prediction of tautomer populations in water under room temperature was done. The results of modeling, quantum chemistry benchmarking and commercial tool comparison were published in the article in Journal of Computer-Aided Molecular Design that is shown below.

Journal of Computer-Aided Molecular Design
https://doi.org/10.1007/s10822-018-0101-6

Assessment of tautomer distribution using the condensed reaction graph approach

T. R. Gimadiev[1,2] · T. I. Madzhidov[2] · R. I. Nugmanov[2] · I. I. Baskin[2,3] · I. S. Antipin[2] · A. Varnek[1]

Received: 20 September 2017 / Accepted: 18 January 2018

Abstract
We report the first direct QSPR modeling of equilibrium constants of tautomeric transformations ($\log K_T$) in different solvents and at different temperatures, which do not require intermediate assessment of acidity (basicity) constants for all tautomeric forms. The key step of the modeling consisted in the merging of two tautomers in one sole molecular graph ("condensed reaction graph") which enables to compute molecular descriptors characterizing entire equilibrium. The support vector regression method was used to build the models. The training set consisted of 785 transformations belonging to 11 types of tautomeric reactions with equilibrium constants measured in different solvents and at different temperatures. The models obtained perform well both in cross-validation ($Q^2 = 0.81$ RMSE = 0.7 $\log K_T$ units) and on two external test sets. Benchmarking studies demonstrate that our models outperform results obtained with DFT B3LYP/6-311++G(d,p) and ChemAxon Tautomerizer applicable only in water at room temperature.

Keywords QSPR · Support vector regression · Condensed graphs of reaction · Tautomerism

Introduction

Handling tautomers is a real challenge in cheminformatics [1–7]. The main problem is related to the fact that different tautomers of the same chemical compound are described by different descriptor vectors, which may affect their registration in chemical databases, the results of similarity searching, building and application of quantitative structure–activity/property relationships (QSAR/QSPR) models for any physico-chemical or biological property. According to the estimations by Sitzmann et al. [8], tautomerism is possible for more than 2/3 of unique structures in the Chemical Structure DataBase (CSDB) of the National Cancer Institute containing some 103.5 million structure records. In order to handle tautomers in chemical databases, some enumeration rules were suggested [8–10]. However, duplicates corresponding to different tautomers are still present in public and proprietary databases [11, 12]. The canonicalization rules are implemented in free or commercial tools [13–21] capable of enumerating all possible tautomers corresponding to a given chemical structure.

Different scenarios of accounting for tautomerism in the modeling studies have been reported in the literature. In the case of structure-based drug design, several tautomers with energy values within a given energetic window are considered. However the relative stability of tautomers, which is usually estimated using quantum chemical [22, 23] or force field methods [24, 25], is rarely taken into account in scoring functions. Typically, descriptors used in QSAR/QSPR studies or in similarity-based virtual screening are calculated for the canonical tautomeric form, although some efforts to account for tautomeric equilibrium have been recently reported. Thus, fuzzy topological pharmacophoric triplets [26, 27] accounting for tautomers' populations are a part of the ISIDA descriptors [28], which were largely used in structure–activity modeling and virtual screening [29–31].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10822-018-0101-6) contains supplementary material, which is available to authorized users.

✉ A. Varnek
varnek@unistra.fr

[1] University of Strasbourg, 1 rue Blaise Pascal, 67000 Strasbourg, France

[2] Kazan Federal University, 18 Kremlyovskaya Str., Kazan, Russian Federation 420008

[3] Department of Physics, Moscow State University, Moscow, Russian Federation 119991

Published online: 29 January 2018

For a particular case of two tautomers, their relative populations can be derived from the equilibrium constant K_T :

$$K_T = \frac{[tautomer2]}{[tautomer1]} \quad (1)$$

For two tautomers one could consider thermodynamic cycle (Fig. 1) where it was noticed that two tautomeric forms have the same anion and protonated cation (it does not mean that interconversion of tautomers goes through naked anion or protonated cation). According to it, K_T can be expressed in terms of acidity (basicity) constants $K_a(K_b)$ of individual tautomers:

$$K_T = \frac{K_a(tautomer1)}{K_a(tautomer2)} \quad (2a)$$

$$K_T = \frac{K_b(tautomer1)}{K_b(tautomer2)} \quad (2b)$$

or in logarithmic form:

$$\log K_T = pKa(2) - pKa(1) \quad (3a)$$

$$\log K_T = pK_b(2) - pK_b(1) \quad (3b)$$

Alternatively, the logarithm of the equilibrium constant can be calculated as a difference between Gibbs free energies (ΔG) of tautomers

$$\log K_T = \Delta G(2) - \Delta G(1) \quad (4)$$

The fractions of the "reactant" (T1) and the "product" (T2) tautomers at equilibrium can be easily calculated as

$$T1 = \frac{1}{K_T + 1} \; and \; T2 = 1 - T1 \quad (5)$$

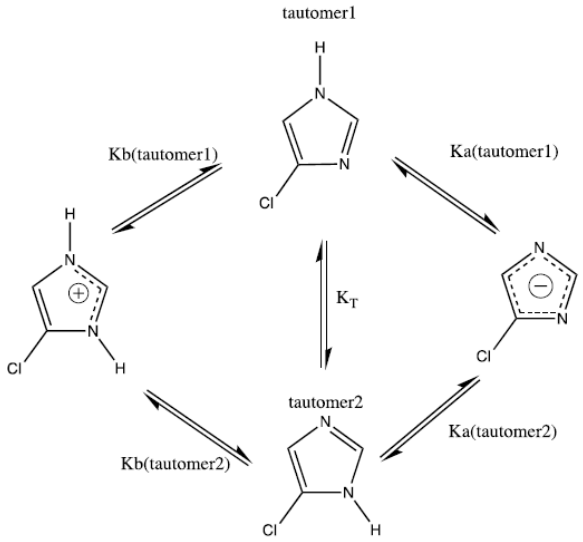


Fig. 1 Thermodynamic cycle showing relationships between basicity, acidity and tautomersation equilibrium constant

Thus, the major tautomer is defined by the sign of $\log K_T$: its negative values correspond to domination of the "reactant", otherwise this is a "product" which dominates.

Thus, the logarithm of equilibrium constant $\log K_T$ can be quantitatively estimated using either the relative stability of tautomers according to Eq. 4 or available data on the basicity (acidity) of tautomeric forms according to Eq. 3a and 3b. The former strategy is usually performed with the help of quantum chemical approaches (HF, DFT or MP2) where solvent effects are taken into account using implicit solvation models [32] or so called cluster-continuum approach [33] combining explicit and implicit solvent. The latter usually relies on the tools [14, 34] implementing QSPR models for *pK*a. Both approaches may lead to significant errors in K_T assessment because of summation of uncertainties in ΔG (Eq. 4) or *pK*a (Eq. 3a and 3b) predictions.

An obvious weak point of the existing tools is related to the treatment of solvent effects. The implicit solvation models used in quantum chemistry tools for free energy calculations could hardly reach precision greater than 2–3 kcal/mol [35–37] which corresponds to error in $\log K_T$ about 1.5–2.2 log units. It should also be noted that available models for *pK*a and tautomer population [14, 34] prediction consider only aqueous solutions. All these seriously limit performance of existing tools.

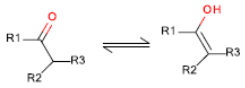
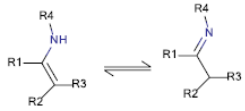
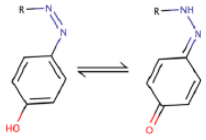
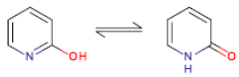
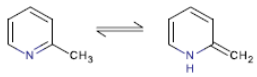
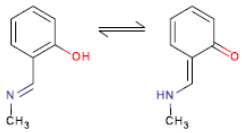
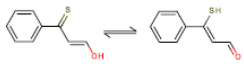
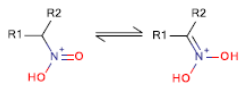
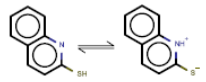
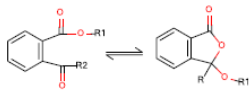
In this paper, we report predictive QSPR models obtained directly for equilibrium constants of different types of tautomeric transformations in different solvents and at different temperatures. We considered only transformations involving a single "reactant" and a single "product". These two species were merged in one sole molecular graph—condensed graph of reaction (CGR)—for which molecular descriptors were generated and further used in building QSPR models for $\log K_T$. Below we demonstrate that such direct $\log K_T$ modeling allows modeling tautomeric equilibria in any solvent and at any temperature, and to overcome the limitations of existing methods relying on Eqs. 3a and 3b and 4.

Computational details

Data

The dataset consists of 786 tautomeric transformations extracted from the Palm's handbook [38] representing unique collection of thermodynamic and kinetic data on chemical reactions. This set includes 367 unique structural transformations in solution at atmospheric pressure. Out of them, 267 equilibria proceeded at single combinations solvent/temperature, whereas for the other $\log K_T$ values were measured in different solvents and/or temperatures. The considered transformations belong to 11 types (Table 1) for which the left-hand ("reactant") and the right-hand ("product") tautomers are defined conventionally. For example, in

Table 1 Types of tautomeric equilibria and related solvents considered in this study

Type	Typical reaction	Number of reactions in different solvents			
		Total	Water	Mixtures with water	Non-aqueous
Keto-Enol (I)		288	4	71	213
Amino-Imino (II)		190	6	36	148
Azo-Hydrazone (III)		3	1	0	2
Pyridol-Pyridone (IV)	Example: 	44	28	1	15
Pyridinoid-Pyridonoid (V)	Example: 	4	4	0	0
Phenol-Imine Keto-Amine (VI)	Example: 	33	0	25	8
Thione-Enol – Keto-Thiol (VII)	Example: 	10	0	10	0
Amine-Thione – Imine-Thiol (VIII)		21	12	0	9
Nitro-Aci (IX)		8	8	0	0
Neutral Form - Zwitterion (X)		18	14	0	4
Chain-Ring (XI)		120	3	40	77
Total		**739**	**80**	**183**	**476**

keto-enol tautomeric transformations ketones were drawn on left-hand side of reaction equation and enoles on the right-hand side. All data were stored in the database built with the ChemAxon InstantJChem [39]. The structures were standardized with the ChemAxon Standardizer [40].

Considered equilibrium constants were measured in 23 pure solvents and 5 water-organic solvent with different proportions of components, see Fig. 2 *bottom*. For duplicated reactions extracted from different sources, an average $\log K_T$ value was calculated. Analysis of duplicates shows that the difference of reported experimental values can approach 0.6 log units, see Table 2.

The $\log K_T$ values vary in the range from −8.2 to 12.82 with a strong peak around 0 (Fig. 2 *top*). Some 97% of data falls into the range [−5;+5] and 75% into the range [−1;+1]. The equilibrium constants were measured at the temperature range from 233 to 373 K, with the majority of experiments performed at room temperature (Fig. 2 *middle*).

Collected data were split into training set and external validation set containing 739 and 46 equilibria, respectively. The external set consisted of two subsets TEST1 (20 equilibria) and TEST2 (26 equilibria). The former contained only non-unique transformations (i.e., tautomeric equilibria occurred in the training set but proceeded under different conditions), whereas the latter contained only unique transformations. Each subset included all types of tautomeric transformations mentioned in Table 1 and presented by at least five reactions. In such a way, these two sets allowed us to assess predictive performance of developed models both for new equilibria (TEST2) and for known equilibria proceeding under new conditions (TEST1). Both sets were chosen randomly under condition of occurrence of 0–4 representatives from each tautomerism type. Both training and test sets are available for the users at our WEB server http://cimm.kpfu.ru/Tautomers_JCAMD_2018.

Condensed graphs of reaction (CGR)

The CGR [44, 45] were used to represent each tautomeric transformation as a single graph. CGR contains both conventional chemical bonds (e.g. single, double, triple, aromatic *etc*.) and so called "dynamic" bonds characterizing chemical transformations, i.e. breaking or forming a bond or changing bond order (see Table 3). Previously this approach was used in the modeling of rate constants of various reactions (S_N2 [45–47], bimolecular elimination [48], Diels–Alder [49]), optimal condition prediction [50] as well as for detection of erroneous atom-to-atom mapping [51]. Since we use hydrogen suppressed molecular graphs (all hydrogens are implicit), conventional CGR [44, 45] can't describe zwitterionic tautomerism related to cleavage/formation of X–H bonds only (X is a heavy atom) and are not accompanied by transformations of any other bonds. In order to describe structural reorganization in CGR for this type of tautomers, we have introduced *dynamic atoms*, characterizing changes of atomic charges upon the transformation (Fig. 3b).

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 tautomeric transformations were extracted in RDF format [52]; (2) atom-to-atom mapping was performed using the ChemAxon/Standartizer tool [53]; (3) the errors of mapping were fixed, first, using Indigo [54], then in *in-house* software [55]; and (4) CGRs corresponding to the transformations were built using the *in house* CGR-Designer tool and stored in SDF format (see detailed explanations in Supporting Information).

Notice that CGR corresponding to a given tautomeric transformation depends on the direction of this reaction, i.e., which tautomer is taken as a reactant and which one as a product. Here, we strictly used canonical representation of tautomeric transformations shown in Table 1.

Descriptors

Descriptor vector for each reactions resulted from concatenation of structural descriptors and parameters describing experimental conditions (solvent and temperature).

Chemical transformations were encoded by the ISIDA descriptors generated for CGR using the ISIDA Fragmenter [44] program. ISIDA descriptors represent subgraphs of different topologies and sizes. Each subgraph is considered as a descriptor whereas its occurrence in a molecule is the descriptor's value. In this study, two types of subgraphs were considered: sequences of atoms and/or bonds and augmented atoms (atoms with first, second, etc coordination spheres). The length of the fragments varied from n_{min} to n_{max} atoms with n_{min} = 1–4 and n_{max} = 2–10. By default, the algorithm searches the shortest possible path between two atoms in case of sequences, but all path exploration option has also been tested. For each fragmentation type, descriptors vector included either all generated fragments or only fragments containing dynamic bonds or atoms. In such a way, 616 different sets of descriptors were generated. The size of descriptor vectors varied from 15 to 10,000.

Each solvent was described by 15 descriptors that represent polarity, polarizability, H-acidity and basicity: Catalan SPP [56], SA [57], and SB constants [56], Camlet–Taft constants α [58], β [59], and π* [60], four functions depending on the dielectric constant ε (Born $f_B = (\varepsilon - 1)/\varepsilon$ and Kirkwood $f_K = (\varepsilon - 1)/(2\varepsilon + 1)$ functions, $f_1 = (\varepsilon - 1)/(\varepsilon + 1), f_2 = (\varepsilon - 1)/(\varepsilon + 2)$), three

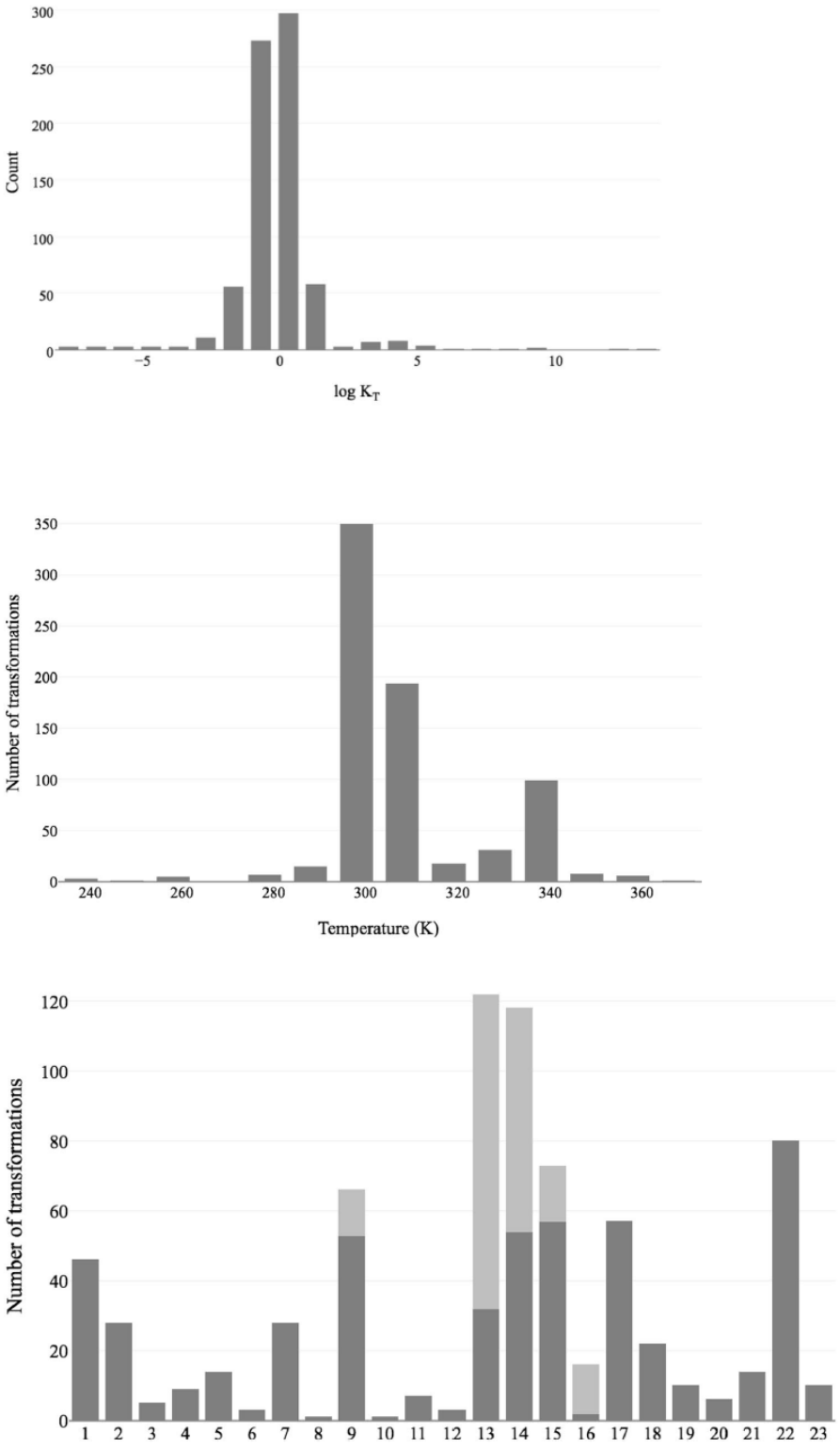
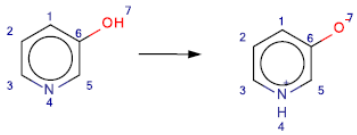
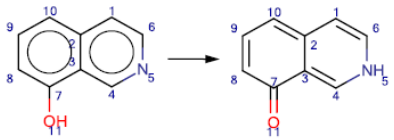
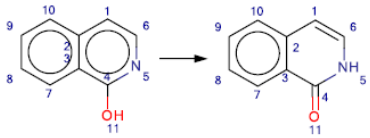


Fig. 2 (top) Distribution of $\log K_T$ values in studied equilibria. (middle) Distribution of temperature of $\log K_T$ measurements. (bottom) Distribution of different solvents and their mixtures with water. Dark grey and light grey correspond to pure solvents and their mixtures with water, respectively. Numbers on horizontal axis correspond to the following solvents: tetrachloromethane (1), nitrobenzene (2), ethyl ether (3), toluene (4), acetonitrile (5), 2,2,4-trimethylpentane (6), acetone (7), nitromethane (8), DMSO (9), oxolane (10), 1,2-dichloroethane (11), DMFA (12), toluene (13), 1,4-dioxane (14), methanol (15), propan-1-ol (16), bromobenzene (17), chloroform (18), cyclohexane (19), ethane-1,2-diol(20), pyridine (21), water (22), benzene(23)

Table 2 Variation of $\log K_T$ values extracted from different sources. All measurements were performed in water at 20 °C

Transformation	$\log K_T$	Reference
	−0.08	[41]
	0.1	[42]
	0.38	[41]
	−0.06	[42]
	4.26	[43]
	4.85	[41]

functions depending on the refractive index measured by D-line of sodium spectra at 20 degrees Celsius n_D^{20} [$g_1 = (n^2-1)/(n^2+2)$, $g_2 = (n^2-1)/(2n^2+1)$, $h = (n^2-1)(\varepsilon-1)/(2n^2+1)(2\varepsilon+1)$]. Remaining 7 descriptors reflect polarity and polarizability of the bulk of solvent. In order to describe water-organic solvent mixtures descriptor specifying the molar ratio of organic solvent in water was added. The inverse absolute temperature, 1/T (in Kelvin degrees) was also used as descriptor.

Building and validation of models

The modeling was performed using support vector regression (SVR) method implemented in the libSVM program [61]. The optimal method parameters and descriptors types were selected using the genetic algorithm driven SVR optimizer tool [62]. The models performances were estimated in fivefold cross-validation procedure repeated 10 times after the data reshuffling (10 × 5-CV). Ten best models corresponding to different descriptors types and/or hyperparameters of SVR were selected according to determination coefficient Q^2 (see Table S1 in Supporting Information). Selected models (Table 3) were then applied to assess the equilibrium constants for the tautomers from the external test sets. Thus for each equilibrium, 10 $\log K_T$ values were obtained followed by calculation of their arithmetic average. In such a way, ensemble of selected models forms the consensus model (CM).

In the calculations on the test sets, a combination of fragment control and bounding box approaches [27, 63] were used as an applicability domain for each individual model.

Predictive performance of models was evaluated in cross-validation and on the external test set. As performance metrics, determination coefficient (for CV procedure denoted as Q^2 and for external test set as R^2), and root mean squared error (*RMSE*) were used.

$$Q^2\ (or\ R^2) = \frac{1 - \sum_{i=1}^{n} (Y_{\exp,\ i} - Y_{pred,\ i})^2}{\sum_{i=1}^{n} (Y_{\exp,\ i} - <Y>_{\exp})^2}$$

$$RMSE = \left[\frac{1}{n \sum_{i=1}^{n} (Y_{\exp,\ i} - Y_{pred,\ i})^2} \right]^{1/2}$$

Here Y_{exp} and Y_{pred} are, respectively, experimental and predicted values of the equilibrium constant $\log K_T$, n is the number of data points.

The success rate of major tautomer predictions (MT) was also estimated.

DFT calculations

Identification of the lowest energy tautomers was performed using *in-house* Perl-based script in several steps. First, all conformers (max 1000) for each reagent and each product were generated by means ChemAxon's JChem Calculator plugin [14]. Then, they were used as input in semi-empirical calculations using PM6 hamiltonian [64] implemented into the MOPAC program [65]. Ten lowest energy structures for each tautomer were then re-optimized using DFT [66] with PBE functional [67] and 3z basis set implemented into the Priroda 11 program [68]. At the second step, Gibbs free energies of the structures in a given solvent and at a given temperature were calculated using the DFT B3LYP/6-311 ++ G(d,p) [69] method implemented in Gaussian09 program, version C.01 [70]. Solvent effects were assessed using IEF-PCM formalism [71] of Tomasi's polarizable continuum model [72]. The SMD parameters [73] for non-electrostatic part of solvation energy were used. Equilibria occurred in solvent mixtures were not considered. Geometry of each structure was fully optimized with normal optimization criteria followed by Hessian calculation to be sure that true local minimum of potential energy surface was localized. Thermochemical corrections to a given temperature were used to obtain Gibbs free energy of formation. The pressure was set to 1 atmosphere. Finally, the $\log K_T$ value has been calculated according to Eq. 4.

The above workflow, unfortunately, doesn't guarantee generation of conformers with intramolecular hydrogen bond which may significantly affect resulting free energies, and, hence, estimated $\log K_T$ values. In order to avoid this problem, all selected conformers of a given tautomer were visually inspected. If a formation of intramolecular

Table 3 Outliers detected on Fig. 4

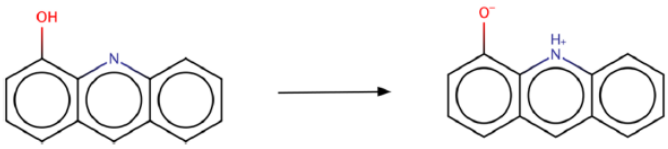
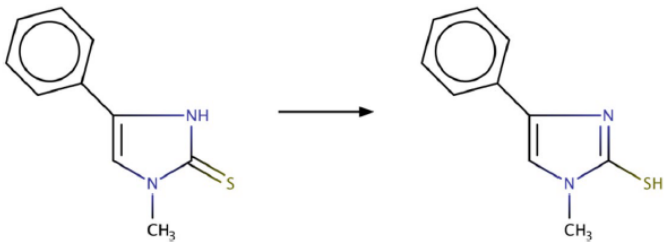
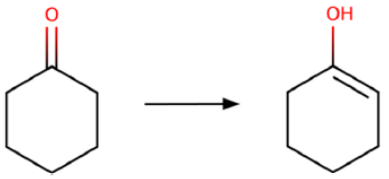
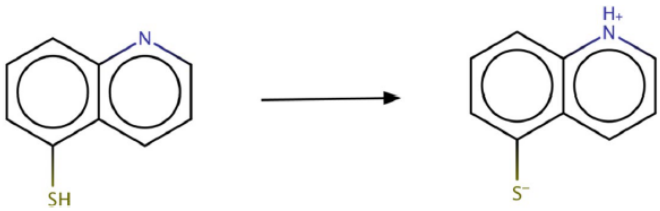
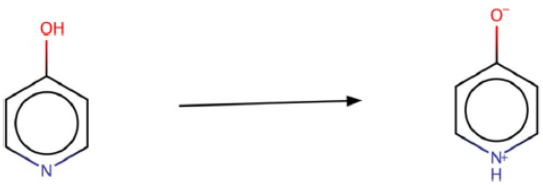
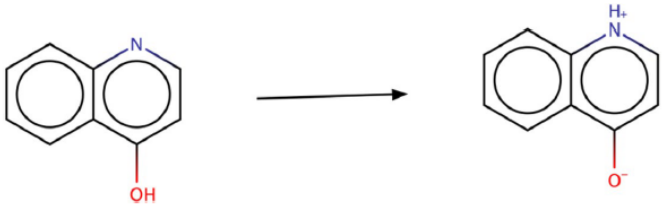
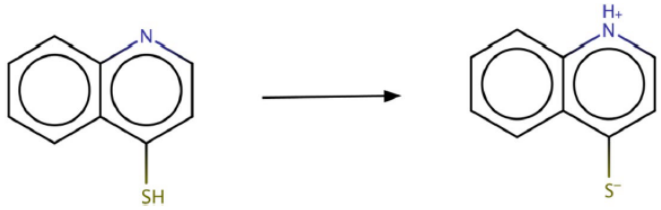
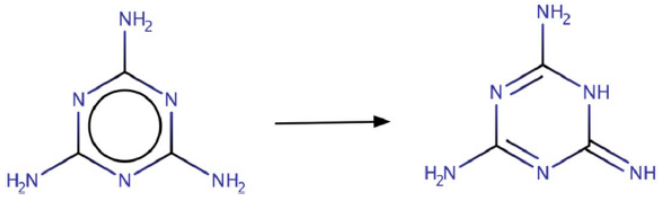
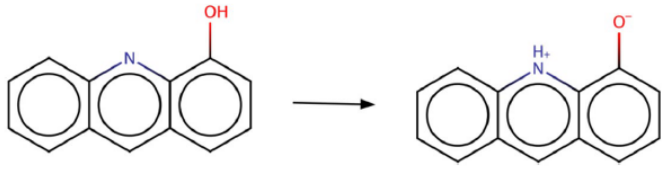
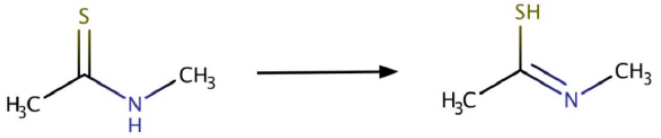
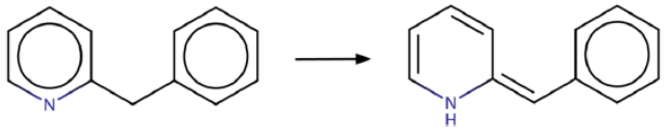
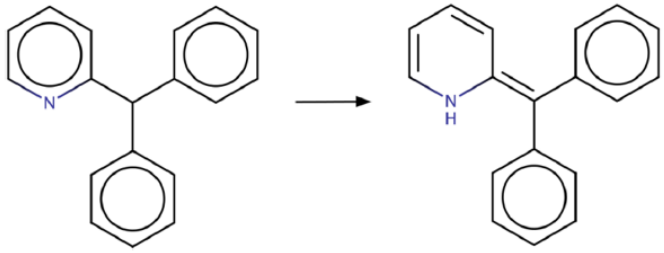
No	Tautomeric transformation	Predicted logK_T	Experimental logK_T
1		1.99	− 0.48 [42]
2		− 1.82	0.4 [42]
3		− 5.21	− 7.6 [74]
4		− 2.38	− 5.4 [75]
5		− 2.45	− 4.9 [75]
6		4.22	1.18 [76]
7		0.93	3.34 [76]
8		0.55	4.38 [76]

Table 3 (continued)

No	Tautomeric transformation	Predicted logK_T	Experimental logK_T
9		2.68	5.04 [76]
10		1.68	5.3 [77]
11		0.32	7.0 [78]
12		4.71	8.3 [79]
13		3.65	9.6 [79]
14		7.15	11.88 [80]
15		9.16	12.82 [80]

hydrogen bond(s) was possible, related conformers were manually prepared and their free Gibbs energies were compared with that calculated for the best structure generated with a help of the above workflow.

ChemAxon calculations

ChemAxon Calculator plugin (script *cxcalc*) was used to assess relative populations of tautomers in water from which

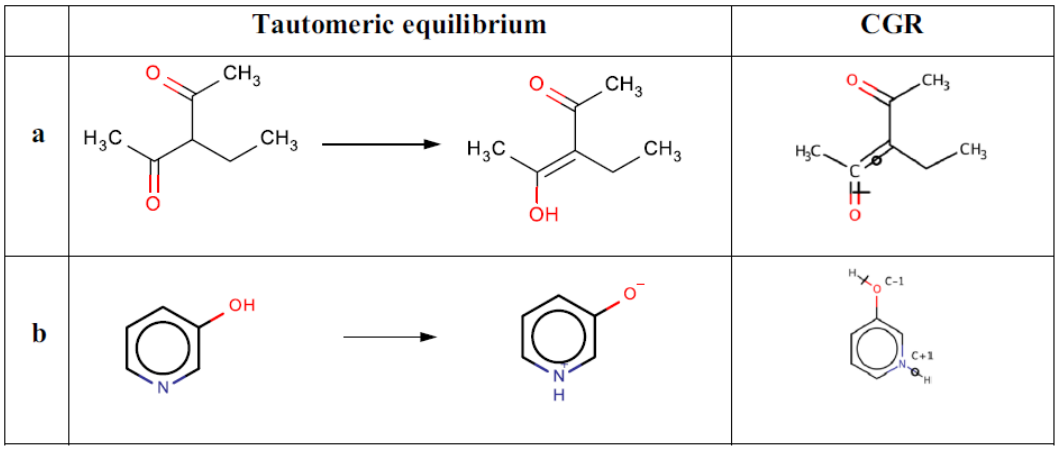


Fig. 3 Examples of tautomeric transformations and related Condensed Graphs of Reaction. The equilibrium (*a*) is described by CGR with two dynamic bonds, one of which is formed (denoted by circle) and another is broken (denoted by a crossed line). A CGR corresponding to the equilibrium (*b*) has two dynamic atoms: oxygen (labeled by "c-1") and nitrogen (labeled by "c+1") which change their charge from 0 to −1 and from 0 to +1, respectively

logK_T can easily be calculated (see Eq. 1). Both reagent and product were used as starting structures for calculation of tautomers populations. Tautomerisation involving atoms separated by no more than eight bonds was allowed. If predicted population of a given tautomer was found exactly 0 or 100%, logK_T tends to infinity and RMSE value can't be calculated. In this case only success rate of major MT was estimated.

Results and discussion

Consensus SVR model

Consensus model performs well in cross-validation calculations performed both on the entire set (RMSE = 0.7 logK_T units, $Q^2 = 0.81$, MT = 82%) and on the subset of unique transformations (RMSE = 0.96, $Q^2 = 0.83$, MT = 82%). Notice the RMSE values are close to the noise in experimental data (some 0.6 log units).

Equilibria for which deviations of predicted logK_T values from the experimental ones were larger than 3 RMSE were considered as outliers (see Fig. 4; Table 3). Their close inspection allowed us to identify the reasons of such discrepancy.

- Equilibrium involves a tautomer possessing rarely occurred (≤ 2) structural patterns (items 3, 10, 12, 13 and 14 in Table 3). In this case, the number of instances of a given type in the training set is not sufficient to learn the contribution of these patterns.
- Equilibrium has the highest or the lowest logK_T values for a given type of tautomerism (4, 5, 13 and 15). When such an example is included in a test set at a given cross-validation fold, its equilibrium constants is outside of the range of logK_T for the corresponding training set. Thus, data extrapolation leads to greater errors.
- Experimental value of logK_T seems to be erroneous. Predicted logK_T = 0.32 for 4-acridinol in water at 20 °C (item 11, Table 3) is largely underestimated compared to experimental logK_T = 7.00 reported in [78]. However, reference [42] for this equilibrium in 8.91% ethanol–water solution at 20 °C reports logK_T = 0.4, which is pretty close to the predicted value. Obviously, the addition of a small amount of ethanol can hardly change the equilibrium constant by several orders of magnitude.
- Erroneous annotation of equilibrium type leading to wrong CGR structure. Thus, in the reference [38] the equilibria 1, 2, 6–9 were annotated as zwitterionic. However, de facto, 1, 2, 7 and 8 belong to the pyridole-pyridone type whereas 6 and 9 belong to the aminethiol-iminethiol type.

Detailed description of these outliers is provided in Table S2 in Supporting Information.

Consider the performance of the consensus model varies as a function of tautomerism type (Table 4). The model performs well on 7 out of 11 subsets ($Q^2 = 0.5$–0.93), but fails to predict logK_T for the azo-hydrazone, pyridol-pyridone, thione-enol–keto-thiol and classical form—zwitterion subsets ($Q^2 < 0.5$). This could be explained by both small subset sizes and their structural diversity. Although the pyridole-pyridone subset is relatively big

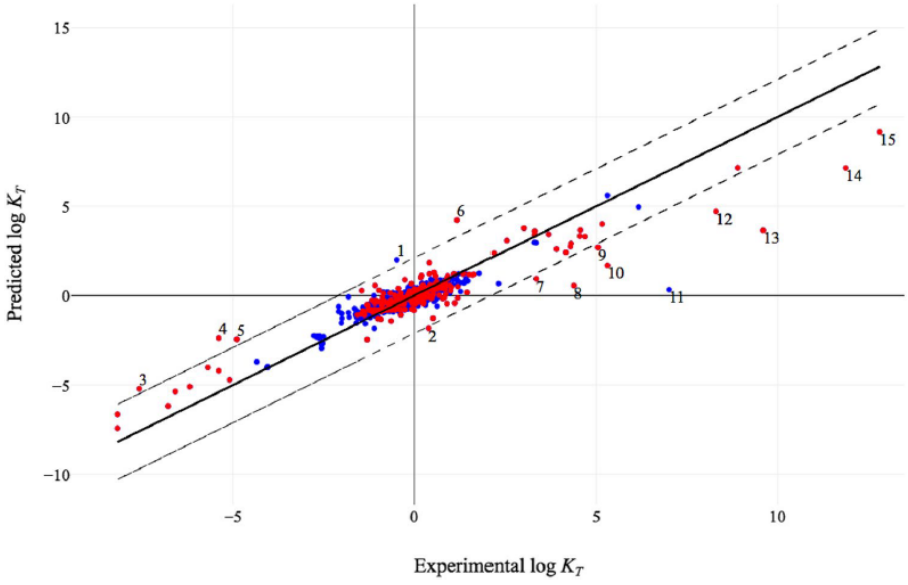


Fig. 4 Consensus model performance for logK_T achieved in 10×5-CV. Data points corresponding to reactions occurred only once (unique reactions) or several times are colored in red and blue, respectively. Dashed lines correspond to 3* RMSE deviation from ideal model (solid line). The outliers for which deviation from the experiment exceeds 3* RMSE are listed in Table 3

Table 4 Performances of local and global models for particular tautomeric types

Type of tautomerism	*N*	Global model			Local models		
		RMSE	Q^2	MT (%)	RMSE	Q^2	MT (%)
Keto-enol	288	0.49	0.65	81	0.39	0.77	84
Amino-imino	190	0.49	0.82	84	0.32	0.92	92
Azo-hydrazone	3	0.86	−0.85	66			
Pyridol-pyridone	44	1.31	0.46	80	1.60	0.40	86
Pyridinoid-pyridonoid	4	3	0.75	100			
Phenol-imine–keto-amine	33	0.26	0.73	75			
Thione-enol–keto-thiol	10	0.26	−0.03	70			
Amine-thione–imine-thiol	21	1.82	0.93	76			
Nitro-aci	8	0.77	0.5	100			
Classical form—Zwitterion	18	1.6	0.4	83			
Ring-chain	120	0.34	0.73	86	0.34	0.74	83

(44 instances), it contains two subclasses corresponding to products with and without charge separation. Related CGR structures and, hence, generated fragment descriptors for these subclasses differ. This implicitly reduces the subset size used for the model development.

In parallel to the "global" model built on the entire dataset, some "local" consensus models each corresponding to particular type of tautomeric equilibrium have also been obtained. Only subsets containing more than 40 transformations were considered: keto-enol, amino-imino, pyridol-pyridone, and ring-chain tautomerism. The local models perform similarly (for ring-chain) or slightly better (for keto-enol and amino-imino) than the global one (Table 4). For pyridol-pyridone the local model is not satisfactory ($Q^2 = 0.40$) similarly to the global one.

External validation of SVR consensus model

The developed consensus model was validated on two external sets—TEST1, containing transformations occurred in the training set but under different conditions, and TEST2 containing only unique transformations not present in the training set (Fig. 5). Statistical parameters obtained for TEST1 were close to those observed in cross-validation: RMSE = 0.66 log units, $R^2 = 0.55$ and MT = 70%. However, the model failed to predict TEST2 transformations: RMSE = 1.63, $R^2 = 0.74$ and MT = 58%. Twelve equilibria in TEST2 were found outside the model applicability domain; their discarding lead to significant improvement of the statistical parameters: RMSE = 1.3 and $R^2 = 0.79$.

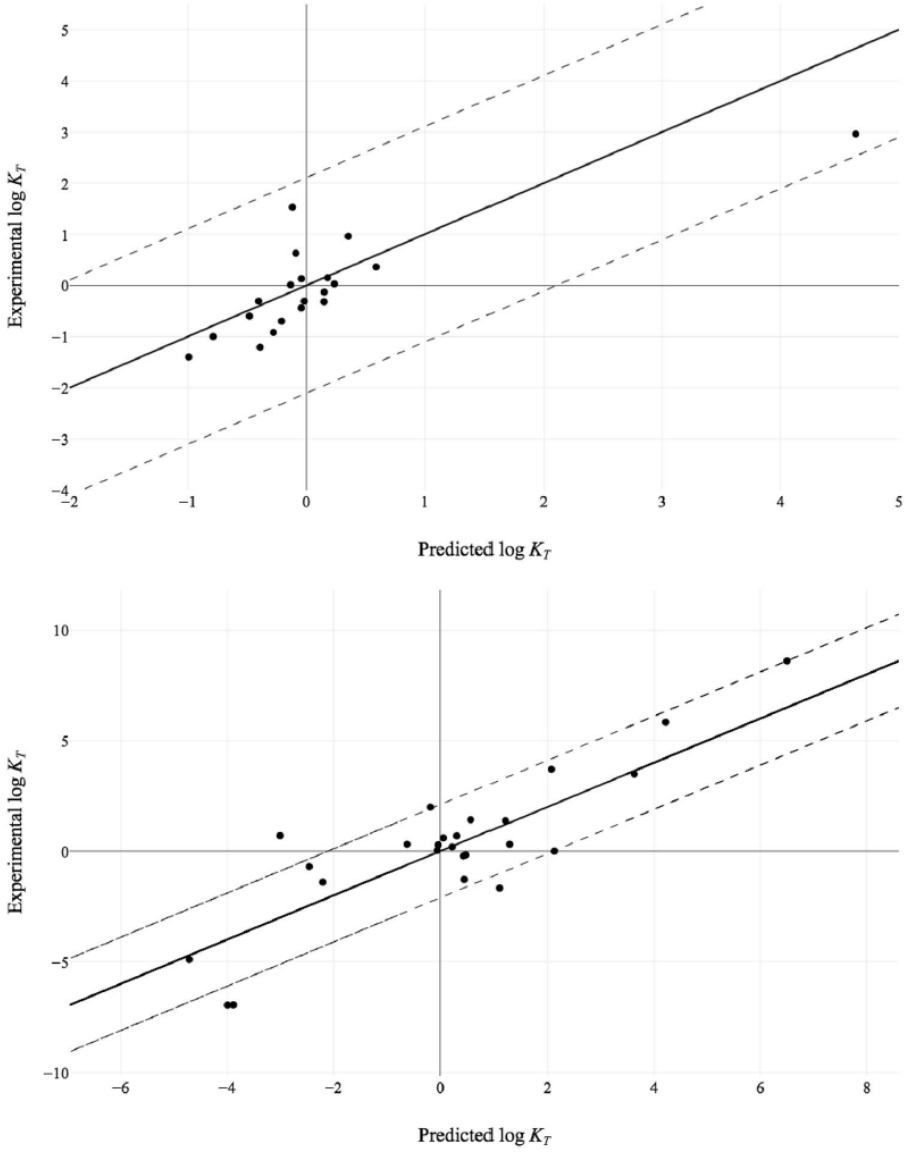


Fig. 5 Consensus model performance for logK_T for TEST1(top) and TEST2 (bottom). Dashed lines correspond to 3* RMSE deviation from ideal model (solid line)

Benchmarking studies

For the comparison purpose, performance of SVR models have been compared with that of DFT calculations applied to TEST1 and TEST2 instances, see computational workflow in the Method section. The logK_T values predicted in DFT calculations for TEST1 and TEST2 transformations are provided in Tables S5 and S6 in Supporting Information. Results given in Table 5 clearly demonstrate better performance of SVR model to assess logK_T.

Another comparison was performed with the results obtained with ChemAxon Tautomerizer plugin. This popular tool enumerates for a given query all its tautomers and predicts their relative ratio in water at ambient temperature.

Table 5 Comparison of the predictive performance of SVR models and DFT calculations

Method	Dataset	*N*	RMSE	R^2	MT (%)
DFT	TEST1	20	1.1	−0.3	65
	TEST2	26	3.00	0.13	54
SVR	TEST1	20	0.66	0.55	70
	TEST2	26	1.63	0.74	58

The number of data points (*N*), determination coefficients (R^2) and root-mean squared errors (RMSE in log*K* units) and success rate of major tautomer prediction (MT, %)

A subset of 57 tautomeric transformations in water has been selected from the training set. For each transformation, reactants and products were used as queries for enumerating tautomers. Equilibrium constant was calculated using Eq. 1. Calculations revealed two distinct subsets of instances. The first one (WATER1) is composed of 32 transformations for which the reagent/product ratio varied between 0% and 100% and remained practically the same for *query* = reactant and *query* = product. Another subset (WATER2) contained 25 instances which were not recognized by ChemAxon as species in equilibrium and, therefore, the reagent/product tautomer population was equal to either 0 or 100% (see Tables S3 and S4 in Supporting Information). For these instances, $\log K_T$ could not be assessed and only MT was estimated. Results given in Table 6 show that predictive performance of SVR models is much better than that of ChemAxon.

Conclusions

In this study, we have built for the first time direct QSPR models of tautomeric equilibrium constant without intermediate calculation of acidity (basicity) constants or free energies of individual tautomers. Representing each tautomeric transformation with a single Condensed Graph of Reaction significantly simplifies their structural complexity which, in turn, facilitates descriptors calculations. As a complement to structural descriptors, we also used descriptors of reaction conditions. This allowed us to build the consensus SVR models predicting $\log K_T$ for tautomeric transformations in different solvents and at different temperatures. For the dataset containing 739 tautomeric transformations of 11 tautomerism types, reasonable prediction performance was observed in cross-validation: RMSE = 0.7 $\log K_T$ units which is similar to experimental noise estimated as 0.6 $\log K_T$ units. For some particular transformations (phenol-imine–keto-amine, thione-enol–keto-thiol and ring-chain) error of predictions drops to about 0.3 $\log K_T$ units. We have demonstrated that our model outperforms the results obtained with DFT B3LYP/6-311++G(d,p) and ChemAxon Tautomerizer.

Table 6 Comparison of the predictive performance of SVR models and ChemAxon calculations

Method	Dataset	N_{eq}	RMSE	R^2	MT (%)
ChemAxon	WATER1	32	5.4	−0.14	84
	WATER2	25	–[a]	–[a]	28
SVR	WATER1[b]	32	1.83	0.87	97
	WATER2[b]	25	1.87	0.76	76

See footnote for Table 5

[a]RMSE and Q^2 could not be calculated (see text)

[b]Cross-validated predictions were used

The consensus model is available for the users *via* the Web server http://cimm.kpfu.ru. It should be noted that this model should not be applied to heterocyclic compounds, because the training set extracted from the Palm's handbook [38] does not contain them.

Last but not least: in this work, only equilibria involving two tautomeric forms were considered. We believe, however, that our approach can be extended to compounds possessing any number of tautomers whenever reliable information about their populations in solution is available.

Acknowledgements This study was supported by Russian Science Foundation, Grant No. 14-43-00024. TG thanks the IDEX UniStra program for the fellowship.

References

1. Greenwood JR, Calkins D, Sullivan AP, Shelley JC (2010) Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. J Comput Aided Mol Des 24:591–604. https://doi.org/10.1007/s10822-010-9349-1
2. Clark T (2010) Tautomers and reference 3D-structures: the orphans of in silico drug design. J Comput Aided Mol Des 24:605–611. https://doi.org/10.1007/s10822-010-9342-8
3. Pospisil P, Ballmer P, Scapozza L, Folkers G (2003) Tautomerism in computer-aided drug design. J Recept Signal Transduct Res 23:361–371. https://doi.org/10.1081/RRS-120026975
4. Oellien F, Cramer J, Beyer C et al (2006) The impact of tautomer forms on pharmacophore-based virtual screening. J Chem Inf Model 46:2342–2354. https://doi.org/10.1021/ci060109b
5. Martin Y (2009) Let's not forget tautomers. J Comput Aided Mol Des 23:693–704. https://doi.org/10.1007/s10822-009-9303-2
6. Warr W (2010) Tautomerism in chemical information management systems. J Comput Aided Mol Des 24:497–520. https://doi.org/10.1007/s10822-010-9338-4
7. Sayle RA (2010) So you think you understand tautomerism? J Comput Aided Mol Des 24:485–496
8. Sitzmann M, Ihlenfeldt W-D, Nicklaus MC (2010) Tautomerism in large databases. J Comput Aided Mol Des 24:521–551. https://doi.org/10.1007/s10822-010-9346-4
9. Guasch L, Sitzmann M, Nicklaus MC (2014) Enumeration of ring-chain tautomers based on SMIRKS rules. J Chem Inf Model 54:2423–2432. https://doi.org/10.1021/ci500363p
10. Szegezdi J, Csizmadia F (2007) Tautomer generation. pKa based dominance conditions for generating dominant tautomers. 234th National Meeting of the ACS, Boston, MA, 19–23 August 2007
11. Trepalin SV, Skorenko AV, Balakin KV et al (2003) Advanced exact structure searching in large databases of chemical compounds. J Chem Inf Comput Sci 43:852–860. https://doi.org/10.1021/ci025582d
12. Guasch L, Yapamudiyansel W, Peach ML et al (2016) Experimental and chemoinformatics study of tautomerism in a database of commercially available screening samples. J Chem Inf Model 56:2149–2161. https://doi.org/10.1021/acs.jcim.6b00338
13. Advanced Chemistry Development Inc (2015) ACD/Tautomers
14. ChemAxon JChem Calculator Plugins 15.8.3
15. Molecular Networks GmbH Computerchemie. MN Tautomer
16. Schrödinger LLC LigPrep tautomeriser
17. Xemistry GmbH. CACTVS,
18. OpenEye Scientific Software. QUACPAC

19. BIOVIA. BIOVIA Pipeline Pilot
20. Harańczyk M, Gutowski M (2007) Quantum mechanical energy-based screening of combinatorially generated library of tautomers. TauTGen: a tautomer generator program. J Chem Inf Model 47:686–694. https://doi.org/10.1021/ci6002703
21. Kochev NT, Paskaleva VH, Jeliazkova N (2013) Ambit-tautomer: an open source tool for tautomer generation. Mol Inform 32:481–504. https://doi.org/10.1002/minf.201200133
22. Garcia-Viloca M, Alhambra C, Truhlar DG, Gao J (2003) Hydride transfer catalyzed by xylose isomerase: mechanism and quantum effects. J Comput Chem 24:177–190
23. Stigliani J-L, Arnaud P, Delaine T et al (2008) Binding of the tautomeric forms of isoniazid-NAD adducts to the active site of the Mycobacterium tuberculosis enoyl-ACP reductase (InhA): a theoretical approach. J Mol Graph Model 27:536–545. https://doi.org/10.1016/j.jmgm.2008.09.006
24. Todorov NP, Monthoux PH, Alberts IL (2006) The influence of variations of ligand protonation and tautomerism on proteinвЄ'ligand recognition and binding energy landscape. J Chem Inf Model 46:1134–1142. https://doi.org/10.1021/ci050071n
25. Rastelli G, Thomas B, Kollman PA, Santi DV (1995) Insight into the specificity of thymidylate synthase from molecular dynamics and free energy perturbation calculations. J Am Chem Soc 117:7213–7227. https://doi.org/10.1021/ja00132a022
26. Bonachéra F, Parent B, Barbosa F et al (2006) Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. J Chem Inf Model 46:2457–2477. https://doi.org/10.1021/ci6002416
27. 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. https://doi.org/10.2174/157340908785747465
28. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) ISIDA property-labelled fragment descriptors. Mol Inform 29:855–868. https://doi.org/10.1002/minf.201000099
29. Horvath D, Marcou G, Varnek A (2013) Do not hesitate to use tversky and other hints for successful active analogue searches with feature count descriptors. J Chem Inf Model 53:1543–1562. https://doi.org/10.1021/ci400106g
30. Ruggiu F, Gizzi P, Galzi J-L et al (2014) QSPR modelling—a valuable support in HTS quality control. Anal Chem. https://doi.org/10.1021/ac403544k
31. Brown JB, Okuno Y, Marcou G et al (2014) Computational chemogenomics: is it more than inductive transfer?. J Comput Aided Mol Des 28:597–618. https://doi.org/10.1007/s10822-014-9743-1
32. Cramer CJ, Truhlar DG (1999) Implicit solvation models: equilibria, structure, spectra, and dynamics. Chem Rev 99:2161–2200. https://doi.org/10.1021/cr960149m
33. Pliego JR, Riveros JM (2001) The cluster—continuum model for the calculation of the solvation free energy of ionic species. J Phys Chem A 105:7241–7247. https://doi.org/10.1021/jp004192w
34. Milletti F, Storchi L, Sforna G et al (2009) Tautomer enumeration and stability prediction for virtual screening on large chemical databases. J Chem Inf Model 49:68–75. https://doi.org/10.1021/ci800340j
35. Soteras I, Orozco M, Luque FJ (2010) Performance of the IEF-MST solvation continuum model in the SAMPL2 blind test prediction of hydration and tautomerization free energies. J Comput Aided Mol Des 24:281–291. https://doi.org/10.1007/s10822-010-9331-y
36. Nicholls A, Wlodek S, Grant JA (2010) SAMPL2 and continuum modeling. J Comput Aided Mol Des 24:293–306. https://doi.org/10.1007/s10822-010-9334-8
37. Ribeiro RF, Marenich AV, Cramer CJ, Truhlar DG (2010) Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models. J Comput Aided Mol Des 24:317–333. https://doi.org/10.1007/s10822-010-9333-9
38. Palm VA (1978) Tables of rate and equilibrium constants of heterolytic organic reactions. VINITI, Moscow
39. ChemAxon (2015) InstantJChem 15.7.27.0
40. ChemAxon (2015) Standardizer, JChem 15.8.3.0
41. Mason SF (1958) 131. The tautomerism of N-heteroaromatic hydroxy-compounds. Part III. Ionisation constants. J Chem Soc 674. https://doi.org/10.1039/jr9580000674
42. Mason SF (1957) The tautomerism of N-heteroaromatic hydroxy-compounds. Part II. Ultraviolet spectra. J Chem Soc 5010. https://doi.org/10.1039/jr9570005010
43. Albert A, Phillips JN (1956) 264. Ionization constants of heterocyclic substances. Part II. Hydroxy-derivatives of nitrogenous six-membered ring-compounds. J Chem Soc 1294. https://doi.org/10.1039/jr9560001294
44. Varnek A, Fourches D, Hoonakker F et al (2005) Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. J Comput Aided Mol Des 19:693–703. https://doi.org/10.1007/s10822-005-9008-0
45. 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
46. 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
47. 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. https://doi.org/10.1134/S0022476614060043
48. Madzhidov TII, Bodrov AVV, 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. https://doi.org/10.1134/S002247661507001X
49. 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. https://doi.org/10.15372/JSC20170402
50. 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. https://doi.org/10.1021/acs.jcim.6b00319
51. 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
52. 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 Comput Sci 32:244–255. https://doi.org/10.1021/ci00007a012
53. ChemAxon (2013) Standardizer, JChem 6.0.0
54. EPAM Systems (2015) Indigo
55. Madzhidov TI, Nugmanov RI, Gimadiev TR et al (2015) Consensus approach to atom-to-atom mapping in chemical reactions. Butlerov Commun 44:170–176
56. 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. https://doi.org/10.1002/jlac.199519950234
57. 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 1997:1941–1949. https://doi.org/10.1002/jlac.199719970921
58. Kamlet MJ, Taft RW (1976) The solvatochromic comparison method. I. The.beta.-scale of solvent hydrogen-bond

acceptor (HBA) basicities. J Am Chem Soc 98:377–383. https://doi.org/10.1021/ja00418a009
59. Taft RW, Kamlet MJ (1976) The solvatochromic comparison method. 2. The.alpha.-scale of solvent hydrogen-bond donor (HBD) acidities. J Am Chem Soc 98:2886–2894. https://doi.org/10.1021/ja00426a036
60. Kamlet MJ, Abboud JL, Taft RW (1977) The solvatochromic comparison method. 6. The.pi.* scale of solvent polarities. J Am Chem Soc 99:6027–6038. https://doi.org/10.1021/ja00460a031
61. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(27):1–27:27. https://doi.org/10.1145/1961189.1961199
62. Horvath D, Brown J, Marcou G, Varnek A (2014) An evolutionary optimizer of libsvm models. Challenges 5:450–472. https://doi.org/10.3390/challe5020450
63. Mathea M, Klingspohn W, Baumann K (2016) Chemoinformatic classification methods and their applicability domain. Mol Inform 35:160–180. https://doi.org/10.1002/minf.201501019
64. Stewart J (2007) Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. J Mol Model 13:1173–1213. https://doi.org/10.1007/s00894-007-0233-4
65. Stewart JJP (2008) MOPAC2009
66. Laikov DN (1997) Fast evaluation of density functional exchange-correlation terms using the expansion of the electron density in auxiliary basis sets. Chem Phys Lett 281:151–156
67. Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77:3865–3868
68. Laikov DN, Ustynyuk YA (2005) PRIRODA-04: a quantum-chemical program suite. New possibilities in the study of molecular systems with the application of parallel computing. Russ Chem Bull 54:820–826
69. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648–5652. https://doi.org/10.1063/1.464913
70. Frisch MJ, Trucks GW, Schlegel HB et al (2009) Gaussian 09 Revision C.01, Gaussian Inc., Wallingford
71. Tomasi J, Mennucci B, Cances E (1999) The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level. J Mol Struct Theochem 464:211–226
72. Tomasi J, Mennucci B, Cammi R (2005) Quantum mechanical continuum solvation models. Chem Rev 105:2999–3094. https://doi.org/10.1021/cr9904009
73. Marenich AV, Cramer CJ, Truhlar DG (2009) Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J Phys Chem B 113:6378–6396. https://doi.org/10.1021/jp810292n
74. Kjellin G, Sandström J, Willadsen T et al (1969) Tautomeric cyclic thiones. Part IV. The thione-thiol equilibrium in some azoline-2-thiones. Acta Chem Scand 23:2888–2899. https://doi.org/10.3891/acta.chem.scand.23-2888
75. Bell RP, Smith PW (1966) The enol content and acidity of cyclopentanone, cyclohexanone, and acetone in aqueous solution. J Chem Soc B Phys Org 241. https://doi.org/10.1039/j29660000241
76. Albert A, Barlin GB (1959) Ionization constants of heterocyclic substances. Part III. Mercapto-derivatives of pyridine, quinoline, and isoquinoline. J Chem Soc 2384. https://doi.org/10.1039/jr9590002384
77. Angyl SJ, Angyal CL (1952) The tautomerism of N-hetero-aromatic amines. Part I. J Chem Soc 1461. https://doi.org/10.1039/jr9520001461
78. Albert A, Phillips JN (1956) Ionization constants of heterocyclic substances. Part II. Hydroxy-derivatives of nitrogenous six-membered ring-compounds. J Chem Soc 1294. https://doi.org/10.1039/jr9560001294
79. Kjellin G, Sandström J, Sæthre LJ et al (1973) The thione-thiol tautomerism in simple thioamides. Acta Chem Scand 27:209–217. https://doi.org/10.3891/acta.chem.scand.27-0209
80. Chua S-O, Cook MJ, Katritzky AR (1973) Tautomeric pyridines. Part XIV. The tautomerism of 2-benzyl-, 2-benzhydryl-, and 2-anilino-pyridine. J Chem Soc Perkin Trans 2:2111. https://doi.org/10.1039/p29730002111

Conclusive remarks

In this section we reported the first attempt to model tautomerisation equilibrium directly without calculation of acidities of individual tautomers. It was possible since each equilibrium was encoded by CGR. Logarithms of equilibrium constants were learned directly from the data by standard machine-learning techniques. Two external test sets were collected to prove the possibility of model to predict equilibrium constants in new reaction conditions or containing new structural moieties correspondingly. External validation and cross-validation demonstrated good performance of the model. The prediction quality was close to noise in experimental data.

Along with global model built on whole training set, local models built on particular type of tautomerisation were tested. In most of cases, local model lead to lower RMSE than the global model. However some tautomerism types are predicted much worse than others. This can be explained by small number of datapoints and by its structural diversity.

The comparison with other methods as ChemAxon Tautomerizer plugin or quantum chemistry calculations showed that developed approach has higher precision in prediction on external test.

The model was published on-line on the site http://cimm.kpfu.ru/predictor.

Summing up the results shown in the Chapters 5-8, we could conclude that the efficiency of CGR-based and solvent descriptors in modeling of reaction characteristics. Prediction errors were found at the level of the data noise. The developed approach was used to build models to predict rates of reaction of different classes: substitution (S_N2), elimination (E2), cycloaddition (Diels-Alder) reactions, which were built for the first time; as well as tautomeric equilibrium constant model utilizing new methodology of direct prediction without intermediate building of acidity models.

Chapter 9.

Modeling of reaction rates of some bioorthogonal reactions

The last modeling challenge was modeling of biorthogonal reactions. The project was initiated within a collaboration with Prof. Frederic Taran, CEA. The initial idea was to develop a model to predict reaction rate for dipolar cycloaddition of sydnones to strained alkynes. The previous Diels-Alder model was useless for this set, as it does not contain reactions of 1,3 dipolar cycloaddition. The training set provided by group of Prof. Taran contained 18 reaction. In this chapter we will describe the problems we faced and the solution proposed.

The ability to selectively form and break chemical bonds in chemically complex and uncontrollable biological media is a long-standing goal of chemists interested in modifying biological materials. Biorthogonal chemical reactions [222], which are reactions that do not interfere with biological processes, address precisely this challenge and therefore they are of major importance in the fields of chemical biology and biochemistry. To fulfill the requirements of bioorthogonality, reaction partners must be stable and inert towards the plethora of chemical functionalities found in living systems while reacting selectively, efficiently and rapidly with each other under physiological conditions with no or innocuous by-products.

Biorthogonal reactions play key roles in modern biochemistry [223, 224]. 1,3-dipolar cycloaddition reactions like copper-catalyzed azide–alkyne cycloaddition (CuAAC) are one of the most extensively studied and used biorthogonal transformations. But they were catalyzed by Cu salts, which resulted in a dramatic toxicity of this systems for living organisms [225]. New reaction systems based on strain promoted cycloadditions were explored [226]. The catalyst addition is not required for them but reaction have rather low rate [227].

Recently sydnones were shown to be a very promising class of biorthogonal 1,3-diene that can participate cyclization reaction with cyclic alkynes resulting in pyrazole formation [228]. This reaction proceeds in two step: cyclization with formation of reactive intermediate and retro-Diels-Alder reaction with CO_2 cleavage (Figure 44). Thermal cycloaddition of sydnones with alkynes require quite harsh conditions and proceeds with low regioselectivity [228] while under copper catalysis reaction proceed in mild conditions with high yield and selectivity [229–231]. Strained alkynes readily react

with sydnones and iminosydnones without catalyst [232–234]. Substituents in sydnone ring [6, 235] as well as alkyne nature [233] were shown to have great impact on the reaction rate switching it from moderate to ultra-fast. Very recently detailed stopped-flow kinetics study of reaction of some sydnones have been performed that supported two-stage mechanism and rates of cycloaddition and retro-Diels-Alder steps were measured [6]. It was shown that for fast reactions of fluorosydnones the second, CO_2 release step, was rate-limiting.

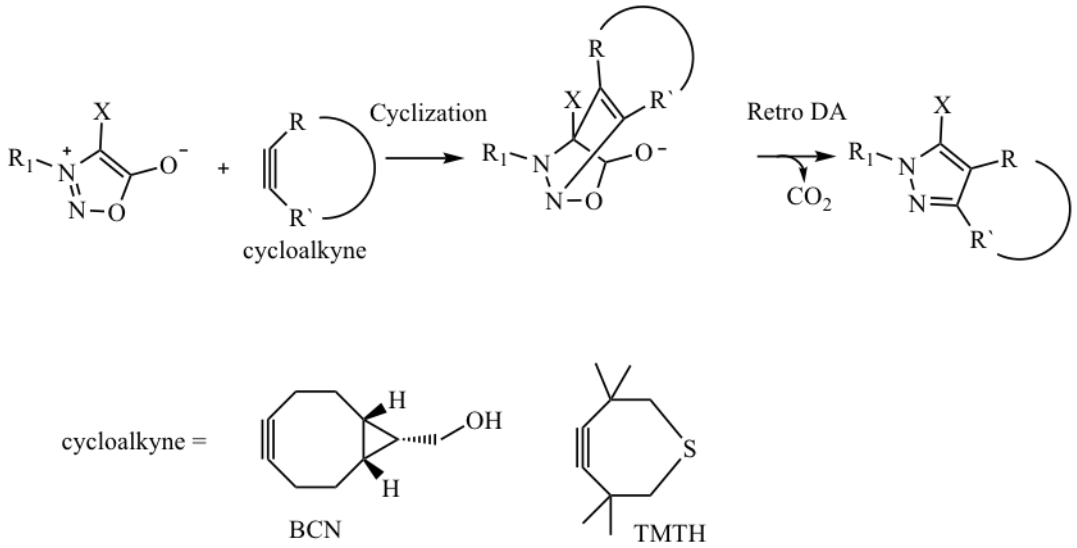


Figure 44. Studied reactions of sydnones with strained alkynes.

Murphy et al [233] performed a DFT study of N-phenyl sydnone and diaryl-1,2,4,5- tetrazine reaction with different strained alkenes and alkynes: norbornene, bicyclo-[6.1.0]-nonyne (BCN), 1,3- and 3,3- dimethylcyclopropene (Cp(1,3) and Cp(3,3)), difluorocyclooctyne (DIFO), trans-cyclooctene (TCO), biarylazacyclooctynone (BARAC) and dibenzoazacyclooctyne (DIBAC). A two steps mechanism of sydnone-alkyne reaction was computationally supported. Using experimental data on N-phenyl sydnone – BCN cycloaddition the rate of other reactions based on activation free energy was assessed. The reaction rate changes in several orders of magnitude – from 10^{-7} to 10 $M^{-1}s^{-1}$ – depending on the nature of dienophile (alkyne or alkene). For N-phenyl sydnone the most active reaction partners were BARAC and DIBAC due to large strain. However due to steric hindrance they have low reaction rates with diaryltetrazine. This observation was used to identify two mutually orthogonal reactions, i.e. two sets of alkyne-diene pair, namely sydnone-(BARAC or DIBAC) and diaryltetrazine-norbornene, that has almost no

cross reactivity. It was supported experimentally. Similarly, in other work of the same authors [236] on azide and dimethyltetrazine another mutually orthogonal pair was found. The article of Gordon et al. [237] was devoted to the design of the most reactive cyclooctynes for the click 1,3-dipolar cycloaddition with azides. They found that generally electronically depleted alkynes and steric hindrance in them improved the reaction rate, steric hindrance being the most important. The importance of steric strain of alkene or alkyne moieties to accelerate kinetics of 1,3-cycloaddition to organic azides or tetrazines could be explained by distortion/interaction model [238, 239]. Strained alkenes and azides are predistorted toward the Diels–Alder transition structure and thus interact easier.

Thus, by now, major attention was devoted to design of alkynes or alkenes with improved reactivity toward strain promoted 1,3-cycloaddition. In this part of the work we wanted to create predictive model for reaction rate of sydnones with alkynes and to shed light how substituents in sydnone ring influence its reactivity. We collected some experimantal data on sydnone-alkyne cycloaddition that was used for assessment of the error of computational prediction of reaction rate and revealing the importance of electronic and steric effects of substituents of sydnone ring for reaction rate.

Moreover, in [233], it was shown that the second step on reaction pathway of sydnone-strained alkyne cycloaddition has almost no barrier which contradicts with kinetic experiments of Liu et al [6]. However, in these two works different sydnones and alkynes were studied. Potentially, fast fluoro-N-phenylsydnone reaction with BCN might have completely different reaction path than medium-rate reaction of unsubstituted N-phenylsydnone. In this work we performed QSPR modeling and quantum chemical study of fluorosydnone reaction pathway.

9.1. Data set description

The data on cycloaddition reaction rate were taken from publications of F. Taran's group [229, 234], Figure 44. The collected dataset contains reaction rates at ambient temperatures of 15 reactions of sydnones and BCN, 3 reactions of sydnones with TMTH (see Table 7).

Table 7. Dataset of sydnone-alkyne strain-promoted cycloaddition reaction rate constants. R-group position are shown on Figure 44.

No.	**R**	**X**	**Cycloalkyne**	**Rate constant (tolerance), $M^{-1}*sec^{-1}$**	**Reference**
1	p-MeO C_6H_4	H	BCN	0.006 (0.001)	[235]
2	p-Me C_6H_4	H	BCN	0.032 (0.001)	[6]
3	C_6H_5	H	BCN	0.027 (0.002)	[235]
4	p-CO_2H C_6H_4	H	BCN	0.059 (0.001)	[235]
5	p-CF_3 C_6H_4	H	BCN	0.199 (0.002)	[235]
6	p-NO_2 C_6H_4	H	BCN	0.289 (0.012)	[235]
7	C_6H_5	CH_3	BCN	0.018 (0.002)	[235]
8	C_6H_5	C_6H_5	BCN	0.027 (0.001)	[235]
9	C_6H_5	CF_3	BCN	0.008 (0.001)	[235]
10	p-Me C_6H_4	Cl	BCN	0.872 (0.034)	[6]
11	p-Me C_6H_4	Br	BCN	0.592 (0.021)	[6]
12	p-Me C_6H_4	I	BCN	0.306 (0.008)	[6]
13	p-CO_2H C_6H_4	Br	BCN	0.798 (0.065)	[235]
14	p-CO_2H C_6H_4	Cl	BCN	1.593 (0.034)	[235]
15	p-Me C_6H_4	F	BCN	42	[6]
16	p-Me C_6H_4	F	TMTH	1500	[6]
17	p-F C_6H_4	F	TMTH	3500	[6]
18	p-CF_3 C_6H_4	F	TMTH	12000	[6]

From the table one can see that reactions with TMTH have several magnitudes higher speed than with BCN and their rate were measured only for the most reactive fluorosydnone. We removed from dataset 3 reactions, that have the same reaction rates in two articles, but have difference in structure (reactions 10,11,12 in [234] have CH_3 group in para position of phenyl group, but in previous article [229] they have not). So, out of six reactions in the article [229] we kept only 3.

9.2. QSRR modeling

In the beginning of the project the attempt to model rate constants using QSAR approach was done. To do it some new, confidential data on reaction rate with BCN was given to us by F. Taran's group. Totally we had 34 data on measurement of rate constants of different sydnones and iminosydnones with BCN. The only factor influencing the reaction rate was the chemical structure of the sydnones, since the alkyne and reaction conditions were kept constant. Thus, there is no need in application of CGR approach for the reaction rate prediction and standard QSPR approach could be used. The target property of the modeling was the rate constant logarithm. Various structural descriptors of different type were applied. The following combinations of descriptors/modeling procedure were explored:

- ISIDA fragment descriptors of different type with SVM machine learning method,
- Quantum-chemical CODESSA descriptors with Multiple Linear Regression using CODESSA Pro program [240, 241],
- Quantum-chemical descriptors based on Bader's Quantum Theory of Atoms in Molecules [242] with Multiple Linear Regression using CODESSA 3 program [243],
- Conceptual DFT theory [244] indices (see Chapter 4.3.3.4 for details) calculated for key atoms of sydnone ring on the basis of DFT/PBE/3z by Priroda [165] with Multiple Linear Regression.

However all approaches have shown poor performance ($Q^2<0.5$) on 10-fold cross-validation. We explain it by following factors:

- Small and heterogeneous data set - the number of points in the dataset is not enough to learn all the factors that influence on reaction. The data set appeared to be too heterogeneous, most reactions are slow and 4 had a much bigger reaction rate constant than most of the others in the dataset.
- Obvious feature that influence reaction speed – one can notice that reaction rate increases when halogen atom is present in X position. Thus upon descriptor selection usually presence of certain halogen becomes an important feature. However such model could not predict correctly test set reactions where a sydnone

contained another halogen atom. Such effect and small number of high-speed representatives makes impossible to learn other features from this dataset.

Our attempt to manually select descriptors according to knowledge of mechanism failed as well and we never managed to build a model with moderate performance. Thus we came to conclusion that the dataset has low modelability for QSPR approach application due to its composition. Thus, we turned toward quantum-chemical calculations to estimate the reaction rate.

The work was divided into two parts: development of the workflow for direct calculation of transition state Gibbs free energy with following reaction rate constant calculation and revealing the structural effects responsible for reaction rate. The tasks were solved using dataset given in Table 7.

9.3. Details of quantum chemical calculations

In this sections the computational approaches used for calculation will be described.

9.3.1. Energy and geometry optimization

Density Functional Theory (DFT) [118, 119] calculations were made in Priroda 11 program [165] with PBE exchange and correlation functional [126] and built-in triple-zeta split valence basis set (called 3z, equivalent of Schäfer's TZVP basis [167]). Relativistic effects were neglected since for molecules under study it has minor importance. Priroda11 is probably one of the fastest DFT code due to efficient evaluation of density functional exchange-correlation terms based on the expansion of the electron density over auxiliary basis set [166]. The program was used since computational efficiency was of major importance for the study.

Geometry optimization for reagents, intermediates and products as well as saddle point optimization were performed using built-in quasi-Newton-Raphson procedure and BFGS hessian update scheme. Scanning along plausible reaction coordinate was used to localize good structural guess for transition state. The scanning procedure was constrained local optimization with one coordinate N1…C_{alkyne} set externally. Its value changed from 2.1 to 3.1 with step size 0.1 Å. Geometry optimization was followed by frequency calculation to control correct structure of hessian: discussed structures of reagents, products and intermediate had no imaginary frequency, transition state

geometry had one large imaginary frequency. The correctness of transition states was also checked by intrinsic reaction coordinate (IRC) following approach.

For 3 reactions full reaction paths were calculated. First stage transition state (TS1) was localized and optimized using aforementioned procedure. The intermediate was found during the intrinsic reaction coordinate downhill movement, followed by local optimization. The intermediate geometry was used to localize a second stage transition state (TS2). The N1-O5 bond elongation was used as reaction coordinate, as it is the most sensitive to geometry changes occurring in the intermediate. For scan procedure step size was 0.01 Å, since TS2 is very structurally similar to intermediate. Geometry of TS2 was found by saddle point optimization of the highest energy structure obtained during the scanning procedure. IRC procedure supported that TS2 belong to reaction minimal energy pathway, descent from TS2 in one direction reproduces the intermediated while in the other direction it converges to the final product molecules.

Solvation free energy was calculated for some structures by IEF-PCM model [4, 169] with SMD parameters for non-electrostatic terms [170] using Gaussian program [171]. In this case geometry optimization and hessian calculation was performed in Gaussian program [171] too using PBEPBE/6-311++G(d,p) method.

9.3.2. Procedure for transition state detection

The workflow used for semi-automatic first transition state detection is schematically represented on Figure 45. The detailed descriptions of each step are given below.

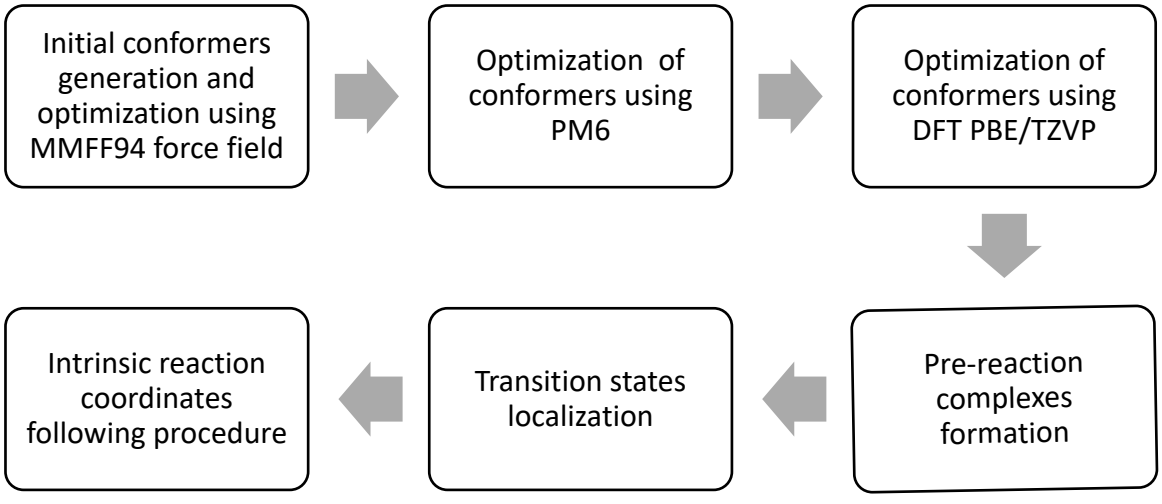


Figure 45. Schematic representation of the workflow for TS1 detection.

Initial conformers generation

All structures were drawn following the same numeration of sydnone ring atoms. ChemAxon Calculator Plugin *cxcalc* was used to generate up to 500 conformers with diversity limit 0.1 Å with subsequent optimization using MMFF94 force field.

PM6 optimization

All structures obtained from previous step were optimized using PM6 semi-empiric method using MOPAC 11 program [245]. Duplicated conformers (with RMSD <0.1 Å) were removed using simple Perl script that aligns molecules according to principal components of inertia with following RMSD calculation (developed at N.N.Vorozhtsov Institute of Organic Chemistry, Novosibirsk, Russia, available at http://limor1.nioch.nsc.ru/quant/program/conformers/conformers.html). Remaining conformers were ranked by energy and 100 most stable conformers were chosen for further calculations.

Reagent structure optimization using DFT

All structures obtained from PM6 optimization were optimized in PBE/3z in Priroda11. The lowest energy conformer was selected to continue with the next step. The selected conformer geometry was optimized until no frequency of the hessian was imaginary.

Pre-reaction complexes formation

There are 4 possible orientations of reagents to form TS1, that could be considered as 2 pairs of enantiomers, Figure 46. However when CO_2 is cleaved the product is formed as racemic mixture of two enantiomers.

Four possible orientations of reagents were generated using in-house Python script in a way that distance between reaction center atoms (C3…C1 and N1…C2 or C2…C1 and N1…C2, depending on orientation, atomic numbering is given on Figure 46) was 3.1 Å, i.e. much greater than in transition state. Guess for first transition state was found using scanning procedure in Priroda11 varying distance between reaction center atoms from 3.1 Å to 2.1 Å with step size 0.1 Å. For the structure with the lowest energy hessian was calculated, and saddle point optimization started.

Pair of enantiomeric structures of transition states should have the same energies. However, due to some fluctuations, structures of them obtained using 4 different orientations of reagents usually had slightly different energies. For further analysis the lowest energy transition state among four was selected. The selected structure hessian was calculated and the presence of a single large imaginary frequency was checked. The correctness of transition state was checked using Intrinsic Reactional Coordinate following (IRC) procedure.

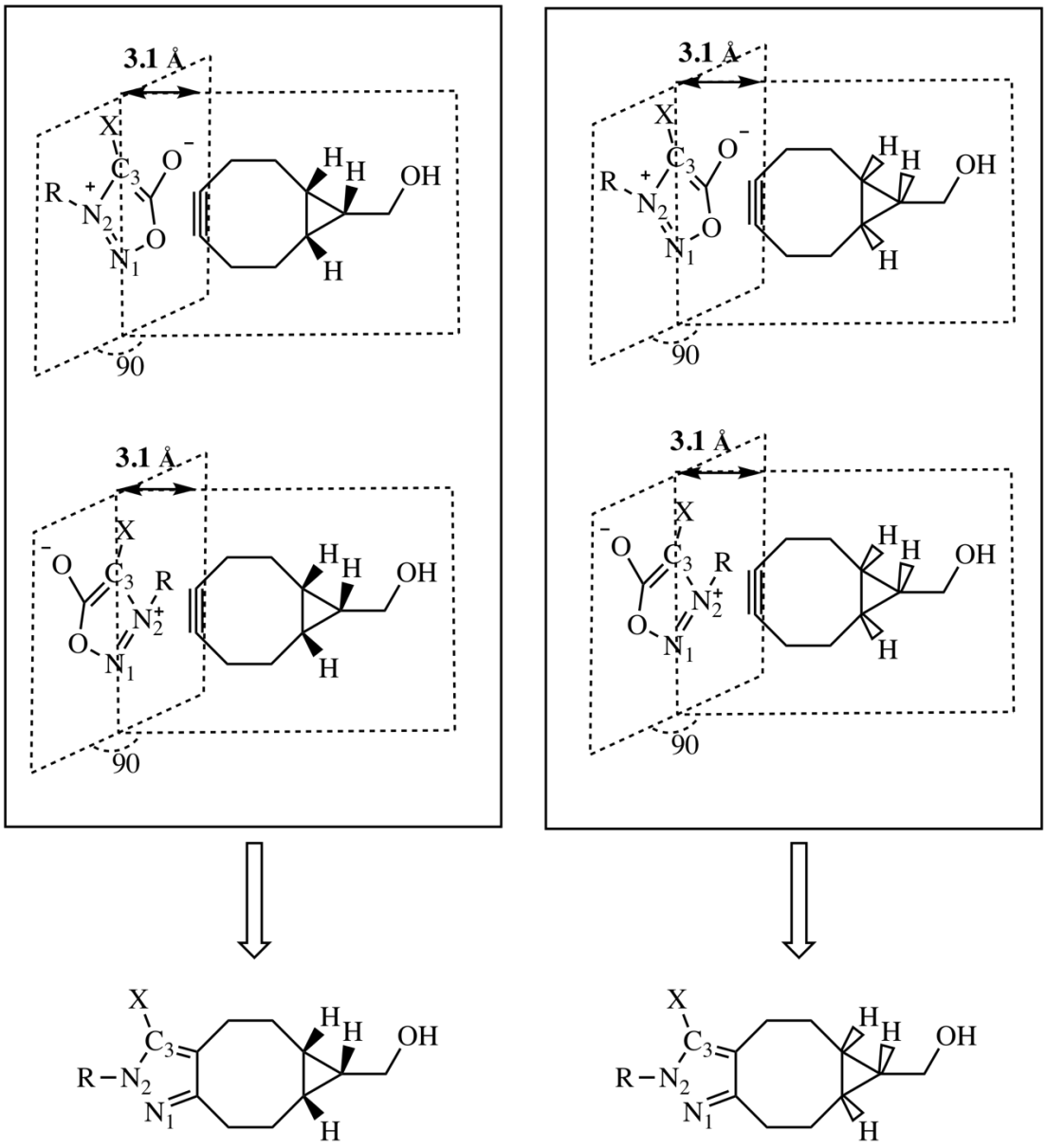


Figure 46. Four possible pre-reaction complexes diene and dienophile

9.3.3. Activation free energy calculation

Free energy of molecule and transition state formation was calculated using built-in thermochemical calculation in rigid ideal rotator and harmonic oscillator approximation at 298 K in Priroda11 program, including zero-point vibrational energy corrections [165].

Formula Eq. 6 was used for calculation of activation free energy on the basis of quantum chemical calculations:

Eq. 6 $$\Delta\Delta G^{\neq} = \Delta G_{298}^{TS} - (\Delta G_{298}^{sydnone} + \Delta G_{298}^{alkyne})$$

Calculation of activation free energy based on experimentally measured reaction rate was done using formula Eq. 7 based on Eyring equation of Transition state theory Eq. 8:

Eq. 7 $$k = \kappa \frac{K_B T}{h} e^{\frac{-\Delta\Delta G^{\neq}}{RT}}$$

Eq. 8 $$\Delta\Delta G^{\neq} = -RT\ln\left(\frac{k*h}{\kappa K_B T}\right)$$

where k - rate constant; κ – transmission coefficient (here, κ=1); T - Temperature in Kelvin; K_B -Boltzmann constant; $\Delta\Delta G^{\neq}$ - Difference of free energies of reagents and TS at 298.15 K; h - Planck constant; R - gas constant.

9.3.4. Conceptual DFT indices

Different Conceptual DFT [244] indices were used in the work to reveal structural factors responsible for reaction rate. Electrophilicity index, ω, measures the stabilization energy when system acquires an additional electronic charge [246]:

$$\omega = \frac{\mu^2}{2\eta}$$

where μ is electronic chemical potential, and η – chemical hardness. They could be expressed in terms of HOMO (ε_{HOMO}) and LUMO (ε_{LUMO}) energies as:

$$\mu = (\varepsilon_{HOMO} + \varepsilon_{LUMO})/2$$

$$\eta = \varepsilon_{LUMO} - \varepsilon_{HOMO}$$

The HOMO and LUMO energies were obtained within DFT scheme [118] for sydnone molecules.

Fukui indices are used to characterize atom ability to share/withdraw electronic charge. For calculation of Fukui nucleophilicity F_A^-, elecrophilicity F_A^+ and radical attack

susceptibility F_A^0 indices of atom A single point calculations of molecules with added and removed electron were done:

$$F_A^- = P_A(N) - P_A(N-1)$$
$$F_A^+ = P_A(N+1) - P_A(N)$$
$$F_A^0 = 0.5 * P_A(N+1) - P_A(N-1)$$

where $P_A(M)$ – Hirshfield charge on the atom A in molecule with M electrons, N - number of electrons in neutral molecule. Geometry of cation and anion-radical molecules were approximated to the one corresponding to the lowest energy structure of the neutral molecule.

9.4. Reaction pathway investigation

The full reaction path was explored for reactions **2, 15** and **16** in Table 7 whose kinetic measurements using stopped-flow technique was performed in reference [6]. The procedure included scanning along reaction coordinate, optimization of transition state, intrinsic coordinate following to unite transition state with reagents and products. Energetic profile of reaction is shown on Figure 47. N1-O5 length could serve a good approximation for reaction coordinate, it smoothly elongates along the reaction path: 1.379 Å in reagents, 1.414 Å in TS1, 1.546 Å in intermediate and 1.721 Å in TS2 of reaction 15.

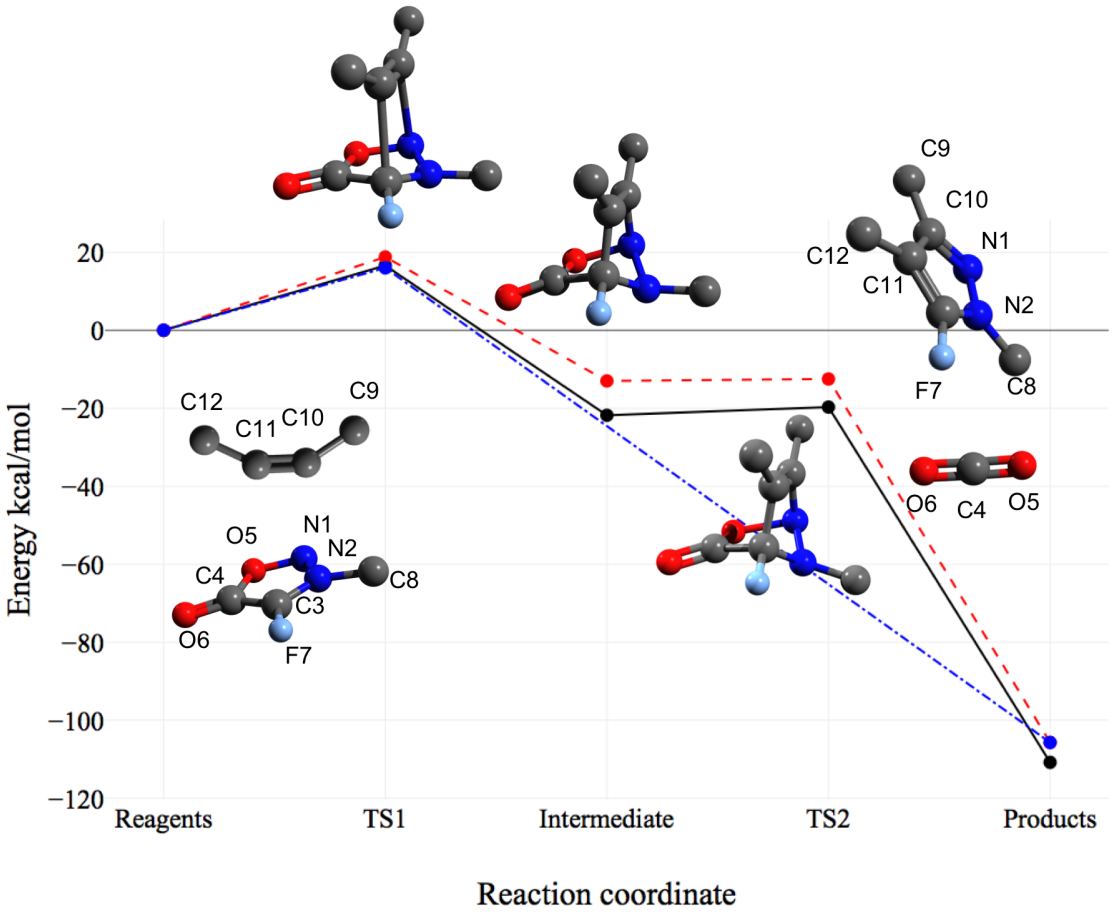


Figure 47. Reaction pathways for reaction **2** (dashed red line), **15** (solid black line) and **16** (dashed blue line). Relative free energies at 298K of molecules with respect to reagents are shown. Structure of reagents, transition states and intermediate for reaction **15** is shown. Substituent R of sydnone and almost all atoms of BCN are omitted for the sake of clarity. Bond orders correspond to molecule representation on Figure 44.

One can see from Figure 47 that the limiting step of reaction **15** is the first transition state (TS1), intermediate is lower in free energy than reagents by 21 kcal/mol. According to our calculation, intermediate is very unstable since it is separated from products by transition state (TS2) with a tiny barrier (some 1 kcal/mol). Intermediate decomposition is exergonic by some 85 kcal/mol. The same is true for the much slower reaction **2**. Solvent effect that was accounted by IEF-PCM model [4] shown almost no influence on the reaction barrier due to compensation. In reaction **16** between fluorosydnone and TMTH TS2 was not localized at all.

Liu at al [6] experimentally observed the existence of an instable intermediate during the reaction. However, they estimated the second step reaction rate in the range 0.02 to 0.98 s^{-1} depending on alkyne and concluded that this second step becomes rate

limiting. The discrepancy with the computed free energy clearly points toward yet undefined entropic effects that are balancing the two steps of the reaction. A possible hypothesis is that the solvent water molecules play an active role in the studied reactions.

With the present set of approximations, our results show that there is no qualitative difference between reaction paths for fast reaction of fluorosydnones (X=F) and much slower reaction of unsubstituted (X=H) sydnones: both are characterized by an unstable intermediate. This is also in qualitative agreements with the observations reported by Narayanam et al [233] for unsubstituted sydnones with strained alkynes.

9.5. Assessment of rate constant

The study of Liu et al [6] reported that the first step of the reaction is strongly affected by the nature of the sydnone. The main result is that the rate of the first step of the reaction is so fast, if it involves a fluorosydnones, that the second step became the limiting. This second step seems marginally affected by the nature of the sydnone and much more by the nature of the strained alkyne, ranging from 0.02 s^{-1} to 1 s^{-1}. However, these variations are order of magnitudes smaller than those that affect the reaction rate of the first step when varying the nature of the sydnone, ranging from 42 $Mole^{-1}.s^{-1}$ to 12000 $Mole^{-1}.s^{-1}$.

This explains the paradoxe that the effective rate of reaction (rate of product formation) seems to be mostly driven by the nature of the sydnone and the rate of the first step (see Table 7). The first step is so fast and the transformation is so irreversible that it produces a large accumulation of the instable intermediate. In turn, this produces a massive imbalance that pushes toward the generation of the final products. Moreover, according to description of HPLC experiment in author's previous studies [235] rate constants reported in Table 7 represent the speed of conversion of reagents. The latter depends only on first step rate constant.

Hence, from practical point of view the first step rate constant is the most relevant.

Experimental measurements of rate constants are time-consuming and expensive while the variability of structure is rather big. For practical application it is important to find a way to quantitatively assess the rate constants of some reaction. We attempted to find activation free energies for all reactions under study and compare predicted values with experimental ones. The goal was to assess the quality of quantum-chemical

estimation of reaction kinetic characteristics. For all reaction under study TS1 was identified and its activation energy was calculated as difference between free energy of TS1 and reagents. However, due to large number of reactions under study, conformational lability of interacting compounds and many possible orientations of reagents to form transition state, an automatic approach for the detection of TS1 was needed (described in details in Chapter 4.3.3.2). The approach includes (i) conformational sampling of initial compounds, (ii) subsequent elimination of local minimum geometries by semi-empiric and DFT calculations, (iii) pre-orientation of reagents to form a valid guess for TS1, (iv) energy screening along the reaction coordinate, (v) optimization of TS1, (vi) selection of lowest energy structure of TS1, (vii) manual examination of TS1 structure, correction and recalculation if required.

Free energies of activation, $\Delta\Delta G^{\neq}_{calc}$, were predicted for all reactions under study using described quantum chemical approach. For comparison, "experimental" values of free energies, $\Delta\Delta G^{\neq}_{exp}$, of activation were calculated using Transition State theory and Eyring equation (see Chapter 4.3.3.3) from measured rate constants. The comparison of these two quantities is given on Figure 48.

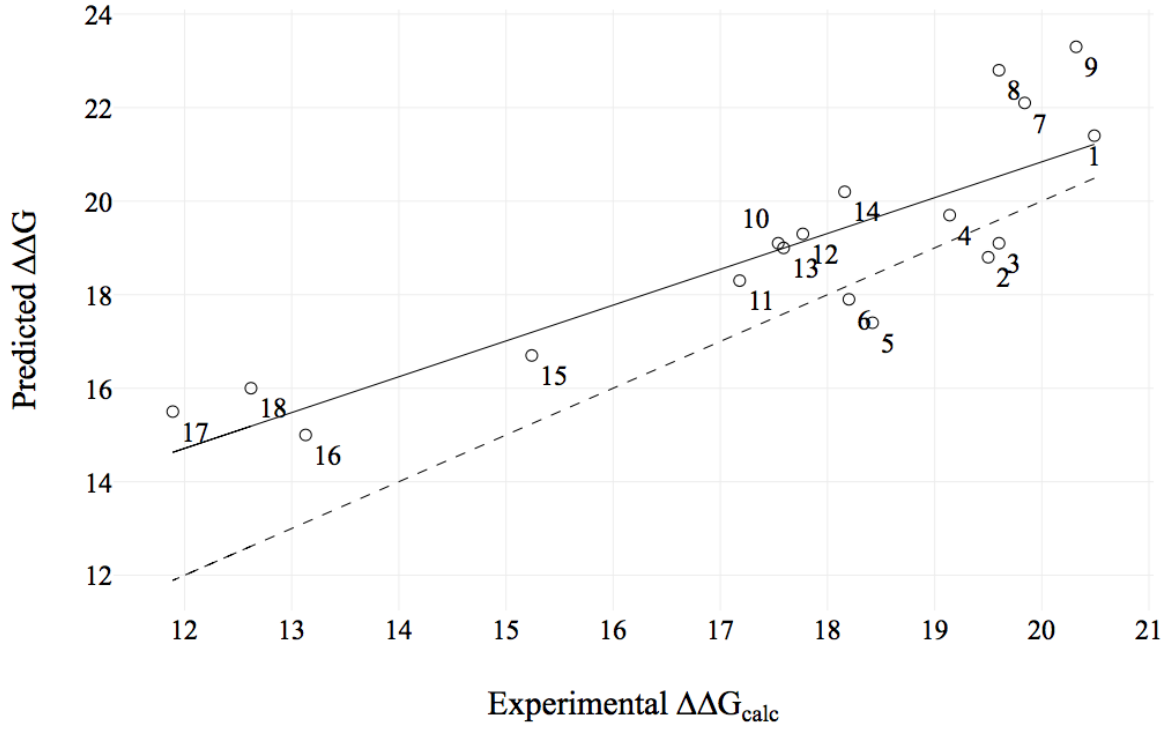


Figure 48. Free energies of activations predicted by quantum chemical calculations and calculated from experimental values using Eyring formula. Numbers corresponds to reactions in Table 7. Dashed line corresponds to perfect prediction, solid line represent linear correlation.

As one can see, quantum chemical calculations in gas phase well describe activation free energy of sydnones cycloaddition to strained alkynes. Root mean squared deviation between predicted and experimental values is 1.97 kcal/mol, the plot predicted vs experimental is shown on Figure 48. Considering that the accuracy of the free energy measures is about 1 kcal/mol, the agreement between the models and the experiments is reasonable and is comparable to what is expected using common DFT functional [11]. Moreover, the calculated value could contain error due to approximate nature of Transition State theory. The saddle point optimization is more error-prone than local minima geometry optimization. One can notice that some systematic overestimation of activation free energy takes place. Systematic error could be taken into account using linear correlation and in this case the error of prediction can be lower. The correlation coefficient, r=0.84 and RMSE (vs correlation line) = 1.53 kcal/mol. Thus, the developed workflow open perspectives for relatively fast and cheap (in comparison with experimental measurement) computational screening of reaction partners to identify the most reactive ones.

9.6 Structural factors responsible for reaction rate

Quantum chemical assessment of reaction rate based on the developed semi-automatic approach is reliable but computational resource-consuming task since the whole calculation could take some days per CPU core. The most complicated is the localization and optimization of transition state. For real application one needs to reduce search space and find factors responsible for high reactivity. Our attempt to build QSPR models failed since the resulting models have mediocre predictive ability according to cross-validation procedure.

Conceptual Density Functional Theory (DFT) indexes [244] are often used to analyze cycloaddition reactions [117, 247]. In order to reveal structural factors influencing on the rate constant we looked for correlations within the largest series of reactions of different sydnones with the same alkyne – BCN. However, for reaction under study neither sydnone electrophilicity index [246] nor chemical potential (calculated as half-sum of HOMO and LUMO energies) nor chemical hardness (calculated as HOMO – LUMO gap) correlates with reaction rate. Hereafter HOMO and LUMO energies are

obtained using DFT scheme, however our test has shown that the conclusions are not affected if Hartree-Fock calculations (HF/6-311+G**) are used instead. Atomic charges do not correlate with reaction rate as well.

Analysis of orbital symmetries and energies shows that electrons are transferred from HOMO of alkyne to LUMO of sydnone (Figure 49) in agreement with earlier publication by Liang et al [236] studied azide and tetrazine cycloaddition to alkynes. In the series of reactions with BCN there is no valid correlation of LUMO of sydnone with reaction rate, Figure 50a. Unlike the work of Domingo [117] no correlation with charge transfer was found. So even within this congeneric group there is no evidence that orbital interaction is the major factor affecting reaction rate.

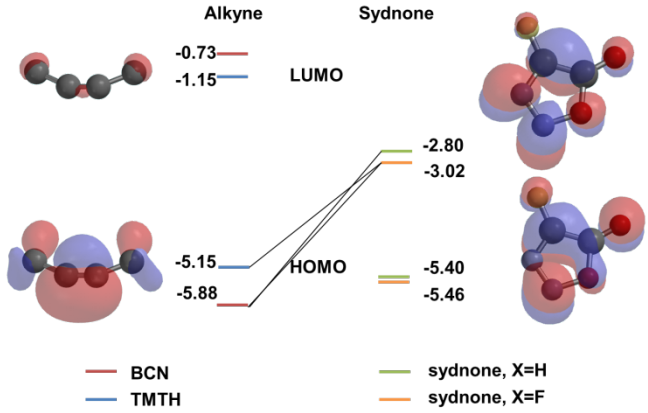


Figure 49. Comparison of HOMO and LUMO energies and symmetries for reactions **2**, **15**, **16**. HOMO and LUMO energies are obtained using DFT scheme on PBE/3z (TZVP) level.

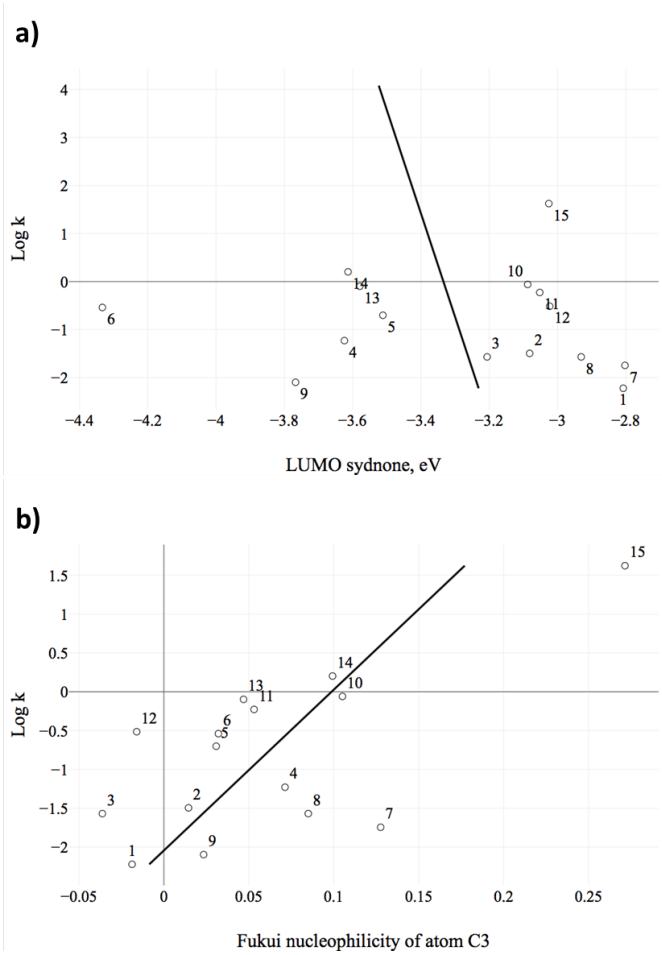


Figure 50. Dependency of rate constant on (a) LUMO energy of sydnones and (b) Fukui nucleophilicity on C3 atom. Reaction numbers corresponding to Table 7 is shown next to points.

We suggested that sydnone-alkyne reaction could be charge controlled. Fukui nucleophilicity, electrophilicity, radical attack susceptibility using Conceptual DFT [244] for 5 core atoms of sydnones were calculated and used to find correlations with log*k* of sydnones – BCN cycloaddition. Only one unbiased correlation was found between log*k* and nucleophilicity of C3 atom with $R^2 = 0.43$ (Figure 50b). However, one could notice that correlation was mainly caused by fluorosydnone-BCN reaction **15**. The deletion of the latter point will lead to dramatic drop of correlation coefficient. Thus, charge control does not play the major role in these cycloaddition reactions.

The result leads us to conclusion that there is a complex interplay of structural factors that cannot be caught by simple linear correlations. To avoid any additional effect reaction rate in small congeneric series of reactions of sydnones having different

substituents X and the same R with BCN (reactions **3**, **10-12**, **15**) and fluorosydnones (X=F) having different R with TMTH (reactions **16-18**) was analyzed.

Results for reaction of BCN with halo-sydnones (**3**, **10-12**, **15**) given on Figure 51 in blue show significant variation of electrostatic charges on C3 atom, which change from negative (X = H) to positive value (X = F), whereas the charge on N2 atom slightly varies as a function of substituents. In this series, the logarithm of the cycloaddition rate (log*k*) well correlates with the C3 charges (Figure 51a), r=0.98 and RMSE = 0.13 log*k*, p value $4 \cdot 10^{-4}$. At the same time no correlation with LUMO energy was observed (Figure 51b).

Variation of aryl substituents at N3 atom in sydnones in reaction **16-18** (with TMTH, in red on Figure 51a), doesn't lead to significant variations of charge at C3 atom (Figure 51a). One could notice that the rate of reaction with TMTH is affected to charge variation in greater extent than BCN. Thus electron nature of substituents on sydnone ring will have stronger influence on reaction with TMTH. On the other hand, LUMO energy in this series decreases from -3.57 to -3.05 eV which could partially explain considerable raise of the rate constant (Figure 51b). One should take into account that correlation coefficient is not big enough (0.86) and 3 points are too few to make robust conclusions.

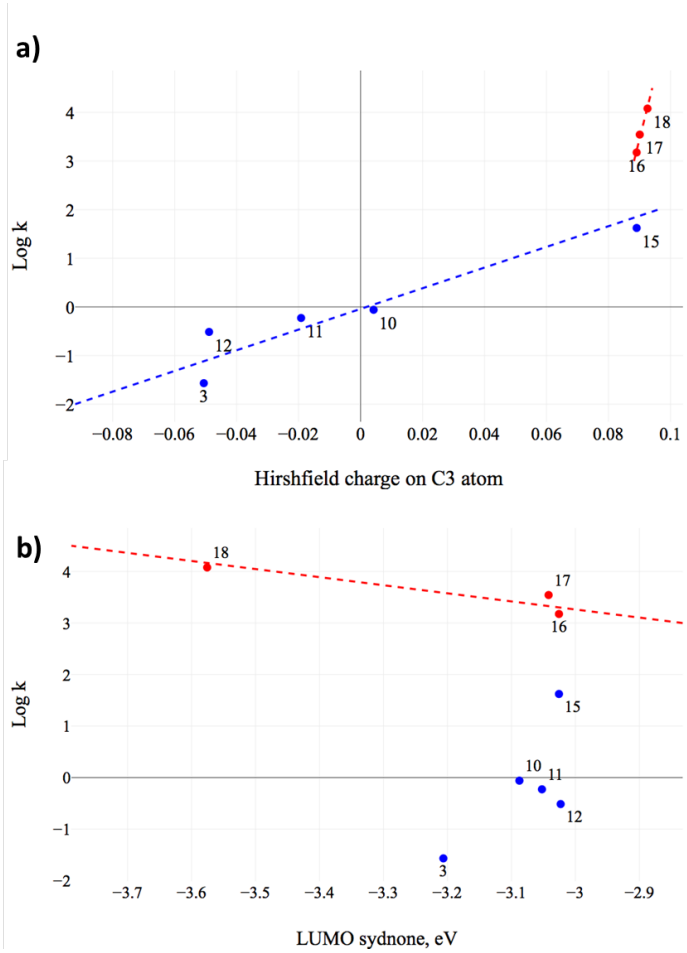


Figure 51. Dependency of rate constant of congeneric series of halo-sydnones on (a) charge of C3 atom of sydnone and (b) LUMO of sydnone. Correlations are shown as dashed lines. Red point and lines correspond to reactions with TMTH, blue ones to reactions with BCN.

Larger steric strain of alkyne in TMTH compared to BCN affects both charges on triple bond carbon atoms (-0.663 and -0.652 in BCN and -0.636 for both atoms in TMTH) and HOMO energy (-5.88 eV in BCN and -5.15 eV in TMTH, Figure 49). The latter effect favors larger reactivity of TMTH.

DFT calculations of the transition state of the cycloaddition of sydnones with BCN reveals an increase of the distance between alkyne C atom and C3 atom (C_{alkyne}-C3) with the size of the substituent at C4 atom: 2.283, 2.345 and 2.348 Å for X = H, F and I, respectively, Figure 4.3.6-4. This could be explained by steric repulsion between substituent X and aliphatic CH_2 group next to triple bond in the alkyne moiety in the transition state. The shortest contact H…I (3.486 Å) in transition state TS1 is almost equal to the sum of the Pauling vdW radii of H and I (3.4 Å). While fluorine atom is

almost located in the plane of sydnone ring in TS1, iodine atom deviates from it avoiding sterical clashes, Figure 52. The sums of valence angels for reactions **15**, **2**, **10**, **11**, **12** are 353.6, 353.2, 351.9, 351.5, 351.2 correspondingly. This measure can evaluate steric tension in TS1. As one can see sydnone with F is even more planar in TS1 than sydnone with H. We can consider that negative charge on F atom and positive charge on hydrogens of alkyne can form weak hydrogen bond in this phase, that reduces energy of TS1.

These observations are opposite to suggestions by Liang et al [236] that electronically more reactive electrophiles should be sterically more encumbered. Indeed, on one hand, iodo-sydnone, sterically more encumbered than fluoro-sydnone, is less reactive than the latter due to weaker electron acceptor ability and, on the other hand, steric repulsion of halogen with alkyne molecule. But, indeed, fluorine change to more bulky electron-withdrawing substituent could lead to drastic loss of reaction rate. Thus, reaction **9** involving sydnone with X=CF_3 is drastically slower than **10** with X=Cl, despite atomic charges on C3 are close (C3 charge is 0.0093 when X=CF_3, 0.0042 for X=Cl, 0.0891 for X=F).

In the context of rational design of highly reactive substrates, our calculations result in the following conclusions:

1. Strong electron acceptor X at C3 atom increases its positive charge and, thus, improves its affinity to alkyne in cycloaddition. However, this substituent should not be bulky because it may cause steric hindrance to alkyne in the transition state. In this context, fluorine substituent at C4 represents an optimal choice. However, other flat acceptors could be a good alternative as well.

2. Strong electron acceptor at N3 atom weakly affects charge distribution in sydnone moiety, but decreases its LUMO energy, which, in turn, favors cycloaddition, especially in reaction with TMTH. Thus, one can expect that nitro-groups or other strong electron withdrawing substituents in benzene ring R will strongly favor reaction.

3. Compared to BCN, more sterically strained TMTH has higher HOMO energy which explains its better reactivity.

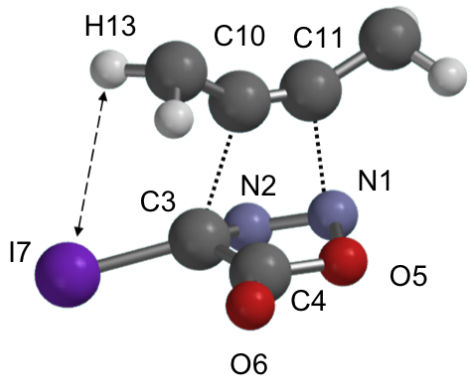


Figure 52. Transition state of cycloaddition reaction between sydnone 1b (X=I) and BCN. Only few atoms of both reactants are shown for clarity. Dashed line shows steric contacts H…I between halogen atom of sydnone and H atoms attached to one of carbons at the triple bond of alkyne, dotted lines indicate new single bonds forming in this reaction.

Conclusions

Extensive quantum chemical study of sydnone-alkyne cycloaddition reaction have been performed. Quantum-chemical exploration of reaction pathway have supported previous findings [233] on similar systems that despite reaction have formally two-step mechanism the second step has a tiny barrier and intermediate is located substantially lower in energetic scale than reagents. The intermediate was observed experimentally on the same systems by Liu et al [6]. They discovered that the second step of the reaction was the limiting step, in contrast to our results. Yet, for practical application, the effective reaction rates are reporting the consumption of the reagents which is controlled by the rate of the first step of the reaction. Besides, this rate of the first step is much more variable with the nature of the reagents than the second step. This is confirmed by our quantum-chemical calculations well reproducing experimental data on reaction kinetics based on Transition State theory. The root-mean squared deviation between quantum-chemically calculated activation free energies and the value that was recalculated from experimental data on reaction rate constant using Eyring equantion on 18 reaction of sydnone-alkyne cycloaddition is lower than 2 kcal/mol. We did not have yet a satisfactory understanding of the mechanism of the second step of the reaction. But, since our results are not changed using an implicit solvation model, it is possible that some solvent water molecules play an active role in the process.

From the other hand, such a good reproducibility of activation free energy and as a consequence reasonable prediction of rate constant by quantum chemical calculations gives a powerful tool for relatively cheap screening of possible reagents. Application of developed workflow could be used to computationally prove hypothesis on reactivity of certain pairs of sydnone and alkyne. Only the most promising candidates could be synthesized and their reaction rate could be experimentally measured.

To reduce search space in this work the structural factors important for reaction rate were analyzed. The absence of clear correlation within whole dataset of reactions could be explained by complex interplay of effects responsible for reaction rate and some experimental noise in reaction rate constant measurement. However, in restricted series of reactions some correlations could be found that made possible to reveal three major factors affecting reaction rate. First, large positive charge on C3 atom guarantees faster reactions (atom numbering according to Figure 47). This charge itself is mainly affected by electron withdrawing ability of substituent at this atom (X) and partially on the electron withdrawing ability of substituent at N2 atom (R). On the other hand, bulky substituents at C3 atom lead to sterical hindrance to approaching reagents and complicate transition state formation. Thus, only sterically unencumbered electron withdrawing groups could be used at C3 atom to gain reaction speed. Since LUMO of sydnone is participating in orbital interactions, its energy is an important factor influencing rate constant. The lower the energy of LUMO the greater is rate of reaction. Mostly LUMO energy is affected by π-electron withdrawing ability of substituent R. Thus, usage of stronger electron acceptors will favor reaction. Being augmented by proposed workflow for fast screening of reaction rate these recommendations provide an efficient tool to the design of more active agents for bioorthogonal click reactions.

The quality of correlation between free energy of activation calculated quantum-chemically and estimated from experimental value of rate constant using Eyring equation shows from one side that developed quantum-chemical approach reproduces the energy rather well in absolute scale (less than 2 kcal/mol). From the other side, having such good description the correlation coefficient is not large (r is 0.84, that means that determination coefficient would be at most 0.6). The reason could be noise in data and in this case it will explain why we failed to build QSPR model on the dataset – high level of noise in

addition to other problem such as imbalances and heterogeneity of dataset prevented building predictive model.

Chapter 10.

General conclusions

1. The first database of chemical reaction kinetic and thermodynamic properties was collected. Information on more than 10 000 chemical reactions involving structure, solvent, temperature, and rate constants of bimolecular nucleophilic substitution, bimolecular elimination, cycloaddition (Diels-Alder reactions) and tautomeric equilibrium constants was annotated from reference books and PhD thesis defended in Kazan Federal University.

2. The workflow for reaction standardization and curation was proposed and the tools required for it was developed. Some tools are based on approach of Condensed Graph of Reaction. The latter was extended by notion of dynamic atom to extend its applicability to wider range of reactions. The approach to store CGRs in SDF format was proposed. The toolbox developed includes CGR hashing, reaction center detection and hashing, workflows for AAM checking and conditions verification.

3. The workflow for reaction property modeling was proposed. It incorporates usage of Condensed Graph of Reaction based fragment descriptors for encoding chemical transformation in combination with solvent and temperature descriptors to represent reaction conditions. Having descriptor vector one can use any machine learning method for model creation (SVM and RF were tried). Using developed approach the model predicting the rate or equilibrium constants of reactions involving various reagents, which occur in many organic solvents and water-organic mixtures were built the first time. It is shown that the RMSE of prediction is comparable with the level of experimental noise. The analysis of prediction errors also shows that the quality of the model is sufficiently high for the identification of data errors and objects with the unique structure with respect to this set of reactions.

4. New type of descriptors for chemical reactions based on mixture representations of reagents and products using SiRMS approach was proposed. Different types of structural descriptors of chemical reactions were benchmarked on cleaned E2 reaction dataset. It showed that for this particular dataset three best structural descriptors are SiRMS mixture descriptors, CGR-based ISIDA fragment descriptors and Morgan fingerprint-based difference reaction fingerprint.

5. The study has clearly shown the importance of correct validation scheme for unbiased estimation of predictive performance of chemical reaction. Two different

strategies that are superior to classic cross-validation were proposed: the first one based on calculating predictive performance metrics only on point for which reaction property was measured in only one condition, the second one based on stratified product-out cross-validation. It was shown that these strategies avoid too optimistic estimation of models performance.

6. Predictive models for rate constants of bimolecular nucleophilic substitution, bimolecular elimination, and Diels-Alder reactions, as well as tautomeric equilibrium constants were built. The models were published on-line on server developed specially for chemical reactions.

7. To predict sydnone-alkyne cycloaddition reaction rate constant the workflow based on quantum-chemical calculations and semi-automatic identification of transition state was developed. QSPR modeling of this dataset failed. Using quantum chemistry approach, activation free energy of reactions under study were reproduced with some 2 kcal/mol accuracy. To speed up selection of optimal reagents for these cycloadditions, most important factors affecting the reaction rates were reported.

References

1. Palm VA (1978) Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions. VINITI, Moscow
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. Drucker H, Burges CJC, Kaufman L, et al (1997) Support vector regression machines. In: Mozer MC, Jordan JI, Petsche JI (eds) Adv. Neural Inf. Process. Syst. MIT Press, pp 155–161
4. Tomasi J, Mennucci B, Cances E (1999) The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level. J Mol Struct THEOCHEM 464:211–226.
5. Marenich A V, Cramer CJ, Truhlar DG (2009) Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J Phys Chem B 113:6378–96. doi: 10.1021/jp810292n
6. Liu H, Audisio D, Plougastel L, et al (2016) Ultrafast Click Chemistry with Fluorosydnones. Angew Chemie Int Ed 55:12073–12077. doi: 10.1002/anie.201606495
7. Varnek A, Baskin II (2011) Chemoinformatics as a Theoretical Chemistry Discipline. Mol Inform 30:20–32. doi: 10.1002/minf.201000100
8. Engkvist O, Norrby P-O, Selmi N, et al (2018) Computational prediction of chemical reactions: current status and outlook. Drug Discov Today in print. doi: https://doi.org/10.1016/j.drudis.2018.02.014
9. Baskin II, Madzhidov TI, Antipin IS, Varnek AA (2017) Artificial intelligence in synthetic chemistry: achievements and prospects. Russ Chem Rev 86:1127–1156. doi: 10.1070/RCR4746
10. Blakemore DC, Castro L, Churcher I, et al (2018) Organic synthesis provides opportunities to transform drug discovery. Nat Chem 10:383–394. doi: 10.1038/s41557-018-0021-z
11. Peverati R, Truhlar DG (2014) Quest for a universal density functional: the

accuracy of density functionals across a broad spectrum of databases in chemistry and physics. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 372:

12. Cohen ER, Cvitas T, Frey JG, et al (1993) Quantities, Units and Symbols in Physical Chemistry, IUPAC Green Book, 3rd Editio. IUPAC& RSC Publishing, Cambridge
13. Beguin J (1983) Tyrocinium Chymicum [1669 edition facsimile]. Reprint of the English translation by Richard Russel, with orig. imprint: London : T. Passenger, 1669. Heptangle Books, Gillette, N.J.
14. Struebing H, Ganase Z, Karamertzanis PGPG, et al (2013) Computer-aided molecular design of solvents for accelerated reaction kinetics. Nat Chem 5:952–957. doi: 10.1038/nchem.1755
15. Manion JA, Huie RE, Levin RD, et al NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version), Release 1.6.8, Data version 2015.09, National Institute of Standards and Technology. In: Gaithersburg, Maryl.
16. Mellouki W Chemical Kinetics Database on oxygenated VOCs gas-phase reactions.
17. Polishchuk P, Madzhidov T, Gimadiev T, et al (2017) Structure–reactivity modeling using mixture-based representation of chemical reactions. J Comput Aided Mol Des 0:1–11. doi: 10.1007/s10822-017-0044-3
18. Oprisiu I, Varlamova E, Muratov E, et al (2012) QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids. Mol Inform 31:491–502. doi: 10.1002/minf.201200006
19. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Model 28:31–36. doi: 10.1021/ci00057a005
20. Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 29:97–101. doi: 10.1021/ci00062a008
21. (2008) SMIRKS - A Reaction Transform Language. Daylight theory manual, Chapter 7. . 2017:

22. Heller S, McNaught A, Stein S, et al (2013) InChI - the worldwide chemical structure identifier standard. J Cheminform 5:7. doi: 10.1186/1758-2946-5-7
23. Grethe G, Goodman JM, Allen CH (2013) International chemical identifier for reactions (RInChI). J Cheminform 5:45. doi: 10.1186/1758-2946-5-45
24. (2017) Downloads of InChI Software.
25. 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
26. Holliday GL, Murray-Rust P, Rzepa HS (2006) Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. J Chem Inf Model 46:145–157. doi: 10.1021/ci0502698
27. Chen WL, Chen DZ, Taylor KT (2013) Automatic reaction mapping and reaction center detection. Wiley Interdiscip Rev Comput Mol Sci 3:560–593. doi: 10.1002/wcms.1140
28. Chen L (2003) Reaction Classification and Knowledge Acquisition. In: Gasteiger J (ed) Handb. Chemoinformatics From Data to Knowl. WILEY-VCH, Weinheim, pp 348–388
29. Balaban AT (1967) Chemical graphs. 3. Reactions with cyclic 6-membered transition states. Rev Roum Chim 12:875–902.
30. Arens JF (1979) A formalism for the classification and design of organic reactions. I. The class of (− +)n reactions. Recl des Trav Chim des Pays-Bas 98:155–161. doi: 10.1002/recl.19790980403
31. Arens JF (1979) A formalism for the classification and design of organic reactions. II. The classes of (+ −)n + and (− +)n − reactions. Recl des Trav Chim des Pays-Bas 98:395–399. doi: 10.1002/recl.19790980606
32. Arens JF (1979) A formalism for the classification and design of organic reactions III. The class of (+ -)nC reactions. Recl des Trav Chim des Pays-Bas 98:471–483. doi: 10.1002/recl.19790980902
33. Hendrickson JB (1974) The Variety of Thermal Pericyclic Reactions. Angew Chemie Int Ed English 13:47–76. doi: 10.1002/anie.197400471
34. Hendrickson JB (1997) Comprehensive System for Classification and Nomenclature of Organic Reactions. J Chem Inf Comput Sci 37:852–860. doi:

10.1021/ci970040v

35. Zefirov NS (1987) An Approach to Systematization and Design of Organic Reactions. Acc Chem Res 20:237–243. doi: 10.1021/ar00139a001
36. Tratch SS, Zefirov NS (1998) A Hierarchical Classification Scheme for Chemical Reactions. J Chem Inf Comput Sci 38:349–366. doi: 10.1021/ci960098u
37. Vléduts GÉ (1963) Concerning one system of classification and codification of organic reactions. Inf Storage Retr 1:117–146. doi: 10.1016/0020-0271(63)90013-5
38. Vladutz G (1986) Do we still need a classification of reactions? In: Willett P (ed) Mod. Approaches to Chem. React. Search. Gower, London, pp 202–220
39. Kikho Y (1971) Formal definition of some notions of quantitative oraganic chemistry [in Russian]. Reaktsionnaya Spos. Org. Soedin. (Reactivity Org. Compd. 8:
40. Fujita S (1986) Description of organic reactions based on imaginary transition structures. 2. Classification of one-string reactions having an even-membered cyclic reaction graph. J Chem Inf Comput Sci 26:212–223. doi: 10.1021/ci00052a010
41. Jauffret P, Tonnelier C, Hanser T, et al (1990) Machine learning of generic reactions: 2. toward an advanced computer representation of chemical reactions. Tetrahedron Comput Methodol 3:335–349. doi: 10.1016/0898-5529(90)90060-L
42. Gimadiev TR, Madzhidov TI, Nugmanov RI, 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
43. Dugundji J, Ugi I (1973) An algebraic model of constitutional chemistry as a basis for chemical computer programs. In: Comput. Chem. Springer-Verlag, Berlin/Heidelberg, pp 19–64
44. Gasteiger J, Jochum C (1978) EROS A computer program for generating sequences of reactions. In: Org. Compunds. Springer-Verlag, Berlin/Heidelberg, pp 93–126
45. Jochum C, Gasteiger J, Ugi I (1980) The Principle of Minimum Chemical Distance(PMCD). Angew Chemie Int Ed English 19:495–505. doi: 10.1002/anie.198004953
46. Todeschini R, Consonni V (2000) Handbook of Molecular Descriptors. doi:

10.1002/9783527613106

47. Baskin I (2008) Chapter 1. Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. In: Varnek A (ed) Chemoinformatics Approaches to Virtual Screen. Royal Society of Chemistry, Cambridge, pp 1–43
48. Baskin II, Skvortsova MI, Stankevich I V., Zefirov NS (1995) On the Basis of Invariants of Labeled Molecular Graphs. J Chem Inf Model 35:527–531. doi: 10.1021/ci00025a021
49. Gonzalez-Diaz H, Vilar S, Santana L, Uriarte E (2007) Medicinal Chemistry and Bioinformatics - Current Trends in Drugs Discovery with Networks Topological Indices. Curr Top Med Chem 7:1015–1029. doi: 10.2174/156802607780906771
50. Karelson M, Lobanov VS, Katritzky AR (1996) Quantum-Chemical Descriptors in QSAR/QSPR Studies. Chem Rev 96:1027–1044. doi: 10.1021/cr950202r
51. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967. doi: 10.1021/ja00226a005
52. Bonachéra F, Parent B, Barbosa F, et al (2006) Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes. J Chem Inf Model 46:2457–2477. doi: 10.1021/ci6002416
53. Hammett LP (1937) The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. J Am Chem Soc 59:96–103. doi: 10.1021/ja01280a022
54. Zefirov NS, Palyulin VA (2002) Fragmental Approach in QSPR. J Chem Inf Comput Sci 42:1112–1122. doi: 10.1021/ci020010e
55. Artemenko N V., Baskin II, Palyulin VA, Zefirov NS (2003) Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds. Russ Chem Bull 52:20–29. doi: 10.1023/A:1022467508832
56. Jelfs S, Ertl P, Selzer P (2007) Estimation of p K a for Druglike Compounds Using Semiempirical and Information-Based Descriptors. J Chem Inf Model 47:450–459. doi: 10.1021/ci600285n
57. Varnek A, Tropsha A (2008) Chemoinformatics Approaches to Virtual Screening.

doi: 10.1039/9781847558879

58. Smolenski E. (1964) No Title. Zhurnal Fiz Khimii 1288–1291.
59. Wiswesser WJ (1982) How the WLN began in 1949 and how it might be in 1999. J Chem Inf Model 22:88–93. doi: 10.1021/ci00034a005
60. Rosenkranz HS, Klopman G (1986) Mutagens, carcinogens, and computers. Prog Clin Biol Res 209A:71–104.
61. Benson SW, Buss JH (1958) Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. J Chem Phys 29:546–572. doi: 10.1063/1.1744539
62. Adamson GW, Bush JA, McLure AHW, Lynch MF (1974) An Evaluation of a Substructure Search Screen System Based on Bond-Centered Fragments. J Chem Doc 14:44–48. doi: 10.1021/c160052a011
63. Zhang Q-Y, Aires-de-Sousa J (2005) Structure-Based Classification of Chemical Reactions without Assignment of Reaction Centers. J Chem Inf Model 45:1775–1783. doi: 10.1021/ci0502707
64. Avidon V V., Pomerantsev IA, Golender VE, Rozenblit AB (1982) Structure-activity relationship oriented languages for chemical structure representation. J Chem Inf Model 22:207–214. doi: 10.1021/ci00036a006
65. Schneider G, Neidhart W, Giller T, Schmid G (1999) "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. Angew Chemie Int Ed 38:2894–2896. doi: 10.1002/(sici)1521-3773(19991004)38:19<2894::aid-anie2894>3.0.co;2-f
66. Free SM, Wilson JW (1964) A Mathematical Contribution to Structure-Activity Studies. J Med Chem 7:395–399. doi: 10.1021/jm00334a001
67. Hansch C, Fujita T (1964) p -σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J Am Chem Soc 86:1616–1626. doi: 10.1021/ja01062a035
68. Fujita T, Iwasa J, Hansch C (1964) A New Substituent Constant, p, Derived from Partition Coefficients. J Am Chem Soc 86:5175–5180. doi: 10.1021/ja01077a028
69. Randic M (1992) Representation of molecular graphs by basic graphs. J Chem Inf Model 32:57–69. doi: 10.1021/ci00005a010
70. Kramer S, De Raedt L, Helma C (2001) Molecular feature mining in HIV data.

In: Proc. seventh ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '01. ACM Press, New York, New York, USA, pp 136–143

71. Asai T, Abe K, Kawasoe S, et al (2002) Efficient substructure discovery from large semi-structured data. 2nd SIAM Int. Conf. Data Min.
72. Graham DJ, Malarkey C, Schulmerich M V. (2004) Information Content in Organic Molecules: Quantification and Statistical Structure via Brownian Processing. J Chem Inf Comput Sci 44:1601–1611. doi: 10.1021/ci0400213
73. Sanderson DM, Earnshaw CG (1991) Computer Prediction of Possible Toxic Action from Chemical Structure; The DEREK System. Hum Exp Toxicol 10:261–273. doi: 10.1177/096032719101000405
74. 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
75. Glavatskikh M, Madzhidov T, Solov'ev V, et al (2016) Predictive Models for the Free Energy of Hydrogen Bonded Complexes with Single and Cooperative Hydrogen Bonds. Mol Inform 35:629–638. doi: 10.1002/minf.201600070
76. Fiorella R, Vitaly S, Gilles M, et al (2014) Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules. Mol Inform 33:477–487. doi: 10.1002/minf.201400032
77. Glavatskikh M, Madzhidov T, Solov'ev V, et al (2016) Predictive Models for Halogen-bond Basicity of Binding Sites of Polyfunctional Molecules. Mol Inform 35:70–80. doi: 10.1002/minf.201500116
78. Solov'ev V, Varnek A, Tsivadze A (2014) QSPR ensemble modelling of the 1:1 and 1:2 complexation of Co2+, Ni2+, and Cu2+ with organic ligands: relationships between stability constants. J Comput Aided Mol Des 28:549–564. doi: 10.1007/s10822-014-9741-3
79. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) ISIDA Property-labelled fragment descriptors. Mol Inform 29:855–868. doi: 10.1002/minf.201000099
80. 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
81. Kuz'min VE, Polischuk PG, Artemenko AG, et al (2008) Quantitative structure-

affinity relationship of 5-HT1A receptor ligands by the classification tree method. SAR QSAR Environ Res 19:213–244. doi: 10.1080/10629360802085090

82. Muratov EN, Artemenko AG, Varlamova E V, et al (2010) Per aspera ad astra: application of Simplex QSAR approach in antiviral research. Future Med Chem 2:1205–1226. doi: 10.4155/fmc.10.194

83. Polishchuk PG, Muratov EN, Artemenko AG, et al (2009) Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. J Chem Inf Model 49:2481–2488. doi: 10.1021/ci900203n

84. Kovdienko N, Polishchuk P, Muratov E, et al (2010) Application of Random Forest and Multiple Linear Regression Techniques to QSPR Prediction of an Aqueous Solubility for Military Compounds. Mol Inform 29:394–406. doi: 10.1002/minf.201000001

85. Mokshyna E, Polishchuk PG, Nedostup V, Kuzmin VE (2015) Predictive QSPR Modelling for the Second Virial Coefficient of the Pure Organic Compounds. Mol Inform 34:53–59. doi: 10.1002/minf.201400081

86. Polishchuk P, Mokshyna E, Kosinskaya A, et al (2017) Structural, Physicochemical and Stereochemical Interpretation of QSAR Models Based on Simplex Representation of Molecular Structure BT - Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences. In: Roy K (ed) Adv. QSAR Model. Springer International Publishing, Cham, pp 107–147

87. Mokshyna E, Polishchuk P, Nedostup V, Kuz'min V (2016) QSPR-Modeling for the Second Virial Cross-Coefficients of Binary Organic Mixtures. Int J Quant Struct Relationships 1:72–84. doi: 10.4018/IJQSPR.2016070104

88. Mokshyna E, Nedostup VI, Polishchuk PG, Kuzmin VE (2014) 'Quasi-Mixture' Descriptors for QSPR Analysis of Molecular Macroscopic Properties. The Critical Properties of Organic Compounds. Mol Inform 33:647–654. doi: 10.1002/minf.201400036

89. Taft RW (1952) Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters1. J Am Chem Soc 74:3120–3128. doi: 10.1021/ja01132a049

90. Reichardt C, Welton T (2010) Solvents and Solvent Effects in Organic Chemistry,

4th Editio. doi: 10.1002/9783527632220

91. Salin A, Fatkhutdinov A, Il'in A, et al (2014) Solvent Effect on Kinetics and Mechanism of the Phospha-Michael Reaction of Tertiary Phosphines with Unsaturated Carboxylic Acids. Heteroat Chem 25:205–216. doi: 10.1002/hc.21161
92. Sukhachev D V., Pivina TS, Zhokhova NI, et al (1995) QSPR approach to the calculation of rate constants of homolysis of nitro compounds in different states of aggregation. Russ Chem Bull 44:1585–1588. doi: 10.1007/BF01151274
93. Marcou G, Aires de Sousa J, Latino DARS, et al (2015) Expert System for Predicting Reaction Conditions: The Michael Reaction Case. J Chem Inf Model 55:239–250. doi: 10.1021/ci500698a
94. Latino DARS, Zhang Q-Y, Aires-de-Sousa J (2008) Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. Bioinformatics 24:2236–2244. doi: 10.1093/bioinformatics/btn405
95. Elhabiri M, Sidorov P, Cesar-Rodo E, et al (2015) Electrochemical Properties of Substituted 2-Methyl-1,4-Naphthoquinones: Redox Behavior Predictions. Chem – A Eur J 21:3415–3424. doi: 10.1002/chem.201403703
96. 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.
97. Kravtsov AA, Karpov P V, Baskin II, et al (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 441:314–317. doi: 10.1134/S0012500811110048
98. Hu Q-N, Zhu H, Li X, et al (2012) Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints. PLoS One 7:e52901. doi: 10.1371/journal.pone.0052901
99. (2008) Fingerprints - Screening and Similarity. Daylight theory manual, Chapter 6. . 2017:
100. 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 55:39–53. doi: 10.1021/ci5006614
101. Ridder L, Wagener M (2008) SyGMa: Combining Expert Knowledge and

Empirical Scoring in the Prediction of Metabolites. ChemMedChem 3:821–832. doi: 10.1002/cmdc.200700312

102. Faulon J-L, Misra M, Martin S, et al (2008) Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. Bioinformatics 24:225–233. doi: 10.1093/bioinformatics/btm580
103. Hoonakker F (2008) Graphes condensés de réactions, applications à la recherche par similarité, la classification et la modélisation. Louis Pasteur University of Strasbourg
104. De Luca A, Horvath D, Marcou G, et al (2012) Mining chemical reactions using neighborhood behavior and condensed graphs of reactions approaches. J Chem Inf Model 52:2325–2338. doi: 10.1021/ci300149n
105. Patterson DE, Cramer RD, Ferguson AM, et al (1996) Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. J Med Chem 39:3049–3059. doi: 10.1021/jm960290n
106. Onsager L (1936) Electric Moments of Molecules in Liquids. J Am Chem Soc 58:1486–1493. doi: 10.1021/ja01299a050
107. Marcus Y (1998) The Properties of Solvents. Wiley-VCH, Weinheim
108. 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
109. 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
110. 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 1997:1941–1949. doi: 10.1002/jlac.199719970921
111. Kamlet MJ, Taft RW (1976) The solvatochromic comparison method. I. The .beta.-scale of solvent hydrogen-bond acceptor (HBA) basicities. J Am Chem Soc 98:377–383. doi: 10.1021/ja00418a009
112. Taft RW, Kamlet MJ (1976) The solvatochromic comparison method. 2. The

.alpha.-scale of solvent hydrogen-bond donor (HBD) acidities. J Am Chem Soc 98:2886–2894. doi: 10.1021/ja00426a036

113. Kamlet MJ, Abboud JL, Taft RW (1977) The solvatochromic comparison method. 6. The .pi.* scale of solvent polarities. J Am Chem Soc 99:6027–6038. doi: 10.1021/ja00460a031
114. Koppel IA, Palm VA (1972) The influence of the solvent on organic reactivity. In: Chapman NB, Shorter J (eds) Adv. Linear Free Energy Relationships. Plenum Press, London, pp 203–280
115. Wells PR (1963) Linear Free Energy Relationships. Chem Rev 63:171–219. doi: 10.1021/cr60222a005
116. Palm VA (1977) Fundamentals of the Quantitative Theory of Organic Reactions, in Russian. Khimiya, Leningrad
117. Domingo LR, Saez JA (2009) Understanding the mechanism of polar Diels-Alder reactions. Org Biomol Chem 7:3576–3583. doi: 10.1039/B909611F
118. Kohn W, Sham LJ (1965) Self-Consistent Equations Including Exchange and Correlation Effects. Phys Rev 140:A1133–A1138. doi: 10.1103/PhysRev.140.A1133
119. Hohenberg P, Kohn W (1964) Inhomogeneous Electron Gas. Phys Rev 136:B864--B871. doi: 10.1103/PhysRev.136.B864
120. Roothaan CCJ, Bagus PS (1963) Methods in Computational Physics. In: Quantum Mech. Academic Press New York, p 47
121. Møller C, Plesset MS (1934) Note on an approximation treatment for many-electron systems. Phys Rev 46:618–622. doi: 10.1103/PhysRev.46.618
122. Hegarty D, Robb MA (1979) Application of unitary group methods to configuration interaction calculations. Mol Phys 38:1795–1812. doi: 10.1080/00268977900102871
123. Bartlett RJ, Purvis GD (2004) Many-body perturbation theory, coupled-pair many-electron theory, and the importance of quadruple excitations for the correlation problem. Int J Quantum Chem 14:561–581. doi: 10.1002/qua.560140504
124. Mardirossian N, Head-Gordon M (2017) Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. Mol Phys 115:2315–2372. doi: 10.1080/00268976.2017.1333644

125. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648–5652. doi: 10.1063/1.464913
126. Perdew JP, Burke K, Ernzerhof M (1996) Generalized Gradient Approximation Made Simple. Phys Rev Lett 77:3865–3868. doi: 10.1103/PhysRevLett.77.3865
127. Bond D (2007) Computational methods in organic thermochemistry. 1. Hydrocarbon enthalpies and free energies of formation. J Org Chem 72:5555–5566. doi: 10.1021/jo070383k
128. Petersson GA, Bennett A, Tensfeldt TG, et al (1988) A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. J Chem Phys 89:2193–2218. doi: 10.1063/1.455064
129. Greenwood JR, Calkins D, Sullivan AP, Shelley JC (2010) Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. J Comput Aided Mol Des 24:591–604. doi: 10.1007/s10822-010-9349-1
130. Geballe M, Skillman AG, Nicholls A, et al (2010) The SAMPL2 blind prediction challenge: introduction and overview. J Comput Aided Mol Des 24:259–279. doi: 10.1007/s10822-010-9350-8
131. Eyring H (1935) The Activated Complex in Chemical Reactions. J Chem Phys 3:107–115. doi: 10.1063/1.1749604
132. Laidler KJ, King MC (1983) Development of transition-state theory. J Phys Chem 87:2657–2664. doi: 10.1021/j100238a002
133. Minkin VI, Minyaev BY, Simkin RM (1990) Quantum Chemistry of Organic Compounds. Mechanisms of Reactions. doi: 10.1007/978-3-642-75679-5
134. Fernández-Ramos A, Miller JA, Klippenstein SJ, Truhlar DG (2006) Modeling the kinetics of bimolecular reactions. Chem Rev 106:4518–4584. doi: 10.1021/cr050205w
135. Wold S, Sjorstrom M (1978) Linear Free Energy Relationships as Tools for Investigation Chemical Similarity - Theory and Practice. In: Chapman NB, J. S (eds) Correl. Anal. Chem. Recent Adv. Plenum Press, New York, pp 1–54
136. Johnson CD (1975) Linear free energy relations and the reactivity-selectivity principle. Chem Rev 75:755–765.
137. Halberstam NM, Baskin II, Palyulin VA, Zefirov NS (2002) Quantitative Structure

-Conditions - Property Relationships Studies. Neural Network Modelling of the Acid Hydrolysis of Esters. Mendeleev Commun 12:185–186.

138. Zhokhova NI, Baskin II, Palyulin VA, et al (2007) Fragmental descriptors with labeled atoms and their application in QSAR/QSPR studies. Dokl Chem 417:282–284.
139. 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.
140. 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.
141. 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
142. Carey JS, Laffan D, Thomson C, Williams MT (2006) Analysis of the reactions used for the preparation of drug candidate molecules. Org Biomol Chem 4:2337–2347. doi: 10.1039/B602413K
143. Roughley SD, Jordan AM (2011) The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. J Med Chem 54:3451–3479. doi: 10.1021/jm200187y
144. (2015) NameRxn.
145. Schneider N, Lowe DM, Sayle RA, et al (2016) Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. J Med Chem 59:4385–4402. doi: 10.1021/acs.jmedchem.6b00153
146. Chen L, Gasteiger J (1997) Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network. J Am Chem Soc 119:4033–4042. doi: 10.1021/ja960027b
147. Sello G, Termini M (1997) Classification of organic reactions using similarity. Tetrahedron 53:14085–14106. doi: http://dx.doi.org/10.1016/S0040-4020(97)00911-3
148. RDKit: Open-source cheminformatics.
149. 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

150. Skoraczyński G, Dittwald P, Miasojedow B, et al (2017) Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? Sci Rep 7:3582. doi: 10.1038/s41598-017-02303-0
151. Ahneman DT, Estrada JG, Lin S, et al (2018) Predicting reaction performance in C–N cross-coupling using machine learning. Science (80-) 5169:eaar5169. doi: 10.1126/science.aar5169
152. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. doi: 10.1007/BF00994018
153. Breiman L (2001) Random Forests. Mach Learn 45:5–32. doi: 10.1023/A:1010933404324
154. Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106. doi: 10.1007/BF00116251
155. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140. doi: 10.1007/BF00058655
156. 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
157. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. Altern Lab Anim 33:445–59.
158. Netzeva TI, Worth A, Aldenberg T, et al (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim 33:155–73.
159. Mathea M, Klingspohn W, Baumann K (2016) Chemoinformatic Classification Methods and their Applicability Domain. Mol Inform 35:160–180. doi: 10.1002/minf.201501019
160. 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

161. Gaspar HA, Baskin II, Marcou G, et al (2015) GTM-Based QSAR Models and Their Applicability Domains. Mol Inform 34:348–356. doi: 10.1002/minf.201400153
162. Horvath D, Brown J, Marcou G, Varnek A (2014) An Evolutionary Optimizer of libsvm Models. Challenges 5:450–472. doi: 10.3390/challe5020450
163. 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
164. Cohen AJ, Mori-Sánchez P, Yang W (2012) Challenges for Density Functional Theory. Chem Rev 112:289–320. doi: 10.1021/cr200107z
165. Laikov DN, Ustynyuk YA (2005) PRIRODA-04: a quantum-chemical program suite. New possibilities in the study of molecular systems with the application of parallel computing. Russ Chem Bull 54:820–826.
166. Laikov DN (1997) Fast evaluation of density functional exchange-correlation terms using the expansion of the electron density in auxiliary basis sets. Chem Phys Lett 281:151–156. doi: 10.1016/S0009-2614(97)01206-2
167. Schäfer A, Huber C, Ahlrichs R (1994) Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. J Chem Phys 100:5829–5835. doi: 10.1063/1.467146
168. Frisch MJ, Trucks GW, Schlegel HB, et al (2009) Gaussian 09 Revision C.01. Gaussian Inc., Wallingford
169. Tomasi J, Mennucci B, Cammi R (2005) Quantum Mechanical Continuum Solvation Models. Chem Rev 105:2999–3094. doi: 10.1021/cr9904009
170. Cramer CJ, Truhlar DG (1992) An SCF Solvation Model for the Hydrophobic Effect and Absolute Free Energies of Aqueous Solvation. Science (80-) 256:213–217. doi: 10.1126/science.256.5054.213
171. Frisch MJ, Trucks GW, Schlegel HB, et al (2009) Gaussian 09 Revision D.01. Gaussian Inc., Wallingford
172. Python Software Foundation. Python 3.5 language. https://www.python.org.

173. Hagberg AA, Schult DA, Swart PJ (2008) Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds) Proc. 7th Python Sci. Conf. Pasadena, CA USA, pp 11–15
174. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in {P}ython. J Mach Learn Res 12:2825–2830.
175. Pony ORM. https://ponyorm.com. Accessed 21 May 2018
176. The PostgreSQL Global Development Group. PostgreSQL. https://www.postgresql.org. Accessed 21 May 2018
177. McKinney W (2010) Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J (eds) Proc. 9th Python Sci. Conf. pp 51–56
178. Oliphant TE (2006) A guide to NumPy.
179. ChemAxon (2012) ChemAxon Kft., Záhony u. 7, Building HX, 1031 Budapest, Hungary.
180. (2012) Marvin 5.8.2.
181. MDL (2005) CTFile Formats.
182. Morgan HL (1965) The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. J Chem Doc 5:107–113. doi: 10.1021/c160017a018
183. Ihlenfeldt WD, Gasteiger J (1994) Hash codes for the identification and classification of molecular structure elements. J Comput Chem 15:793–813. doi: 10.1002/jcc.540150802
184. James CA (2016) OpenSMILES specification. In: www.opensmiles.org.
185. (2009) ICClassify, The InfoChem Reaction Classification Program.
186. Kraut H, Eiblmaier J, Grethe G, et al (2013) Algorithm for Reaction Classification. J Chem Inf Model 53:2884–2895. doi: 10.1021/ci400442f
187. ChemAxon. (2015) InstantJChem 15.7.27.0.
188. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50:1189–1204.
189. Tropsha A (2010) Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform 29:476–488. doi: 10.1002/minf.201000061
190. Fourches D, Muratov E, Tropsha A (2015) Curation of chemogenomics data. Nat

Chem Biol 11:535. doi: 10.1038/nchembio.1881

191. Fourches D, Muratov E, Tropsha A (2016) Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. J Chem Inf Model 56:1243–1252. doi: 10.1021/acs.jcim.6b00129
192. ChemAxon. (2015) Standardizer, JChem 15.8.3.0.
193. Jochum C, Gasteiger J, Ugi I (1980) The Principle of Minimum Chemical Distance (PMCD). Angew Chemie Int Ed English 19:495–505. doi: 10.1002/anie.198004953
194. Rivest R (1992) The MD5 Message-Digest Algorithm. RFC 1321 1–21.
195. Connors KA (1990) Chemical Kinetics: The Study of Reaction Rates in Solution. John Wiley & Sons, New York
196. Marcou G, Solov'ev V, Horvath D, Varnek A (2012) ISIDA Fragmentor - User Manual.
197. Smith MB, March J (2007) March's advanced organic chemistry: reactions, mechanisms, and structure, Sixth edit. John Wiley & Sons, Hoboken
198. Ingold C (1953) Structure and Mechanism in Organic Chemistry. Cornell University Press, New York
199. Kireeva N, Baskin II, Gaspar H a., et al (2012) Generative Topographic Mapping (GTM): Universal tool for data visualization, structure-activity modeling and dataset comparison. Mol Inform 31:301–312. doi: 10.1002/minf.201100163
200. Chen L, Gasteiger J (1996) Organic Reactions Classified by Neural Networks: Michael Additions, Friedel–Crafts Alkylations by Alkenes, and Related Reactions. Angew Chemie Int Ed English 35:763–765. doi: 10.1002/anie.199607631
201. Bishop CM, Svensén M, Williams CKII (1998) GTM: The Generative Topographic Mapping. Neural Comput 10:215–234. doi: 10.1162/089976698300017953
202. 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
203. Peacock DH (1925) CCXCIII.-The velocity of benzylation of certain amines. Part II. J Chem Soc Trans 127:2177–2180. doi: 10.1039/CT9252702177

204. Matsui T, Tokura N (1971) Solvent Effects on ρ Values of the Hammett Equation. II. Bull Chem Soc Jpn 44:756–761. doi: 10.1246/bcsj.44.756
205. (2009) ICClassify. The InfoChem Reaction Classification Program.
206. Alexander S (1875) Zur Kenntniss der Reihenfolge der Analgerung und Ausscheidung der Jodwasserstoffelemente in organischen Verbindungen. Justus Liebigs Ann Chem 179:296–301. doi: 10.1002/jlac.18751790304
207. Clark T (2010) Tautomers and reference 3D-structures: the orphans of in silico drug design. J Comput Aided Mol Des 24:605–611. doi: 10.1007/s10822-010-9342-8
208. Pospisil P, Ballmer P, Scapozza L, Folkers G (2003) Tautomerism in computer-aided drug design. J Recept Signal Transduct Res 23:361–371. doi: 10.1081/RRS-120026975 [doi]
209. Oellien F, Cramer J, Beyer C, et al (2006) The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening †. J Chem Inf Model 46:2342–2354. doi: 10.1021/ci060109b
210. Martin Y (2009) Let's not forget tautomers. J Comput Aided Mol Des 23:693–704. doi: 10.1007/s10822-009-9303-2
211. Warr W (2010) Tautomerism in chemical information management systems. J Comput Aided Mol Des 24:497–520. doi: 10.1007/s10822-010-9338-4
212. Sayle RA (2010) So you think you understand tautomerism? J Comput Aided Mol Des 24:485–496.
213. Trepalin S V, Skorenko A V, Balakin K V, et al (2003) Advanced Exact Structure Searching in Large Databases of Chemical Compounds. J Chem Inf Comput Sci 43:852–860. doi: 10.1021/ci025582d
214. Kubinyi H Virtual screening-problems and success stories. 4th Eur. Work. Drug Des.
215. Guasch L, Sitzmann M, Nicklaus MC (2014) Enumeration of ring-chain tautomers based on SMIRKS rules. J Chem Inf Model 54:2423–32. doi: 10.1021/ci500363p
216. Sitzmann M, Ihlenfeldt W-D, Nicklaus MC (2010) Tautomerism in large databases. J Comput Aided Mol Des 24:521–51. doi: 10.1007/s10822-010-9346-4
217. Szegezdi J, Csizmadia F (2007) Tautomer generation. pKa based dominance conditions for generating dominant tautomers. 234th ACS Natl. Meet. Boston,

MA, August 19-23, 2007

218. Garcia-Viloca M, Alhambra C, Truhlar DG, Gao J (2003) Hydride transfer catalyzed by xylose isomerase: mechanism and quantum effects. J Comput Chem 24:177–190.

219. Stigliani J-L, Arnaud P, Delaine T, et al (2008) Binding of the tautomeric forms of isoniazid-NAD adducts to the active site of the Mycobacterium tuberculosis enoyl-ACP reductase (InhA): A theoretical approach. J Mol Graph Model 27:536–545. doi: 10.1016/j.jmgm.2008.09.006

220. Todorov NP, Monthoux PH, Alberts IL (2006) The Influence of Variations of Ligand Protonation and Tautomerism on Proteinв€'Ligand Recognition and Binding Energy Landscape. J Chem Inf Model 46:1134–1142. doi: 10.1021/ci050071n

221. Rastelli G, Thomas B, Kollman PA, Santi D V (1995) Insight into the specificity of thymidylate synthase from molecular dynamics and free energy perturbation calculations. J Am Chem Soc 117:7213–7227. doi: 10.1021/ja00132a022

222. Sletten EM, Bertozzi CR (2011) From Mechanism to Mouse: A Tale of Two Bioorthogonal Reactions. Acc Chem Res 44:666–676. doi: 10.1021/ar200148z

223. Lim RK V, Lin Q (2010) Bioorthogonal chemistry: recent progress and future directions. Chem Commun 46:1589–1600. doi: 10.1039/B925931G

224. Gong Y, Pan L (2015) Recent advances in bioorthogonal reactions for site-specific protein labeling and engineering. Tetrahedron Lett 56:2123–2132. doi: 10.1016/j.tetlet.2015.03.065

225. Kennedy DC, McKay CS, Legault MCB, et al (2011) Cellular Consequences of Copper Complexes Used To Catalyze Bioorthogonal Click Reactions. J Am Chem Soc 133:17993–18001. doi: 10.1021/ja2083027

226. Mbua NE, Guo J, Wolfert MA, et al (2011) Strain-Promoted Alkyne–Azide Cycloadditions (SPAAC) Reveal New Features of Glycoconjugate Biosynthesis. ChemBioChem 12:1912–1921. doi: 10.1002/cbic.201100117

227. Sletten EM, Bertozzi CR (2009) Bioorthogonal Chemistry: Fishing for Selectivity in a Sea of Functionality. Angew Chemie Int Ed 48:6974–6998. doi: 10.1002/anie.200900942

228. Browne DL, Helm MD, Plant A, Harrity JPA (2007) A Sydnone Cycloaddition

Route to Pyrazole Boronic Esters. Angew Chemie Int Ed 46:8656–8658. doi: 10.1002/anie.200703767

229. Kolodych S, Rasolofonjatovo E, Chaumontet M, et al (2013) Discovery of Chemoselective and Biocompatible Reactions Using a High-Throughput Immunoassay Screening. Angew Chemie Int Ed 52:12056–12060. doi: 10.1002/anie.201305645
230. Specklin S, Decuypere E, Plougastel L, et al (2014) One-Pot Synthesis of 1,4-Disubstituted Pyrazoles from Arylglycines via Copper-Catalyzed Sydnone–Alkyne Cycloaddition Reaction. J Org Chem 79:7772–7777. doi: 10.1021/jo501420r
231. Decuypere E, Specklin S, Gabillet S, et al (2015) Copper(I)-Catalyzed Cycloaddition of 4-Bromosydnones and Alkynes for the Regioselective Synthesis of 1,4,5-Trisubstituted Pyrazoles. Org Lett 17:362–365. doi: 10.1021/ol503482a
232. Wallace S, Chin JW (2014) Strain-promoted sydnone bicyclo-[6.1.0]-nonyne cycloaddition. Chem Sci 5:1742–1744. doi: 10.1039/C3SC53332H
233. Narayanam MK, Liang Y, Houk KN, Murphy JM (2016) Discovery of new mutually orthogonal bioorthogonal cycloaddition pairs through computational screening. Chem Sci 7:1257–1261. doi: 10.1039/C5SC03259H
234. Bernard S, Audisio D, Riomet M, et al (2017) Bioorthogonal Click and Release Reaction of Iminosydnones with Cycloalkynes. Angew Chemie Int Ed 56:15612–15616. doi: 10.1002/anie.201708790
235. Plougastel L, Koniev O, Specklin S, et al (2014) 4-Halogeno-sydnones for fast strain promoted cycloaddition with bicyclo-[6.1.0]-nonyne. Chem Commun 50:9376–9378. doi: 10.1039/C4CC03816A
236. Liang Y, Mackey JL, Lopez SA, et al (2012) Control and Design of Mutual Orthogonality in Bioorthogonal Cycloadditions. J Am Chem Soc 134:17904–17907. doi: 10.1021/ja309241e
237. Gordon CG, MacKey JL, Jewett JC, et al (2012) Reactivity of biarylazacyclooctynones in copper-free click chemistry. J Am Chem Soc 134:9199–9208. doi: 10.1021/ja3000936
238. Lopez SA, Houk KN (2013) Alkene Distortion Energies and Torsional Effects Control Reactivities, and Stereoselectivities of Azide Cycloadditions to Norbornene and Substituted Norbornenes. J Org Chem 78:1778–1783. doi:

10.1021/jo301267b

239. Liu F, Liang Y, Houk KN (2014) Theoretical Elucidation of the Origins of Substituent and Strain Effects on the Rates of Diels–Alder Reactions of 1,2,4,5-Tetrazines. J Am Chem Soc 136:11483–11493. doi: 10.1021/ja505569a
240. Karelson M, Maran U, Wang Y, Katritzky AR (1999) QSPR and QSAR Models Derived Using Large Molecular Descriptor Spaces. A Review of CODESSA Applications. Collect Czechoslov Chem Commun 64:1551–1571.
241. Katritzky AR, Karelson M, Petrukhin R (2005) CODESSA Pro.
242. Bader RFW (1994) Atoms in Molecules: A Quantum Theory. Clarendon Press, Oxford
243. Katritzky AR, Karelson M, Petrukhin R CODESSA 3 program.
244. Geerlings P, De Proft F, Langenaeker W (2003) Conceptual Density Functional Theory. Chem Rev 103:1793–1874. doi: 10.1021/cr990029p
245. Stewart JJP (2011) MOPAC2011.
246. Parr RG, Szentpály L v., Liu S (1999) Electrophilicity Index. J Am Chem Soc 121:1922–1924. doi: 10.1021/ja983494x
247. Domingo LR, Aurell MJ, Pérez P, Contreras R (2002) Quantitative characterization of the global electrophilicity power of common diene/dienophile pairs in Diels–Alder reactions. Tetrahedron 58:4417–4423. doi: https://doi.org/10.1016/S0040-4020(02)00410-6