Available online at www.sciencedirect.com

ScienceDirect

Focus Article, *Mendeleev Commun.*, 2021, **31**, 769–780

Mendeleev Communications

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

Timur I. Madzhidov,*[a] Assima Rakhimbekova,[a] Valentina A. Afonina,[a] Timur R. Gimadiev,[b] Ravil N. Mukhametgaleev,[a] Ramil I. Nugmanov,[a] Igor I. Baskin[c] and Alexandre Varnek[b,d]

[a] *A. M. Butlerov Institute of Chemistry, Kazan Federal University, 420008 Kazan, Russian Federation. E-mail: Timur.Madzhidov@kpfu.ru*

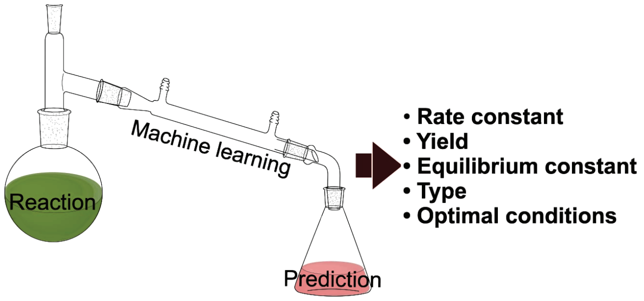
[b] *Institute for Chemical Reaction Design and Discovery, Hokkaido University, 001-0021 Sapporo, Japan*

[c] *Department of Materials Science and Engineering, Technion – Israel Institute of Technology, 3200003 Haifa, Israel*

[d] *Laboratory of Chemoinformatics, UMR 7140 CNRS, University of Strasbourg, 67000 Strasbourg, France*

DOI: 10.1016/j.mencom.2021.11.003

The synthesis of the desired chemical compound is the main task of synthetic organic chemistry. The predictions of reaction conditions and some important quantitative characteristics of chemical reactions as yield and reaction rate can substantially help in the development of optimal synthetic routes and assessment of synthesis cost. Theoretical assessment of these parameters can be performed with the help of modern machine-learning approaches, which use available experimental data to develop predictive models called quantitative or qualitative structure–reactivity relationship (QSRR) modelling. In the article, we review the state-of-the-art in the QSRR area and give our opinion on emerging trends in this field.



Keywords: chemoinformatics, reaction informatics, chemical reaction, reaction yield, reaction rate, reaction conditions, QSAR, QSPR, QSRR.

1. Introduction

Synthesis of the desired chemical compound is the main task of synthetic organic chemistry. Finding an optimal synthetic route requires vast experience, deep knowledge of reaction types, selectivity, and efficiency. The development of a synthetic plan requires answering some quite complicated questions:

(a) What is the optimal order of structural transformations that leads to the desired compound?

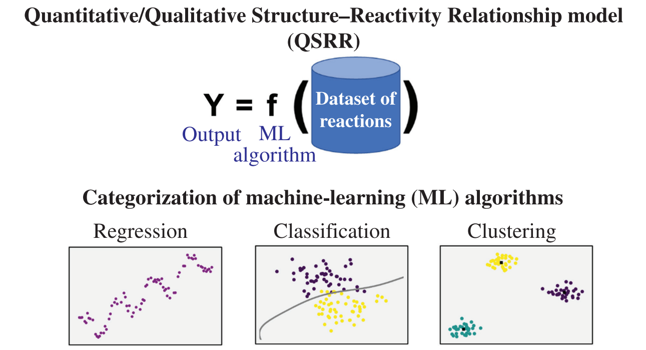
(b) Which are optimal experimental conditions for each step in the route?

(c) What is the expected yield/rate/selectivity for each reaction step?

Low yield or poor selectivity at even one step may cause the entire synthetic route to be discarded. Therefore, the development of models able to predict such reaction characteristics as yield, rate or selectivity is as important as route assessment.

Since seminal work by E. Corey,[1] many attempts were achieved to create computer-aided synthesis design (CASD) systems that find the synthetic pathway leading to the desired molecule. CASD systems should explore, pass through vast spaces of possible synthetic routes and find optimal solutions. Despite the variety of computational retrosynthetic tools reported in the literature so far, they have not come into synthetic chemists' everyday practice. The reason for such failure can be explained by their low computational efficiencies and the small number of considered reaction types.[2] In most cases, suggested by computer synthetic pathways could easily be determined by chemists, thus making earlier CASD tools useless from a practical perspective. Starting from 2016, this situation began to change rapidly thanks to the application of artificial intelligence technologies in chemistry,[3–7] which led to the renaissance of CASD and reaction informatics.

Despite the tremendous success, none of the existing computational CASD tools has embedded models that predict reaction conditions and some important quantitative characteristics of chemical reactions as yield and reaction rate. These reaction parameters are required to assess synthesis feasibility, which, in turn, is essential for ranking possible routes and finding optimal reaction conditions. Historically, theoretical assessments of these characteristics were performed by quantum chemistry or Hammett[8] σ–ρ-like analysis, which is considered an integral part of physical organic chemistry.[9,10] Nowadays, because of the availability of experimental data in chemical reaction databases, such characteristics can be assessed using machine-learning (ML) models which reveal relationships between structural representations of reactions (usually


Quantitative/Qualitative Structure–Reactivity Relationship model (QSRR)
Y = f (Dataset of reactions)
Output
ML algorithm
Categorization of machine-learning (ML) algorithms
Regression
Classification
Clustering

Output	Real value	Class	Cluster
Dataset needs	Samples with real-valued ouptups	Labelled samples	Unlabelled samples
Learning	Supervised (knowledge of target value)	Supervised (knowledge of label)	Unsupervised (no knowledge of label)

Figure 1 Schematic representation of Quantitative Structure–Reactivity Relationship (QSRR) model and classical machine-learning concepts.

represented by descriptor vectors) and their continuous (numerical value) or nominal (class label) properties based on the dataset of known pairs ‘reaction–property’. This is shown schematically in Figure 1. The model is further tested on new data with known properties to assess model performance and robustness. It is called the validation procedure. Hereafter, the models for prediction of reaction characteristics we call Quantitative or Qualitative Structure–Reactivity Relationship (QSRR).

Here, we overview the state-of-the-art application of chemoinformatics and machine learning to model various characteristics of chemical reactions (QSRR) and discuss some prospects in this scientific field. The article is arranged as follows: first, we consider different computer representations of reactions and provide a brief overview of data sources. Then, quantitative (reaction rate constant, yield, and equilibrium constants) and qualitative (reaction conditions or reaction type) machine-learning models are considered. Finally, we give our vision of the perspective of computational tools in synthetic chemistry.

2. Historical survey of structure–reactivity modelling

The first approach to predicting quantitative characteristics of chemical reactions was based on the Linear Free Energy Relationships (LFERs) and was developed by Hammett,[8] Taft,[11] and Palm[12] in the 1930s–1960s. LFERs[13] were extensively used to establish simple linear correlations between substituent or solvent descriptors and chemical reactivity. This approach has been used to study reaction mechanisms, develop synthetic routes, and estimate the biological activity of various organic chemicals.[14,15] However, the utility of this approach is rather limited because: (1) it could only be applied to relatively small congeneric datasets of compounds with the same core[16–18] or the same reaction proceeding in different solvents;[19] (2) it used experimentally measured substituents parameters as descriptors; and (3) as a rule, linear correlations were considered. This article intentionally ignores early LFER approaches since they were thoroughly studied and reviewed in the literature[9,13,20,21] and mostly have historical value or limited applicability.

To overcome the limitations of LFER, new modelling techniques based on the application of different types of molecular descriptors in combination with various machine-learning techniques were developed in the early 1990s.[22] One of the first approaches to quantitative reactivity prediction beyond LFER using a nonlinear machine learning method (neural networks) was published by Halberstam *et al.*,[23] who modelled the reaction rate constant. The major novelty of the proposed approach in comparison with LFER was an ability to predict reaction rates for different reactants/products combinations (hereafter, called transformation) under different reaction conditions. Since then, special descriptors adapted for chemical

Timur I. Madzhidov is a senior researcher, PhD, team leader in Laboratory of Chemoinformatics at the Kazan Federal University, associate professor of the Department of Organic Chemistry of Alexander Butlerov Institute of Chemistry. The area of his scientific interest is chemoinformatics, computational and quantum chemistry, reaction informatics, and AI applications in chemistry.

Assima Rakhimbekova is a postgraduate student and a junior researcher at the Department of Organic Chemistry of the Alexander Butlerov Institute of Chemistry, Kazan Federal University. The area of her scientific interest is chemoinformatics, machine learning, chemical reactions.

Valentina A. Afonina is a postgraduate student at the Department of Organic Chemistry of the Alexander Butlerov Institute of Chemistry, Kazan Federal University. The area of her scientific interest is chemoinformatics, machine learning, and condition prediction.

Timur R. Gimadiev is a postdoctoral researcher at the Institute for Chemical Reaction Design and Discovery (Hokkaido University, Japan), PhD. The area of his scientific interest is chemoinformatics, machine learning, chemical reactions, chemical databases, organic chemistry, computational chemistry, and big data.

Ravil N. Mukhametgaleev is a postgraduate student at the Department of Organic Chemistry of the Alexander Butlerov Institute of Chemistry, Kazan Federal University. The area of his scientific interest is chemoinformatics, and data science.

Ramil I. Nugmanov is a senior researcher, PhD, associate professor of the Department of Organic Chemistry of Alexander Butlerov Institute of Chemistry. The area of his scientific interest is chemoinformatics, computational chemistry, organic chemistry, and big data.

Igor I. Baskin is a Doctor of Science, a senior researcher at the Technion – Israel Institute of Technology. The area of his scientific interest is chemoinformatics, computational chemistry, electrochemistry and materials science, and machine learning.

Alexandre Varnek is a Doctor of Science, Professor at the University of Strasbourg, and Head of the Laboratory of Chemoinformatics at the University of Strasbourg (France). He is also Head of the Chemoinformatics group at ICReDD, Hokkaido University (Japan). The area of his scientific interest is chemoinformatics, computational chemistry, and molecular modelling.

reactions were proposed, among which the difference fingerprints and fragments issued from Condensed Graph of Reaction were the most successful.

Until recently, Quantitative Structure–Reactivity Relationship models were mostly built using in-house datasets manually annotated or collected using the high-throughput technique. In 2013, the first version of the USPTO database[24] collecting organic chemical reactions extracted by text-mining from US patents was released. An updated and sufficiently larger version of this database was released in 2016. In 2016, the Elsevier company, the owner of Reaxys®, shared this database with several scientific groups (including ours) in the framework of Reaxys R&D collaboration aiming at developing novel techniques of reaction mining. The emergence of sufficiently large datasets together with the development of deep learning techniques[25] stimulated the development of powerful retrosynthesis approaches,[3,4] forward (major product) prediction,[26,27] optimal condition selection,[28,29] and AI-based novel reaction discovery techniques.[30] These approaches attracted much attention to chemical reactions mining and led to the renaissance of the reaction informatics area.[2,6,31] Combining such AI-driven approaches with automatic experimentation techniques is an important step toward the development of an entirely self-sufficient robochemist.[32,33]

3. Reaction representations

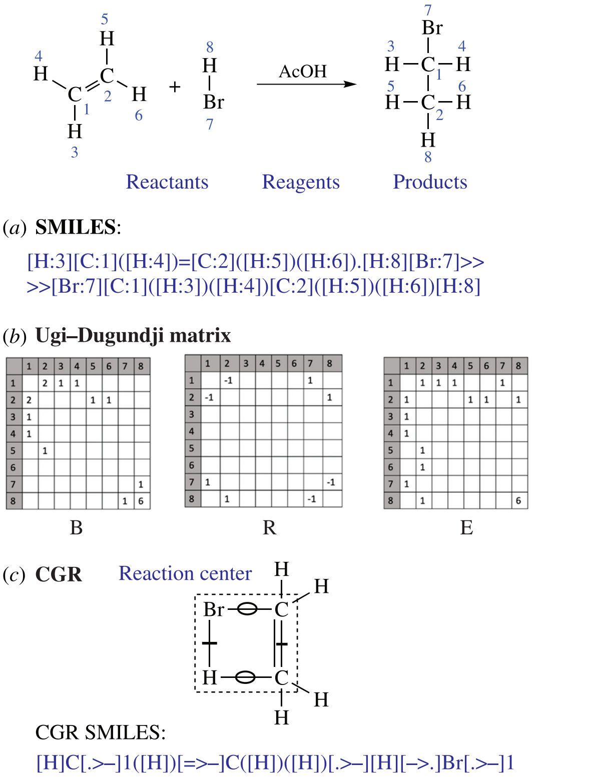
Computer representation of chemical reactions is an important step to process information for further analysis and modelling. For clarity, we will separately discuss structure-based and descriptor-based representations of reactions. The former is required to store or process structural information (molecular graphs of reactants and products), while the latter considers a reaction as a vector in multidimensional descriptor space.

3.1. Structural representation of reactions

In chemoinformatics, molecules can be represented by molecular graphs,[34] which, in turn, can be stored as text strings,[35] matrices, tables.[36] Alternatively, for some tasks, 3D representation[37] of molecules can be more suitable. Encoding chemical reactions is much more complex than that of single structures since reaction description includes several chemical compounds: reactants, reagents, products. Currently, three critical approaches to the representation of chemical reactions can be distinguished: (1) using representations of reactant and product molecules, (2) product–reactant difference, (3) reaction centre-based representation.

The most straightforward way of representing a chemical reaction is described as a set of all (relevant) molecular entities: reactants and products. The easiest and widespread example of the representation is the Reaction SMILES,[38] which captures all involved molecules as SMILES string representations[35] [Figure 2(*a*)]. The standard exchange file of reaction datasets, RDF file format,[36] also represents reactions as a set of molecules in MOL format, together with their associated data. The representation is suitable for storing the reactions in databases and for calculating descriptors for reactions.

Structural changes caused by reaction could be detected by subtracting features of reactants from the corresponding features of products. This idea underlies the Ugi–Dugundji matrix formalism,[39] according to which structures of all molecules involved in the reaction are described by a bonds–electrons matrix of reactants (B) and products (E), while the reaction is described as the difference between these matrices (R-matrix) [Figure 2(*b*)]. Although this representation was used in the computer planning of synthesis[40] to predict reaction pathways[41,42] and to search for the shortest distance between the reactant and product, it is not convenient for storing and searching chemical reactions in databases and modelling reaction characteristics.



(*a*) **SMILES**:

[H:3][C:1]([H:4])=[C:2]([H:5])([H:6]).[H:8][Br:7]>>
>>[Br:7][C:1]([H:3])([H:4])[C:2]([H:5])([H:6])[H:8]

(*b*) **Ugi–Dugundji matrix**

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

(*c*) **CGR**

CGR SMILES:

[H]C[.>–]1([H])[=>–]C([H])([H])[.>–][H][–>.]Br[.>–]1

Figure 2 Different computer representations of reactions: (*a*) reaction SMILES, (*b*) Ugi–Dugundji matrices, (*c*) Condensed Graph of Reaction (CGR) and related SMILES/CGR. In CGR, broken and formed bonds are denoted by a circle and a crossed line, respectively.

In the reaction centre (RC) based approaches, chemical bonds (broken, formed and modified) and/or formal electron flows are specified. A reaction centre assembles a set of atoms associated with the bonds changed during the reaction (formation, cleavage, and bond order changes). It can be identified using the atom-to-atom mapping (AAM) procedure.[43] The RC-based methods are particularly useful for model-driven reaction classification (see Section 6). For example, in the Condensed Graph of Reaction (CGR) approach,[44] a reaction is represented by a single molecular graph, described by both conventional chemical bonds (single, double, aromatic, *etc.*) and so-called dynamic bonds characterizing changed/broken/formed chemical bonds [see Figure 2(*c*)]. Similarly, the changes of atomic charges can be introduced by special dynamic atoms.[45] So far, CGR method has been successfully used for reactions storage,[30,45] search,[28,44,46] analysis,[2,47,48] visualization,[49] and modelling.[50–52] To handle CGRs manipulation, the CGRtools library was developed.[45]

3.2. Reaction descriptors

Chemical reactions represent a complex object because they involve several molecular species of two types (reactants and products) and their properties depend on experimental conditions (solvent, catalyst, temperature, *etc.*). Therefore, reaction descriptors have to include information about reactants, products and small molecule reagents/catalysts, *i.e.*, should reflect the occurring chemical transformations and condition information. There exist three methodologies for computing reaction descriptors: (*a*) reaction centre-free descriptors, and descriptors with (*b*) implicit and (*c*) explicit consideration of a reaction centre (Figure 3).

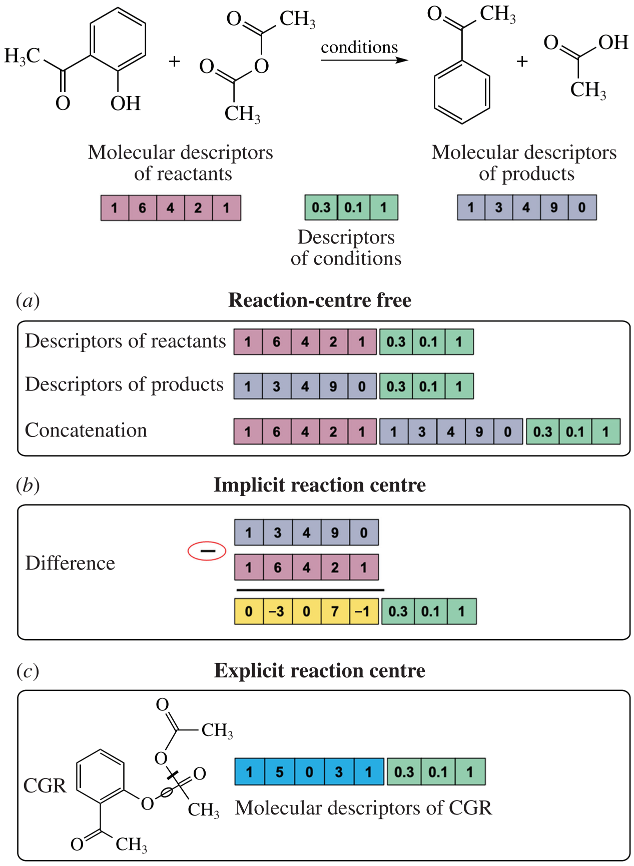

H_3C
O
OH
+
O
CH_3
conditions
Molecular descriptors of reactants
Molecular descriptors of products
1 6 4 2 1
0.3 0.1 1
1 3 4 9 0
Descriptors of conditions
(a) Reaction-centre free
Descriptors of reactants
Descriptors of products
Concatenation
(b) Implicit reaction centre
Difference
–3 0 7 –1
(c) Explicit reaction centre
CGR
1 5 0 3 1
Molecular descriptors of CGR

Figure 3 Descriptors of chemical reaction: (*a*) RC-free descriptors considering molecular descriptors of reactants or products or their concatenation, (*b*) RC-implicit descriptors considering subtraction of products and reactants descriptors, and (*c*) RC-explicit descriptors issued from CGR. The descriptor vectors, marked in pink, purple, and blue, characterize the molecular descriptors of the reactants, products, and CGR. The vector of descriptors, marked in green, characterizes the reaction conditions: additives, temperature, solvents, amounts of reagents, *etc*. The numbers in the figure are illustrative.

In the RC-free methodology, descriptors of reaction are molecular descriptors of reactants and products [Figure 3(*a*)]. The approach is straightforward but applicable only if the reaction type is always the same throughout the dataset. Optionally, descriptors of reaction conditions, like temperature, solvent descriptors, *etc*., can be added. This approach was used in LFER[11,18] and some early publications.[23,53,54] Recently, Sandfort *et al*.[55] used concatenations of 24 types of fingerprints encoding reactants to predict enantioselectivity and yield of reactions. A simple concatenation of descriptors of products and reactants, [Figure 3(*a*)], was used to assess optimal conditions for the Michael reaction[56] and to model the rate constant of E2 reactions.[57]

Popular implicit reaction centre (RC-implicit) descriptors are calculated as a difference between descriptors vectors of product and reactant [Figure 3(*b*)]. These descriptors identify changes in the amount of particular molecular fragments, but not the atoms taking part in reaction centres. The approach has been applied in different reaction classification tasks,[58] enzyme classification based on reactions it catalyzes,[59–61] to predict genome-scale enzyme–metabolite, and drug–target interaction,[62] to predict the potential metabolites of a given parent structure.[63] Besides, such descriptors demonstrated a good performance in comparison with other approaches.[56,57] In principle, RC-free and difference reaction descriptors' calculations require perfectly balanced reactions. Nonetheless, Schneider *et al*.[58] demonstrated that difference fingerprints prepared for unbalanced reactions performed well in classification tasks. Unlike RC-explicit descriptors, the RC-free and RC-implicit descriptors do not need an error-prone and time-consuming atom-to-atom mapping step.

Reaction-centre explicit descriptions can be calculated for Condensed Graph of Reaction (CGR). Usually, SiRMS[57,64] or ISIDA[46,66] descriptors are used for CGR encoding as numerical vector [Figure 3(*c*)]. Such descriptors were used in numerous QSRR studies[28,51,52,56,57,67–70] (see more details in Section 5). Explicit representation of the reaction centre [Figure 3(*c*)] has particular benefits because it allows separating descriptors for both RC and unchanged parts of the molecule, which potentially enhances the predictive ability of models. Notice that RC-explicit descriptors are insensitive concerning molecule ordering and lost reactants/products.

Since kinetic and thermodynamic characteristics depend on experimental conditions (temperature, pressure, reactant concentration, and solvent), related descriptors should complement descriptors characterizing chemical transformations. Temperature and pressure can be used as descriptors. Solvents can be encoded by experimentally measured their dipole moment, refraction index, dielectric permittivity, Catalan SPP,[71] SA, and SB constants,[72–74] Kamlet–Taft α, β, and π* constants,[75–77] some of their physicochemical parameters,[23,50,78] or other descriptors types used to represent molecules (*e.g*., fragment descriptors).[23]

So far, a comparison of different reaction descriptors was performed on a very limited scale. Thus, Polishchuk *et al*.[57] benchmarked mixture SiRMS descriptors and ISIDA fragments generated for CGR on a dataset for reaction rate of E2 reaction. Skoraczyński *et al*.[79] compared different descriptors for yield prediction, but the dataset was noisy and they obtained rather poor results. Since no thorough benchmarking analysis of various reaction descriptors has been performed, we decided to compare RC-free, RC-implicit and RC-explicit descriptors on the datasets of balanced reactions annotated with rate constants of S_N2,[50] E2,[52] Diels–Alder[69] or equilibrium constants of tautomerization reactions.[70] Descriptor vectors were complemented with experimental conditions (solvent parameters and temperature) as described.[51] The rigorous cross-validation strategy ('transformation-out')[80] has been applied to provide a realistic assessment of the models' performance. Benchmarking results show that reactant–product concatenation and CGR-based descriptors are generally the most performant for most databases (see Figure 4, the benchmarking methodology is described in Online Supplementary Materials).

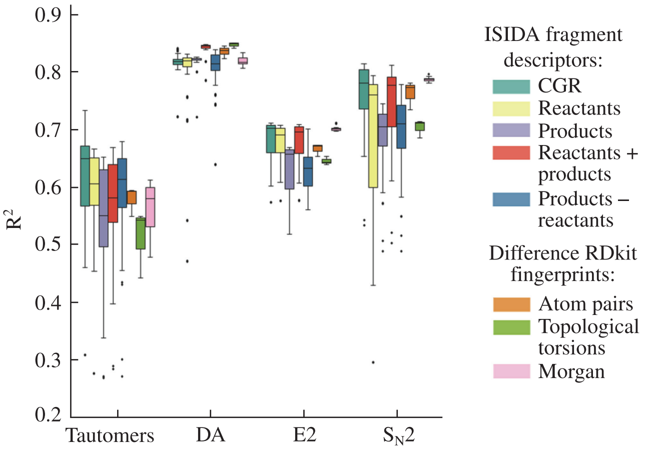


Figure 4 Determination coefficient of the models for rate constants of Diels–Alder (DA), E2 and S_N2 reaction and equilibrium constant of tautomeric equilibria in 'transformation-out' validation[80] as a function of descriptors types. Note that external test set is comprised of reactions with new combinations of reactant–product, absent in training set. ISIDA or RDkit fragment descriptors varied by their topology (paths of atoms and bonds) or sizes.

Table 1 Reaction datasets.[a]

Database name	Number of reactions	Vendor	Availability	Contents
Reaxys[82]	>54 mln	Elsevier	Commercial	S, M, C
CASREACT[81]	~134 mln (from 1840)	American Chemical Society	Commercial	S, M, C
SPRESI[84]	>4.6 mln (1974–2014)	Deepmatter	Commercial	S, C
Pistachio[87]	>9 mln	NextMove Software	Commercial	S, C
ChemInform Reaction Library[93]	~2 mln (1990–2016)	ChemInform (Wiley)	Commercial	S, C
USPTO reaction dataset[24]	~4 mln (1976–2016)	–	Open access	
MIT-400K,[88,94] USPTO derivative	~400 K	–	Open access	
USPTO-50K,[89] USPTO derivative	~50 K	–	Open access	C

[a] S – searchable reactions and their components, M – multistep reactions annotation, C – reaction class annotation.

4. Reaction datasets

The size of chemical reaction data is very big: the largest CASREACT[81] contains more than 130 M reactions, followed by Reaxys[82,83] and SPRESI[84] containing 55+ and 4.6 M reactions, respectively (Table 1). This amount is quickly growing up. Thus, scientific literature analysis platform SciVal[85] reports that about 340 K papers were published in organic chemistry in the last 5 years only! Modern text-mining technologies have been used to extract chemical data directly from the patents to compose USPTO[24,86] and Pistachio[87] reaction databases, containing 4 and 9 M reactions, respectively. USPTO is the only open-source reaction information existing to time. It was used to prepare subsets of USPTO, MIT-400K[88] and USPTO-50K[89] intensively used in various modelling studies. Recently, the ChEMU[90] challenge has been announced to facilitate the development of novel text-mining approaches of reaction data extraction from the patents.

The commercial and open reaction databases listed in Table 1 contain structural representations of reactions and experimental conditions. Among reaction characteristics, the only yield is annotated. There exist no databases collecting kinetic and thermodynamic parameters of chemical reactions, which seriously limits the development of predictive models. In the absence of ready-to-use data, modelling datasets are annotated from literature manually or using high-throughput experimentation technique.[91] We collected the largest database of kinetic reaction characteristics by the digitalization of reference books by Palm.[92] At present, it contains more than 15 000 reaction rates and equilibria constants for different reaction types, which have been used for modelling.[50,52,69,70]

5. Quantitative reaction–property modelling

5.1. Reaction rate constant

Rate reaction constants are probably one of the most important characteristics of chemical reactions. In principle, reaction yield, regio- and enantioselectivity, equilibrium constant can be calculated if the reaction rate constant is known. On the other hand, the reaction rate constant is one of the most complicated reaction characteristics to measure experimentally, requiring prior knowledge of the reaction mechanism.

One of the first attempts to relate reaction rate constants with substituent constants was made by Hammett,[95,96] McDuffie and Dougherty.[97] From the practical point of view, the logarithm of the rate constant ($\log k$) is more useful and theoretically supported by LFER ideology (it linearly related with the free energy of transition state); thus, all models mentioned below actually predict the $\log k$ value. Since then, QSRR modelling of reactions rate constants was performed using linear regression on homogeneous series, either varying the reactants under fixed reaction conditions (solvent, temperature),[16,17,98] or inversely varying conditions to study a reaction between given reactants.[19,99] One of the first studies performed on a large and diverse reaction dataset was reported by Halberstam *et al.*[23] who used artificial neural networks and quantum chemical descriptors to predict rate constants of acid hydrolysis of arbitrary esters under various conditions. To represent reaction conditions, temperature and Koppel–Palm descriptors of solvents[100] were concatenated with structural descriptors. Reasonable performance on the test set was achieved (RMSE = 0.34 $\log k$ units). Using molecular fragments instead of QM descriptors to represent ester structure, Zhokhova *et al.*[101] slightly improved the model's performance (RMSE = 0.31 $\log k$ units). An accuracy of RMSE = 0.58 $\log k$ units on the test set of the models of reaction rate constant of S_N2 reactions proceeding in various solvents at different temperatures was reported by Kravtsov *et al.*[53] They used a descriptor vector combining temperature, local quantum chemical and fragment descriptors for both the nucleophile and electrophile molecules and Koppel–Palm solvent descriptors. Similar workflow applied to S_N1 reaction rate constant[54] resulted in RMSE = 0.61 $\log k$ units on a test set.

A series of models of reaction rate constants were built using ISIDA fragment descriptors[44,65] computed for Condensed Graph of Reaction. It concerns subgraphs of different sizes and topologies (sequence of atoms and bonds or atom-centred fragments).[65] Such descriptors were successfully used by Hoonakker *et al.*[102] in the modelling of the reaction rate of bimolecular nucleophilic substitution reactions (S_N2) in water. Later on, Madzhidov *et al.*[52] and Gimadiev *et al.*[50] considered S_N2 reactions in different solvents (including water-organic binary mixtures) using concatenated ISIDA/CGR fragment and a series of descriptors accounting for temperature and solvents: inverse absolute temperature, the molar ratio of organic solvent, polarity, polarizability, H-acidity and basicity: Catalan constants, Camlet–Taft constants, functions of the dielectric constant, functions of the refractive index. A similar approach was used for modelling kinetics of S_N2 substitution reaction by azide anion,[51] bimolecular elimination reactions[52] and Diels–Alder cycloaddition.[68,69]

In our recent publications[50,57,80] we pointed out a gap in QSRR methodology: conventional cross-validation (CV) scheme for model validation systematically returns overoptimistic model performance characteristics due to reactions proceeding under similar conditions (*e.g.*, same solvents and close temperatures). To address this issue, 'transformation-out' and 'solvent-out' CV schemes have been suggested.

Gimadiev *et al.*[50] and Rakhimbekova *et al.*[103] have raised a problem of the assessment of applicability domain (AD) of QSRR models. An AD is supposed to identify reactions that differ from the training set ones; predictions on the objects outside AD are considered unreliable. It has been shown that simple filters effective for molecular datasets[104] can successfully be applied as AD of models for chemical reactions.

5.2. Reaction yield

Prediction of a reaction yield can help synthetic chemists to select an optimal synthetic route, optimize the efficiency and 'greenness' of the multistep synthetic procedure using a selection of proper conditions and estimate costs of the synthesis. So far, quantitative prediction of yields has been very little explored, primarily due to the lack of available data.

Yields of chemical reactions are known to be quite noisy and subjected to notorious poor reproducibility being a function of the experience and accuracy of chemists. According to the survey of 106 chemists, more than 85% of them were faced with problems of reproducibility of someone else results and about 63% had difficulties with the reproduction of their results.[105] About 8% of submitted to Organic Synthesis journal publications are rejected since yield or selectivity cannot be reproduced within a reasonable range in the laboratory of one of the editors.[105]

For this reason, most of the QSRR models for the prediction of reaction yields were built on rather small and specially collected datasets of reactions of the same type: Buchwald–Hartwig aminations,[55,91,106,107] Suzuki–Miyaura cross-coupling,[106–109] Negishi reactions,[106] deoxyfluorination of alcohols.[110]

Ahneman *et al.*[111] applied the high-throughput experimentation (HTE) technique for collecting a dataset of palladium-catalyzed Buchwald–Hartwig C–N cross-coupling reactions, containing 4140 data points obtained by full enumeration of reactions of 15 aryl and heteroaryl halides, 4 Buchwald ligands, 3 bases, and 23 isoxazole additives. The yield was predicted by the Random Forest method using DFT-calculated electronic, atomic, and vibrational mode descriptors. The model displayed a good performance on a 70/30 train-test random split (RMSE = 7.8 and R^2 = 0.92, average over ten random divisions). The authors also attempted to interpret the data obtained with experimental NMR data and thus to show how ML can be used to give insight into reaction mechanisms. Later on, this study was criticised by Chuang and Keiser[112] for improper computational design. In particular, they demonstrated that replacement of the DFT-calculated chemical features with a random vector of the same length leads to similar model performances. Nonetheless, the dataset from Ahneman *et al.*[111] was actively used in other studies[55,106,107] where different techniques for reaction representation were proposed and compared.

Granda *et al.*[109] reported a neural network-based model that predicted Suzuki–Miyaura reaction yields collected using HTE technique.[113] They obtained a very small error (10%) in yield prediction in both retrospective and prospective tests. Moreover, the model trained on 10% of data was used as a 'brain' of a chemical robot that can perform chemical operations and was applied to explore chemical space to find reactions with high yield.

Huerta *et al.*[106] used the only 0D-, 1D- and 2D-structural descriptors to build classification using the Random Forest model to predict yields of 48 900 Suzuki–Miyaura cross-coupling reactions, 3300 Negishi and 26 500 Buchwald–Hartwig reactions extracted from Reaxys. High (>60%) and low (<40%) yield outcomes were predicted with good quality.

Fu *et al.*[108] used quantum mechanics-based reaction descriptors and deep neural networks to establish relationships between the chemical contexts (reactants and precatalysts), reaction conditions and product yields aiming to determine the most efficient reaction conditions. The model was applied to the 387 Suzuki–Miyaura cross-coupling reactions[114] to find the best catalysts for given reactants and, at the same time, to discover the most favourable catalyst loading and reaction temperature. The model demonstrated good performance on the external validation set (RMSE ≈ 9%).

For now, rare attempts to predict reaction yields on a large-scale, diverse reaction database failed to achieve reasonable performance. Thus, Skoraczyński *et al.*[79] used a dataset of 425 000 reactions from Reaxys to model yields and reaction duration. To simplify the task, predicted values were binarized into 2 classes, high (value >65%) and low yields. The developed classification model achieved an accuracy of 65% which is slightly better than random prediction. Such mediocre prediction quality was interpreted as a consequence of the imperfectness of descriptors,[79] but one should keep in mind ignorance of reaction conditions and a high level of noise in yields values too. Besides, Schwaller *et al.*[107] showed that the distribution of yields reported in the USPTO dataset varies as a function of the reaction mass scale, *i.e.* the amount of isolated product (milligrams or grams). They used complex transformer-based[115] models, which were first trained to predict reaction products from reactants. Then reaction embeddings, called reaction fingerprints,[116] were extracted and applied for yield modelling. The proposed Yield-BERT approach worked well on the above-mentioned Suzuki–Miyaura[113] and Buchwald–Hartwig[111] datasets but failed to accurately learn patent reactions' yields: R^2 achieved 0.117 for gram scale and 0.195 for milligram-scale synthesis because of intrinsic lack of consistency and quality in the patent data.

5.3. Reaction equilibrium constant

Equilibrium constants are not common reaction characteristics to model using QSRR modelling workflow because of the following reasons: (i) they can be estimated using regular quantum chemical calculations, (ii) an equilibrium constant defines reaction outcome for thermodynamically controlled reactions; however, reaction kinetics is, usually, more important, (iii), the problem of low yield is often related to the formation of by-products but not unfavourable thermodynamics.

Prototropic equilibria – acidity and tautomerism – are important equilibrium processes extensively studied in Quantitative Structure–Activity Relationships (QSAR)[117–119] or by quantum chemical studies.[120–122] Tautomeric equilibrium constants can be calculated using acidity values of each tautomeric form:

$$\log K_\mathrm{T} = \mathrm{p}K_\mathrm{a}(\text{tautomer 2}) - \mathrm{p}K_\mathrm{a}(\text{tautomer 1})$$

where K_T is tautomerization equilibrium constant, pK_a – acid dissociation constant for left (tautomer 1) and right-hand-side tautomer (tautomer 2) in equilibrium. Different quantum chemistry,[123–127] and QSAR[128,129] approaches were used to assess equilibrium constant or tautomer distribution using predicted pK_a values of each tautomer. Gimadiev *et al.*[130] were the first who treated tautomer equilibria as a reaction and attempted to predict logK_T directly. They achieved good prediction performance employing fragment descriptors based on CGR representation of equilibrium reaction equation. The proposed model outperformed results obtained in QSAR[128] and quantum chemical studies. Zankov *et al.*[131] developed a conjugated learning approach that linked the modelling workflows for the tautomeric equilibrium constant and acidity, and, thus, predicted tautomeric distribution and acidity of molecules simultaneously. It was found that the proposed approach correctly predicted the acidity of minor tautomers, which is hardly possible in regular QSAR acidity modelling.

5.4. Reaction enantioselectivity

Enantioselectivity is a parameter which characterizes reactions leading to the formation of new chiral centres. The computer-aided design of catalysts leading to high enantioselectivity is an important task due to the high cost of experiments. Earlier, machine learning approaches were used to establish Quantitative Structure–Selectivity Relationship (QSSR)[132] that relate catalyst structure with its selectivity, see recent review by Zahrt *et al.*[133] and references therein.

In most of publications in the QSSR area, the selectivity of different catalysts applied to one particular reaction was considered. In this case, reaction structure can be ignored upon the model building. Since organic molecules used as catalysts for enantioselective catalysis are chiral, an information extracted from their 2D graphs may not be sufficient for the modelling. For this reason, most of QSSR studies were carried out using 3D QSAR techniques[134,135] and 3D descriptors. In CoMFA[136] and other alignment-based 3D QSAR approaches, a catalyst molecule is placed in the rectangular box followed by calculations of van der Waals, electrostatic or other fields (so-called Molecular Interaction Fields) in the nodes of the rectangular grid superposed with the box. These energy values are used as descriptors in machine-learning models. To achieve the descriptors consistency, all considered molecules must be aligned to some frame. The advantage of 3D QSAR approaches is an ability to build a model on a rather small dataset and to interpret the modelling results in terms of physical interactions responsible for catalysts' activity. Lipkowitz and Pradhan[137] applied the CoMFA[136] approach for establishing a relation between enantiomeric excesses in the Diels–Alder reaction catalyzed by 23 copper-containing compounds and their structure. Kozlowski *et al.*[132] used 3D structure of the transition state of the catalyst–reactant complex in order to model the relative free energy of concurrent reactions ($\Delta\Delta G$) calculated based on an experimental ratio of stereoisomers instead of enantiomeric excesses.

In order to exclude the catalysts' alignment step, which requires some manual intervention and introduces some noise in the model, the alignment-free GRIND[138] approach was applied in the modelling of enantioselectivity of Diels–Alder, addition and reduction reactions[139] as well as for the analysis of substrate influence on the enantioselectivity of hydrolysis reaction catalyzed by lipases.[140] Becides 3D QSAR approach, some specific descriptors of the reaction centre,[141,142] quantum chemical descriptors[143,144] or their combinations[145] were also used for enantioselectivity modelling. Note that conventional 3D QSAR approaches rely only on one conformer per catalyst molecule which may significantly deteriorate the model performance. Indeed, Melville *et al.*[146] observed an increase of the predictive ability of the models when a Boltzmann averaging of molecular interaction fields over conformation ensembles was considered. In order to account for conformational ensembles of catalysts, Zahrt *et al.*[147,148] proposed special steric descriptors (called 'average steric occupancy', ASO). The ASO descriptors reflect both conformational flexibility of catalyst and steric hindrance induced in a particular region of space.

Significant progress in the enantioselectivity modelling has been achieved in the study by Zahrt *et al.*[147,148] who considered simultaneously different catalysts and different reactions. They reported a Random Forest[149] model built on a set of 1075 experimental enantioselectivities of 25 aza-Michael addition reactions catalyzed by 43 chiral phosphoric acids. For each catalyst/reaction pair, the ASO descriptors calculated for catalyst, electrophile and nucleophile participating in a reaction were concatenated with special electrostatic descriptors of substituents and NBO charges.

This dataset was used in subsequent works where alternative approaches of reaction-wide enantioselectivity modelling were proposed. Xu *et al.*[150] proposed spherical projection descriptors of molecular stereostructure (SPMS) calculated for 20 conformations of catalyst, nucleophile and electrophile which fed different channels to convolution neural network. Zankov *et al.*[151] represented reaction structure using CGR-based fragment descriptors whereas each conformer of a catalyst molecule was represented by stereochemically sensitive intermolecular interaction descriptors.[152] Entire descriptor vector resulted from concatenation of the above descriptors fed multi-instance learning neural network accounting for multiple conformations.[153] Both approaches[150,151] performed similarly to the Zahrt *et al.* model[147] but they were alignment-free and required no human intervention.

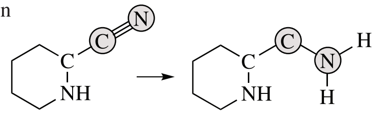
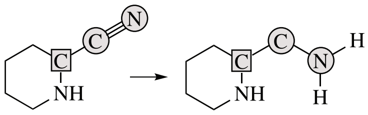
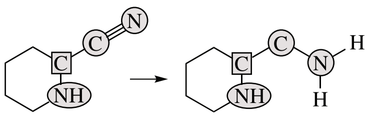
6. Classification modelling

6.1. Reaction type

Classification of reaction types is an important task in many different applications, such as searching for similar reactions, optimal condition selection, or synthesis design. The approaches to classify reaction by types can be divided into 'model-based', 'rule-based', and 'data-based' methods. The model-based methods use predetermined representations of the reaction centre, *i.e.*, formal bond redistribution schemes.[154] Different model-based classification approaches were proposed by Balaban,[155] Dugundji and Ugi,[39] Hendrickson,[156,157] Arens,[158] Zefirov and Tratch,[159,160] and Fujita;[161,162] they differ by representations of the reaction centre. ICClassify approach is one of the most recent model-based classification systems that found its utility in reaction classification or similarity search.[163] It represents a reaction by three levels of hash-codes reflecting reaction centre with the different environment – broad, medium, and narrow (Table 2). At the broad level, only atoms of the reaction centre are considered, at the medium level it also includes the first environment atoms (α-atoms to reaction centre), at the narrowest level it also includes β-atoms (see Table 2). Broad, medium and narrow transformations can be easily encoded by related motifs of Condensed Graph of Reactions and further used for a hash-code generation.[2,48] Obtained hash-codes can be used to perform unsupervised classification by grouping reactions of the same type.

The rule-based approaches are based on manually prepared structural rules which can be formalized by SMIRKS.[38] The popular RXNO reaction ontology is based on arranging similar types of reactions into groups following chemical logic.[164,165] RXNO ontology is implemented in the rule-based reaction classification NameRxn tool by NextMove.[166] This tool was applied to investigate the popularity of various types of reactions in medicinal chemistry.[167] Christ *et al.*[168] proposed a SMIRKS rule-based method for classification reactions and applied them to analyze reaction types from electronic lab notebooks.

Table 2 Levels of consideration of the reaction centre used in the ICClassify approach.[163]

Sphere	Level	Example
Zero	Broad (only the reaction centre atoms are included)	
First	Medium (the reaction centre atoms and α-atoms expect for hydrogens are included)	
Second	Narrow (the reaction centre atoms, α and β-atoms expect for hydrogens and consecutive sp^3-atoms are included)	

Data-based classification approaches use machine learning (ML) to predict reaction classes. Kohonen's self-organizing maps[169] was one of the first approaches used for this purpose. It has been shown that the objects falling into the same node of the map were mostly comprised of reactions of the same type. Similarity-based reaction classification was proposed by Sello and Termini.[170,171] Schwaller *et al.*[116] proposed a data-driven unsupervised classification of chemical reactions by applying BERT transformer-based neural networks trained to restore reaction products from reactants. It has been demonstrated that clusters obtained from BERT embeddings of reactions effectively group reactions of the same type.

The rule- and model-based approaches are often either too strict or make too many errors. Moreover, they ignore reaction conditions that in principle can be important to assign reaction types. For example, the amination of aryl halides in the presence of palladium catalyst (or other coupling catalysts) is called Buchwald–Hartwig reaction, the reaction in the presence of a copper catalyst is usually referred to as Goldberg reaction, and the amination of electron-deficient aryls in the absence of any catalyst is called aromatic nucleophilic substitution (S_NAr) reaction. Therefore, supervised classification of the reaction types using different ML methods represents their valuable alternative. Schneider *et al.*[58] applied supervised classification of the reaction types using different ML methods and concatenated structural and 'agent' fingerprints describing catalysts and solvents. The model classified USPTO reactions annotated according to RXNO ontology using NameRxn.[166] Interestingly that the classifier was able to correctly assign the reaction type when NameRxn failed. Using an in-house dataset containing reactions of 336 types, Ghiandoni *et al.*[172] applied the Random Forest classifier coupled with the Conformal Prediction[173] method to predict reaction type and to estimate the confidence of class assignment. Wei *et al.*[174] developed neural networks for predicting reaction classes (17 reaction types considered) using a concatenation of the fingerprints of the reactants and the reagents. This model is required to determine the SMIRKS pattern corresponding to a particular reaction.

6.2. Optimal reaction conditions

Significant progress in the development of retrosynthetic planning tools reinforces the interest in the assessment of optimal reaction conditions of a given one-step reaction. There exist two main approaches helping to assess optimal reaction conditions: *indirect* prediction using a surrogate model that links structure and conditions with some characteristics (Y = rate, yield, selectivity) and *direct* prediction when conditions are the model's primary outcome. In both approaches, conditions are represented by either a single entity (temperature or catalyst, or solvent) or their combination (*e.g.*, temperature and solvent).

In the first approach, a previously trained QSRR model linking a target property (Y) with some condition parameters X (*e.g.*, solvent descriptors) is used to determine X which maximizes Y. Struebing *et al.*[175] proposed an approach for the selection of solvent for Menshutkin (subtype of S_N2) reactions. Their approach involves building a surrogate linear model, approximating reaction rate dependency on solvent parameters, while the rate constant is calculated through quantum chemistry approaches. Reaction rates on some preselected solvents are calculated, then the best solvent is found based on model prediction for all the rest solvents, and the rate constant calculation is repeated for it. In turn, Fu *et al.*[108] used the QSRR model trained on HTE yield data from Jensen *et al.*[114] for optimal condition selection. They predicted reaction yield using quantum chemical descriptors of reagent and conditions (such as descriptors of catalyst used and its loading, temperature). It has been shown that the model efficiently selects optimal conditions and provides a quite good estimation of yield.

Models for direct conditions prediction are usually based on reactions belonging to the same type. Marcou *et al.*[56] applied the multiclass classification method to predict optimal catalysts and solvents for the Michael reactions. The performances of models only marginally depended on the descriptors used: ISIDA, MOLMAP, CDK, and EED descriptors calculated for reagents only, products only, or entire reaction. Lin *et al.*[28] applied a reaction similarity-based approach to predict optimal catalysts for deprotection reactions using a dataset of 150 K hydrogenation reactions extracted from Reaxys. A good performance was achieved on the external set containing substrates bearing several protective groups. However, this approach can provide only a general type of catalyst, like Pd-containing or Ni-containing, or Lindlar catalyst.

Friedel–Crafts, aldol addition, Claisen condensation, Diels–Alder, and Wittig reaction datasets extracted from the Reaxys database were used to model optimal solvent and catalyst.[176] Benchmarking of different ML techniques showed that the nearest neighbour approach is the most accurate and neural networks slightly less performant.

Gao *et al.*[29] reported a 'universal' approach to predict reaction conditions using some 10 million single-step reactions from the Reaxys database for the model training. The modelling procedure involved a deep neural network that recurrently predicted catalysts, solvents, reagents, and temperature. The nearest neighbour approach performed similarly well but was much slower and resource consuming. The trained neural-network model is available within the ASKCOS system.[177]

7. Concluding remarks

Reaction informatics has a long and rugged history. Synthesis design and reaction characteristics prediction approaches had ups and downs and entered into the renaissance era the last 3–5 years. The main reasons for this are the accumulation of synthetic data into large databases, the development of surprisingly efficient deep learning architectures to predict reaction products and the development of very efficient synthesis planning approaches, which, in turn, was caused by the development of efficient AI algorithms. This has led to the surge of interest in structure–reactivity modelling to predict reaction kinetic, thermodynamic properties, yield, and other reaction outcomes. Although this area is in its infancy, some progress has already been achieved, and some conclusions can be made.

Datasets. Data availability is the most important problem that limits the development of QSRR modelling. Currently, manually annotated databases are mostly commercial, whereas non-commercial databases have been created using automatic text-mining of patents and, hence, quite noisy. Kinetic and thermodynamic properties of chemical reactions are not annotated in commercial databases and are hence sporadic. Negative results are not recorded in chemical reactions: failed reactions with a very low rate or yield, poor selectivity are rarely published and annotated in databases. These data would greatly help develop yield and condition prediction models and are essential for generating reaction feasibility filters required for synthesis planning ('in-scope' filter in paper[4]). Data quality is still the most problematic QSRR modelling issue; even commercial datasets have many flaws, inconsistencies, and errors. Thus, the development of data curation procedures, like the one developed for molecular datasets, is critical. New large and high-quality open datasets of chemical reactions are required for further progress in the field. High-throughput experimentation techniques can also help in the collection of high-quality reaction data.

Descriptors. Many different approaches for descriptor representation of chemical reactions were proposed that either use information about reactants or products separately, combine them into one vector, or use reaction centre representations like CGR for generating descriptors. No benchmarking studies have been performed on a broad scale yet, but according to sporadic publications and our tests, we can conclude that different ways of reaction descriptor calculation have similar performance. CGR-based fragment descriptors and concatenation of reactant–product descriptors have shown the best performance across different balanced reaction datasets in our benchmark. At the same time, if fragment descriptors are applied for modelling, the fragmentation scheme and length of fragments influence drastically the model performance. Among others, difference fingerprints and CGR-based fragment descriptors look most widely used. Direct graph-based machine learning approaches, such as graph convolutional networks[178] and graph embedding schemes, are not widely used for QSRR studies; however, this approach seems to be quite promising. There is only one very recent work applying reaction SMILES for direct property (yield) prediction.[107] At the same time, ML approaches trained directly on SMILES or molecular graphs become quite popular for predicting molecular properties[179,180] and for *de novo* molecule generation.[181–185] In the field of reaction informatics, such approaches, however, are used almost exclusively for product prediction[186,187] or retrosynthetic transformation prediction.[188,189] Limitation of such approaches to use rather big datasets can be overcome using transfer learning techniques, as it was done in the aforementioned yield prediction model.[107] We expect the rise of the application of such approaches for reaction property prediction in the nearest future.

Modelling. At present, the workflow of QSRR modelling is well elaborated and straightforward. Many models were developed, either for reactions with simple mechanisms, like S_N2 and for complex reactions, like Suzuki–Miyaura. Reaction rates and equilibrium constants for single type reactions can be modelled routinely, but the lack of experimental data limits the development of this topic. QSRR models for predicting the reaction yield for diverse reactions are in high demand. They are needed for reaction conditions prediction, reaction feasibility assessment, and synthetic plan prioritization. Unfortunately, yield is a very noisy characteristic for modelling, and successful examples of modelling are mainly based on in-house or HTE datasets. Predictions of reaction conditions are still *terra incognita* in reaction informatics: although several approaches have been suggested, it is unknown which method is the best one. Besides, prediction of reaction conditions is a quite challenging task because: (1) negative examples of reaction conditions are not recorded in databases, (2) only in a few cases all possible conditions were fully enumerated and tested in experiments, (3) if a model predicts some new conditions that were not tested, this does not mean that such conditions are not suitable for the reaction. Only one universal model for condition prediction has been built to date;[29] however, its tests in a real-world scenario was not published yet.

It is already clear that commonly used techniques for QSRR model validation can fail, and several reasons for such failure can be suggested. Although several 'best practices' for model validation have been recently proposed,[80,112] they still need to be accepted by society and implemented into standard modelling workflows. Moreover, almost all published QSRR models ignore the problem of applicability domains, which are well developed and widely used in standard QSAR modelling.

Recent progress in chemoinformatics, machine learning, and artificial intelligence is transforming the field of organic chemistry. A brave new world of synthesis robots[190] is coming[109,191–195] as more efficient synthesis planning, and reaction procedure prediction approaches[196] are proposed. Several initiatives to create such automatized synthesis robots are opened worldwide.[32,197] Also, we believe that very soon, the accuracy of these approaches will achieve such high quality that synthetic chemists will start to apply computer-aided synthesis design and QSRR models in their everyday practice.

Russian Science Foundation (project no. 19-73-10137) is acknowledged for the support.

Online Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.mencom.2021.11.003.

References

1 E. J. Corey, *Chem. Soc. Rev.*, 1988, **17**, 111.

2 I. I. Baskin, T. I. Madzhidov, I. S. Antipin and A. A. Varnek, *Russ. Chem. Rev.*, 2017, **86**, 1127.

3 P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, *Chem. Sci.*, 2020, **11**, 3316.

4 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604.

5 C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281.

6 O. Engkvist, P.-O. Norrby, N. Selmi, Y.-H. Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, *Drug Discov. Today*, 2018, **23**, 1203.

7 S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2016, **55**, 5904.

8 L. P. Hammett, *J. Am. Chem. Soc.*, 1937, **59**, 96.

9 G. S. Hammond, *Pure Appl. Chem.*, 1997, **69**, 1919.

10 E. V. Anslyn and D. A. Dougherty, *Modern Physical Organic Chemistry*, University Science Books, Sausalito, CA, 2006.

11 R. W. Taft, Jr., *J. Am. Chem. Soc.*, 1952, **74**, 3120.

12 V. A. Palm, *Osnovy kolichestvennoi teorii organicheskikh reaktsii (Fundamentals of the Quantitative Theory of Organic Reactions)*, Khimiya, 1977 (in Russian).

13 P. R. Wells, *Chem. Rev.*, 1963, **63**, 171.

14 C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **86**, 1616.

15 R. F. Rekker, *Quant. Struct.-Act. Relat.*, 1992, **11**, 195.

16 F. Ignatz-Hoover, R. Petrukhin, M. Karelson and A. R. Katritzky, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 295.

17 U. A. Chaudry and P. L. A. Popelier, *J. Phys. Chem. A*, 2003, **107**, 4578.

18 H. Zhang, X. Qu and H. Ando, *J. Mol. Struct.: THEOCHEM*, 2005, **725**, 31.

19 A. R. Katritzky, S. Perumal and R. Petrukhin, *J. Org. Chem.*, 2001, **66**, 4036.

20 G. R. Famini and L. Y. Wilson, in *Reviews in Computational Chemistry*, eds. K. B. Lipkowitz and D. B. Boyd, John Wiley & Sons, 2003, vol. 18, pp. 211–255.

21 C. Hansch, A. Leo and R. W. Taft, *Chem. Rev.*, 1991, **91**, 165.

22 C. D. Selassie, in *Burger's Medicinal Chemistry, Drug Discovery and Development*, eds. D. J. Abraham and D. P. Rotella, Wiley, Hoboken, 2003, vol. 1.

23 N. M. Halberstam, I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *Mendeleev Commun.*, 2002, **12**, 185.

24 D. M. Lowe, *PhD Thesis*, 2012, doi: https://doi.org/10.17863/CAM.16293

25 I. I. Baskin, D. Winkler and I. V. Tetko, *Expert Opin. Drug Discov.*, 2016, **11**, 785.

26 P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091.

27 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572.

28 A. I. Lin, T. I. Madzhidov, O. Klimchuk, R. I. Nugmanov, I. S. Antipin and A. Varnek, *J. Chem. Inf. Model.*, 2016, **56**, 2140.

29 H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2018, **4**, 1465.

30 W. Bort, I. I. Baskin, T. Gimadiev, A. Mukanov, R. Nugmanov, P. Sidorov, G. Marcou, D. Horvath, O. Klimchuk, T. Madzhidov and A. Varnek, *Sci. Rep.*, 2021, **11**, 3178.

31 E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chem. Soc. Rev.*, 2020, **49**, 3525.

32 A. Extance, *Chem. World*, 2020, 4012359.

33 C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**, 1566.

34 *Chemoinformatics*: *A Textbook*, eds. J. Gasteiger and T. Engel, Wiley-VCH, Weinheim, 2003.

35 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31.

36 A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland and J. Laufer, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 244.

37 J. Sadowski, in *Handbook of Chemoinformatics: From Data to Knowledge*, ed. J. Gasteiger, Wiley-VCH, 2003, vol. 1, pp. 231–261.

38 *Daylight Theory Manual*, version 4.1, Daylight Chemical Information Systems, Laguna Niguel, CA, 2011, https://www.daylight.com/dayhtml/doc/theory/index.

39 J. Dugundji and I. Ugi, *Top. Curr. Chem.*, 1973, **39**, 19.

40 J. Gasteiger and W. D. Ihlenfeldt, in *Software Development in Chemstry*, ed. J. Gasteiger, Springer, 1990, pp. 57–65.

41 J. Gasteiger and C. Jochum, *Top. Curr. Chem.*, 1978, **74**, 93.

42 J. Gasteiger, M. G. Hutchings, B. Christoph, L. Gann, C. Hiller, P. Löw, M. Marsili, H. Saller and K. Yuki, *Top. Curr. Chem.*, 1987, **137**, 19.

43 W. L. Chen, D. Z. Chen and K. T. Taylor, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2013, **3**, 560.

44 A. Varnek, D. Fourches, F. Hoonakker and V. P. Solov'ev, *J. Comput.-Aided Mol. Des.*, 2005, **19**, 693.

45 R. I. Nugmanov, R. N. Mukhametgaleev, T. Akhmetshin, T. R. Gimadiev, V. A. Afonina, T. I. Madzhidov and A. Varnek, *J. Chem. Inf. Model.*, 2019, **59**, 2516.

46 A. Wagner, F. Hoonakker and A. Varnek, *US Patent 2009/0024575 A1*, 2009.

47 A. de Luca, D. Horvath, G. Marcou, V. Solov'ev and A. Varnek, *J. Chem. Inf. Model.*, 2012, **52**, 2325.

48 V. Delannée and M. C. Nicklaus, *J. Cheminf.*, 2020, **12**, 72.

49 M. Glavatskikh, T. Madzhidov, I. I. Baskin, D. Horvath, R. Nugmanov, T. Gimadiev, G. Marcou and A. Varnek, *Mol. Inf.*, 2018, **37**, 1800056.

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

51 R. I. Nugmanov, T. I. Madzhidov, G. R. Khaliullina, I. I. Baskin, I. S. Antipin and A. A. Varnek, *J. Struct. Chem.*, 2014, **55**, 1026 (*Zh. Strukt. Khim.*, 2014, **55**, 1080).

52 T. I. Madzhidov, A. V. Bodrov, T. R. Gimadiev, R. I. Nugmanov, I. S. Antipin and A. A. Varnek, *J. Struct. Chem.*, 2015, **56**, 1227 (*Zh. Strukt. Khim.*, 2015, **56**, 1293).

53 A. A. Kravtsov, P. V. Karpov, I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *Dokl. Chem.*, 2011, **440**, 299 (*Dokl. Akad. Nauk*, 2011, **440**, 770).

54 A. A. Kravtsov, P. V. Karpov, I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *Dokl. Chem.*, 2011, **441**, 314 (*Dokl. Akad. Nauk*, 2011, **441**, 57).

55 F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem.*, 2020, **6**, 1379.

56 G. Marcou, J. Aires de Sousa, D. A. R. S. Latino, A. de Luca, D. Horvath, V. Rietsch and A. Varnek, *J. Chem. Inf. Model.*, 2015, **55**, 239.

57 P. Polishchuk, T. Madzhidov, T. Gimadiev, A. Bodrov, R. Nugmanov and A. Varnek, *J. Comput.-Aided Mol. Des.*, 2017, **31**, 829.

58 N. Schneider, D. M. Lowe, R. A. Sayle and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 39.

59 Q.-N. Hu, H. Zhu, X. Li, M. Zhang, Z. Deng, X. Yang and Z. Deng, *PLoS One*, 2012, **7**, e52901.

60 Q.-Y. Zhang and J. Aires-de-Sousa, *J. Chem. Inf. Model.*, 2005, **45**, 1775.

61 D. A. R. S. Latino, Q.-Y. Zhang and J. Aires-de-Sousa, *Bioinformatics*, 2008, **24**, 2236.

62 J.-L. Faulon, M. Misra, S. Martin, K. Sale and R. Sapra, *Bioinformatics*, 2008, **24**, 225.

63 L. Ridder and M. Wagener, *ChemMedChem*, 2008, **3**, 821.

64 I. Oprisiu, E. Varlamova, E. Muratov, A. Artemenko, G. Marcou, P. Polishchuk, V. Kuz'min and A. Varnek, *Mol. Inf.*, 2012, **31**, 491.

65 A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. V. Tetko and G. Marcou, *Curr. Comput.-Aided Drug Des.*, 2008, **4**, 191.

66 I. Baskin and A. Varnek, in *Chemoinformatics Approaches to Virtual Screening*, eds. A. Varnek and A. Tropsha, RSC Publishing, 2008, pp. 1–43.

67 D. Horvath, G. Marcou, A. Varnek, S. Kayastha, A. de la Vega de León and J. Bajorath, *J. Chem. Inf. Model.*, 2016, **56**, 1631.

68 M. Glavatskikh, T. Madzhidov, D. Horvath, R. Nugmanov, T. Gimadiev, D. Malakhova, G. Marcou and A. Varnek, *Mol. Inf.*, 2019, **38**, 1800077.

69 T. I. Madzhidov, T. R. Gimadiev, D. A. Malakhova, R. I. Nugmanov, I. I. Baskin, I. S. Antipin and A. A. Varnek, *J. Struct. Chem.*, 2017, **58**, 650 (*Zh. Strukt. Khim.*, 2017, **58**, 685).

70 T. R. Gimadiev, T. I. Madzhidov, R. I. Nugmanov, I. I. Baskin, I. S. Antipin and A. Varnek, *J. Comput.-Aided Mol. Des.*, 2018, **32**, 401.

71 J. Catalán, V. López, P. Pérez, R. Martin-Villamil and J.-G. Rodríguez, *Liebigs Ann.*, 1995, 241.

72 J. Catalán and C. Díaz, *Liebigs Ann.*, 1997, 1941.

73 J. Catalán and C. Díaz, *Eur. J. Org. Chem.*, 1999, 885.

74 J. Catalán, C. Díaz, V. López, P. Pérez, J.-L. G. De Paz and J. G. Rodríguez, *Liebigs Ann.*, 1996, 1785.

75 M. J. Kamlet and R. W. Taft, *J. Am. Chem. Soc.*, 1976, **98**, 377.

76 R. W. Taft and M. J. Kamlet, *J. Am. Chem. Soc.*, 1976, **98**, 2886.

77 M. J. Kamlet, J. L. Abboud and R. W. Taft, *J. Am. Chem. Soc.*, 1977, **99**, 6027.

78 Y. Marcus, *The Properties of Solvents*, Wiley, 1998.

79 G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Sci. Rep.*, 2017, **7**, 3582.

80 A. Rakhimbekova, T. N. Akhmetshin, G. I. Minibaeva, R. I. Nugmanov, T. R. Gimadiev, T. I. Madzhidov, I. I. Baskin and A. Varnek, *SAR QSAR Environ. Res.*, 2021, **32**, 207.

81 *React. – CASREACT*, 2021, http://www.cas.org/support/documentation/reactions.

82 *Reaxys*, 2021, www. reaxys.com.

83 J. Goodman, *J. Chem. Inf. Model.*, 2009, **49**, 2897.

84 *SPRESI*, 2019, http://www.spresi.com/.

85 *SciVal*, 2021, https://www.scival.com/.

86 T. R. Gimadiev, A. Lin, V. A. Afonina, D. Batyrshin, R. I. Nugmanov, T. Akhmetshin, P. Sidorov, N. Duybankova, J. Verhoeven, J. Wegner, H. Ceulemans, A. Gedich, T. I. Madzhidov and A. Varnek, *Mol. Inf.*, 2021, 2100119.

87 *Pistachio*, 2021, https://www.nextmovesoftware.com/pistachio.html.

88 W. Jin, C.W. Coley, R. Barzilay and T. Jaakkola, *arXiv: 1709.04555*, 2017.

89 N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336.

90 D. Q. Nguyen, Z. Zhai, H. Yoshikawa, B. Fang, C. Druckenbrodt, C. Thorne, R. Hoessel, S. A. Akhondi, T. Cohn, T. Baldwin and K. Verspoor, in *Advances in Information Retrieval*, eds. J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva and F. Martins, Springer, Cham, 2020, pp. 572–579.

91 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186.

92 *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions*, ed. V. I. Palm, VINITI, 1978.

93 *ChemInform Reaction Library*, 2021, http://www.cheminform.com/reaction.

94 W. Jin and C. W. Coley, *Rexgen*, 2021, https://github.com/wengong-jin/nips17-rexgen.

95 L. P. Hammett, *Chem. Rev.*, 1935, **17**, 125.

96 L. P. Hammett, *Trans. Faraday Soc.*, 1938, **34**, 156.

97 H. F. McDuffie and G. Dougherty, *J. Am. Chem. Soc.*, 1942, **64**, 297.

98 Q. Zhang, X. Qu, H. Wang, F. Xu, X. Shi and W. Wang, *Environ. Sci. Technol.*, 2009, **43**, 4105.

99 M. Bräuer, J. L. Pérez-Lustres, J. Weston and E. Anders, *Inorg. Chem.*, 2002, **41**, 1454.

100 I. A. Koppel and V. A. Palm, in *Advances in Linear Free Energy Relationships*, eds. N. B. Chapman and J. Shorter, Plenum Press, London, 1972, ch. 5, pp. 203–280.

101 N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov and N. S. Zefirov, *Dokl. Chem.*, 2007, **417**, 282 (*Dokl. Akad. Nauk*, 2007, **417**, 639).

102 F. Hoonakker, N. Lachiche, A. Varnek and A. Wagner, *Int. J. Artif. Intell. Tools*, 2011, **20**, 253.

103 A. Rakhimbekova, T. I. Madzhidov, R. I. Nugmanov, T. R. Gimadiev, I. I. Baskin and A. Varnek, *Int. J. Mol. Sci.*, 2020, **21**, 5542.

104 I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches and A. Varnek, *J. Chem. Inf. Model.*, 2008, **48**, 1733.

105 R. G. Bergman and R. L. Danheiser, *Angew. Chem., Int. Ed.*, 2016, **55**, 12548.

106 F. Huerta, S. Hallinder and A. Minidis, *ChemRxiv*, 2020, https://dx.doi.org/10.26434/chemrxiv.12613214.

107 P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Mach. Learn. Sci. Technol.*, 2021, **2**, 015016.

108 Z. Fu, X. Li, Z. Wang, Z. Li, X. Liu, X. Wu, J. Zhao, X. Ding, X. Wan, F. Zhong, D. Wang, X. Luo, K. Chen, H. Liu, J. Wang, H. Jiang and M. Zheng, *Org. Chem. Front.*, 2020, **7**, 2269.

109 J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377.

110 M. K. Nielsen, D. T. Ahneman, O. Riera and A. G. Doyle, *J. Am. Chem. Soc.*, 2018, **140**, 5004.

111 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186.

112 K. V. Chuang and M. J. Keiser, *Science*, 2018, **362**, aat8603.

113 D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, *Science*, 2018, **359**, aap9112.

114 B. J. Reizman, Y.-M. Wang, S. L. Buchwald and K. F. Jensen, *React. Chem. Eng.*, 2016, **1**, 658.

115 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, *arXiv: 1706.03762*, 2017.

116 P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino and J.-L. Reymond, *Nat. Mach. Intell.*, 2021, **3**, 144.

117 K. Mansouri, N. F. Cariello, A. Korotcov, V. Tkachenko, C. M. Grulke, C. S. Sprankle, D. Allen, W. M. Casey, N. C. Kleinstreuer and A. J. Williams, *J. Cheminform.*, 2019, **11**, 60.

118 A. C. Lee and G. M. Crippen, *J. Chem. Inf. Model.*, 2009, **49**, 2013.

119 F. Luan, W. Ma, H. Zhang, X. Zhang, M. Liu, Z. Hu and B. Fan, *Pharm. Res.*, 2005, **22**, 1454.

120 J. H. Jensen, C. J. Swain and L. Olsen, *J. Phys. Chem. A*, 2017, **121**, 699.

121 F. Eckert and A. Klamt, *J. Comput. Chem.*, 2006, **27**, 11.

122 C. Liao and M. C. Nicklaus, *J. Chem. Inf. Model.*, 2009, **49**, 2801.

123 J. Elguero, C. Marzin, A. R. Katritzky and P. Linda, *The Tautomerism of Heterocycles*, Academic Press, New York, 1976.

124 A. Klamt and M. Diedenhofen, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 621.

125 I. Soteras, M. Orozco and F. J. Luque, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 281.

126 J. R. Greenwood, D. Calkins, A. P. Sullivan and J. C. Shelley, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 591.

127 I. Alkorta and J. Elguero, *J. Phys. Org. Chem.*, 2005, **18**, 719.

128 J. Szegezdi and F. Csizmadia, in *Fall ACS National Meeting*, Boston, August 19–23, 2007.

129 F. Milletti, L. Storchi, G. Sforna, S. Cross and G. Cruciani, *J. Chem. Inf. Model.*, 2009, **49**, 68.

130 T. R. Gimadiev, T. I. Madzhidov, R. I. Nugmanov, I. I. Baskin, I. S. Antipin and A. Varnek, *J. Comput.-Aided Mol. Des.*, 2018, **32**, 401.

131 D. V. Zankov, T. I. Madzhidov, A. Rakhimbekova, T. R. Gimadiev, R. I. Nugmanov, M. A. Kazymova, I. I. Baskin and A. Varnek, *J. Chem. Inf. Model.*, 2019, **59**, 4569.

132 M. C. Kozlowski, S. L. Dixon, M. Panda and G. Lauri, *J. Am. Chem. Soc.*, 2003, **125**, 6614.

133 A. F. Zahrt, S. V. Athavale and S. E. Denmark, *Chem. Rev.*, 2020, **120**, 1620.

134 J. R. Woolfrey, M. A. Avery and A. M. Doweyko, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 165.

135 *3D QSAR in Drug Design*, ed. H. Kubinyi, Springer, Netherlands, 1994.

136 R. D. Cramer, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959.

137 K. B. Lipkowitz and M. Pradhan, *J. Org. Chem.*, 2003, **68**, 4648.

138 M. Pastor, G. Cruciani, I. McLay, S. Pickett and S. Clementi, *J. Med. Chem.*, 2000, **43**, 3233.

139 S. Sciabola, A. Alex, P. D. Higginson, J. C. Mitchell, M. J. Snowden and I. Morao, *J. Org. Chem.*, 2005, **70**, 9025.

140 P. Braiuca, K. Lorena, V. Ferrario, C. Ebert and L. Gardossi, *Adv. Synth. Catal.*, 2009, **351**, 1293.

141 K. C. Harper, E. N. Bess and M. S. Sigman, *Nat. Chem.*, 2012, **4**, 366.

142 J. J. Miller and M. S. Sigman, *Angew. Chem., Int. Ed.*, 2008, **47**, 771.

143 J. D. Oslob, B. Åkermark, P. Helquist and P.-O. Norrby, *Organometallics*, 1997, **16**, 3015.

144 T. T. Metsänen, K. W. Lexa, C. B. Santiago, C. K. Chung, Y. Xu, Z. Liu, G. R. Humphrey, R. T. Ruck, E. C. Sherer and M. S. Sigman, *Chem. Sci.*, 2018, **9**, 6922.

145 Y. Park, Z. L. Niemeyer, J.-Q. Yu and M. S. Sigman, *Organometallics*, 2018, **37**, 203.

146 J. L. Melville, K. R. J. Lovelock, C. Wilson, B. Allbutt, E. K. Burke, B. Lygo and J. D. Hirst, *J. Chem. Inf. Model.*, 2005, **45**, 971.

147 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.

148 J. J. Henle, A. F. Zahrt, B. T. Rose, W. T. Darrow, Y. Wang and S. E. Denmark, *J. Am. Chem. Soc.*, 2020, **142**, 11578.

149 L. Breiman, *Mach. Learn.*, 2001, **45**, 5.

150 L.-C. Xu, X. Li, M.-J. Tang, L.-T. Yuan, J.-Y. Zheng, S.-Q. Zhang and X. Hong, *Synlett*, 2020, doi: 10.1055/S-0040-1705977.

151 D. Zankov, P. Polishchuk, T. I. Madzhidov and A. Varnek, *Synlett*, 2021, doi: 10.1055/A-1553-0427.

152 A. Kutlushina, A. Khakimova, T. Madzhidov and P. Polishchuk, *Molecules*, 2018, **23**, 3094.

153 D. V. Zankov, M. Matveieva, A. Nikonenko, R. Nugmanov, A. Varnek, P. Polishchuk and T. Madzhidov, *ChemRxiv Prepr. 13456277*, 2020, 1.

154 L. Chen, in *Handbook of Chemoinformatics: From Data to Knowledge*, ed. J. Gasteiger, Wiley-VCH, 2003, vol. 1, pp. 348–388.

155 A. T. Balaban, *Rev. Roum. Chim.*, 1967, **12**, 875.

156 J. B. Hendrickson, *Angew. Chem., Int. Ed. Engl.*, 1974, **13**, 47.

157 J. B. Hendrickson and L. Chen, in *Encyclopedia of Computational Chemistry*, John Wiley & Sons, Ltd., 2002, doi: 10.1002/0470845015.cca022.

158 J. F. Arens, *Recl. Trav. Chim. Pays-Bas*, 1979, **98**, 155.

159 S. S. Tratch and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 349.

160 N. S. Zefirov and S. S. Tratch, *MATCH*, 1977, 263.

161 S. Fujita, *J. Chem. Inf. Comput. Sci.*, 1986, **26**, 238.

162 S. Fujita, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 104.

163 H. Kraut, J. Eiblmaier, G. Grethe, P. Löw, H. Matuszczyk and H. Saller, *J. Chem. Inf. Model.*, 2013, **53**, 2884.

164 S. D. Roughley and A. M. Jordan, *J. Med. Chem.*, 2011, **54**, 3451.

165 J. S. Carey, D. Laffan, C. Thomson and M. T. Williams, *Org. Biomol. Chem.*, 2006, **4**, 2337.

166 *NextMove Software*, 2020, https://www.nextmovesoftware.com/namerxn.html.

167 N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli and G. A. Landrum, *J. Med. Chem.*, 2016, **59**, 4385.

168 C. D. Christ, M. Zentgraf and J. M. Kriegl, *J. Chem. Inf. Model.*, 2012, **52**, 1745.

169 L. Chen and J. Gasteiger, *J. Am. Chem. Soc.*, 1997, **119**, 4033.

170 G. Sello, *Tetrahedron*, 1998, **54**, 5731.

171 G. Sello and M. Termini, *Tetrahedron*, 1997, **53**, 14085.

172 G. M. Ghiandoni, M. J. Bodkin, B. Chen, D. Hristozov, J. E. A. Wallace, J. Webster and V. J. Gillet, *J. Chem. Inf. Model.*, 2019, **59**, 4167.

173 V. Vovk, A. Gammerman and G. Shafer, *Algorithmic Learning in a Random World*, Springer, 2005.

174 J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725.

175 H. Struebing, Z. Ganase, P. G. Karamertzanis, E. Siougkrou, P. Haycock, P. M. Piccione, A. Armstrong, A. Galindo and C. S. Adjiman, *Nat. Chem.*, 2013, **5**, 952.

176 E. Walker, J. Kammeraad, J. Goetz, M. T. Robo, A. Tewari and P. M. Zimmerman, *J. Chem. Inf. Model.*, 2019, **59**, 3645.

177 C. Coley, M. Fortunato, H. Gao, P. Plehiers, M. Cameron, M. Liu, Y. Wang, T. Struble, J. Liu and Y. Mo, *GitHub*, 2021, https://github.com/ASKCOS.

178 T. N. Kipf and M. Welling, *arXiv: 1609.02907*, 2016.

179 V. Korolev, A. Mitrofanov, A. Korotcov and V. Tkachenko, *J. Chem. Inf. Model.*, 2020, **60**, 22.

180 S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model.*, 2018, **58**, 27.

181 A. Zhavoronkov, Ya. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, Yu. Volkov, A. Zholus, R. R. Shayakhmetov, A. Zhebrak, L. I. Minaeva, B. A. Zagribelnyy, L. H. Lee, R. Soll, D. Madge, L. Xing, T. Guo and A. Aspuru-Guziket, *Nat. Biotechnol.*, 2019, **37**, 1038.

182 B. Sattarov, I. I. Baskin, D. Horvath, G. Marcou, E. J. Bjerrum and A. Varnek, *J. Chem. Inf. Model.*, 2019, **59**, 1182.

183 D. Merk, L. Friedrich, F. Grisoni and G. Schneider, *Mol. Inf.*, 2018, **37**, 1700153.

184 M. Popova, O. Isayev and A. Tropsha, *Sci. Adv.*, 2018, **4**, doi: 10.1126/sciadv.aap7885.

185 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360.

186 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572.

187 C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370.

188 P. Karpov, G. Godin and I. V. Tetko, in *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, eds. I. V. Tetko, V. Kůrková, P. Karpov and F. Theis, 2019, pp. 817–830.

189 P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, *Chem. Sci.*, 2020, **11**, 3316.

190 M. Peplow, *Nature*, 2014, **512**, 20.

191 C. W. Coley, D. A. Thomas, 3rd, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**, eaax1566.

192 S. Asche, G. J. T. Cooper, G. Keenan, C. Mathis and L. Cronin, *Nat. Commun.*, 2021, **12**, 3547.

193 S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone and L. Cronin, *Science*, 2019, **363**, eaav2211.

194 A. B. Henson, P. S. Gromski and L. Cronin, *ACS Cent. Sci.*, 2018, **4**, 793.

195 B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237.

196 A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller and T. Laino, *Nat. Commun.*, 2020, **11**, 3601.

197 *Dial-a-Molecule EPSRC Grand Challenge Network Website*, 2021, http//generic.wordpress.soton.ac.uk/dial-a-molecule/.

Received: 18th May 2021; Com. 21/6561