DOI: 10.1070/RCR4746

Artificial intelligence in synthetic chemistry: achievements and prospects

Igor I. Baskin,[a, b] Timur I. Madzhidov,[b] Igor S. Antipin,[b] Alexandre A. Varnek [b, c]

[a] Faculty of Physics, Lomonosov Moscow State University
Leninskie gory 1, stroenie 2, 119991 Moscow, Russian Federation
[b] A.M.Butlerov Institute of Chemistry, Kazan (Volga Redgion) Federal University
ul. Kremlevskaya 18, 420008 Kazan, Russian Federation
[c] Laboratory of Chemoinformatics (UMR 7140 CNRS/UniStra), University of Strasbourg
4 rue Blaise Pascal, 67000 Strasbourg, France

The review is devoted to the achievements in analysis of information on chemical reactions using machine learning methods. Four large areas that actively use these methods are outlined: computer-assisted planning of synthesis, analysis and visualization of chemical reaction data, prediction of the quantitative characteristics of reactions and computer-aided design of catalysts.
The bibliography includes 346 references.

Contents

I. Introduction 1127
II. Computer coding of chemical reactions 1129
III. Preliminary processing of reaction data 1131
IV. Computer-assisted methods for planning organic synthesis 1132
V. Analysis and visualization of reaction data 1137
VI. Prediction of quantitative characteristics of the reactions 1142
VII. Modelling of catalytic activity and computer-aided design of new catalysts 1147
VIII. Conclusion 1149

I. Introduction

A huge body of information on chemical reactions has been accumulated to date: the data on more than 80×10^6 chemical reactions are stored in the largest CASReact database (http://www.cas.org/content/reactions). Large arrays of data can also be found in original journal papers, reviews, reference books, theses and patents.[1] Analysis of these data is needed to develop tools to be used by chemists for planning a synthesis. Efficient handling of such huge and intricately arranged amounts of data requires the use of high-performance computers and software that would be based on artificial intelligence techniques.[2]

The term 'artificial intelligence' usually refers to the property of computing systems to solve problems that are traditionally meant for humans.[3] The notion of artificial intelligence appeared rather long ago and was continuously changed with varying ideas of what problems can be solved

I.I.Baskin. Doctor of Physical and Mathematical Sciences, Leading Researcher of the Faculty of Physics at the MSU, Senior Researcher at the KFU.
Telephone: +7(495)939 – 2677, e-mail: igbaskin@gmail.com
Current research interests: chemoinformatics, computational chemistry, machine learning.
T.I.Madzhidov. Candidate of Chemical Sciences, Senior Researcher at the KFU.
Telephone: +7(843)233 – 7371, e-mail: timur.madzhidov@kpfu.ru
Current research interests: chemoinformatics, quantum chemistry, chemical reactions.
I.S.Antipin. Corresponding Member of the RAS, Doctor of Chemical Sciences, Professor, Head of the Department of Organic Chemistry at the KFU.
Telephone: +7(843)233 – 7463, e-mail: iantipin54@yandex.ru
Current research interests: organic chemistry, supramolecular chemistry, host – guest complexes.
A.A.Varnek. Doctor of Science, Professor at the University of Strasbourg, Chief Researcher at the KFU.
Telephone: +33(0)3 – 6885 – 1560, e-mail: varnek@unistra.fr
Current research interests: chemoinformatics, computational chemistry, molecular modelling.

Received 14 April 2017
Translation: Z.P.Svitanko

© 2017 Russian Academy of Sciences and Turpion Ltd.

by computers and what problems necessarily require a human brain. Whereas formerly the artificial intelligence was mainly addressed by philosophers and science-fiction or script writers, today it is the object of vigorous scientific research, rapidly entering everyday life and becoming a major factor in the science-and-technology, economic and even cultural development of the humankind.

This review is devoted to the use of artificial intelligence tools for analyzing information on chemical reactions and development of predictive tools able to answer the following questions, which are important for a synthetic chemist:

— how to synthesize a specified compound;

— what conditions (catalyst, solvent, reactants, temperature, *etc.*) are required for the reaction to be most regio(chemo, stereo)selective and to give the highest yields of products;

— what the rate, the yield and the time for the given reaction will be;

— what major (minor) product is to be expected for this reaction conducted under specified conditions;

— how to find a reaction similar to this one in either a commercial or open-access database in order to get an idea of the procedure for the synthesis of a specified compound;

— what types of reactions are stored in the database and whether there are enough examples of the reaction type of interest.

The development of computing technologies that are suitable for discovering answers to these questions started in the 1960s with studies of Corey *et al.*[4–6] The OCSS (organic chemical simulation of synthesis) program, developed in 1969, was the first computer software meant for synthetic chemists.[4] This program was based on the concept of retrosynthetic analysis, which represents a multistage chart (tree) of the synthesis of an organic compound from precursors.[5] This software was further developed as LHASA (logic and heuristics applied to synthetic analysis) under supervision of Corey[6] and as SECS (simulation and evaluation of chemical synthesis) under supervision of Wipke.[7,8] The cited studies[4–8] had a determining influence on the formation of a whole research area, which was coined computer synthesis.[9,10] The research and development of this area resulted in numerous computer programs for retrosynthetic analysis (SYNCHEM,[11] SYNSUP,[12,13] WODCA,[14] TOSCA,[15] SYNGEN,[16] FORWARD,[17] LILITH,[18,19] FLAMINGO,[20] COMPASS,[21,22] SYNLMA,[23] CHIRON,[24–26] SST,[27] SOS,[28,29] SESAM,[30] AIPHOS,[31] KOSP,[32] RASA,[33] *etc.*), for predicting reaction products (CAMEO,[34,35] EROS,[36,37] PSYCHO,[38] SOPHIA,[39,40] ROBIA,[41,42] *etc.*), for the search for new types of chemical transformations (IGOR,[43,44] SYMBEQ,[45] ARGENT,[46–48] RENE,[49] *etc.*) and for analysis of reaction mechanisms (RAIN,[50] ICAR,[51,52] *etc.*). The key characteristics of these programs, the potential of underlying approaches and particular application examples are comprehensively considered in review publications.[9,10,53–57]

Despite the hopes put on the programs listed above, they still have not turned into widely demanded tools for synthetic chemists. The almost 20-year stagnation in the computer-assisted synthesis is usually attributed to the following factors:

— small amount of encoded knowledge in comparison with the immense body of information pertaining to organic chemistry;

— low computer performance at that time and small amount of data computers could process;

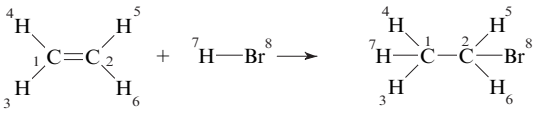
— limited range of task setting and solution using computers in comparison with the everyday demands of synthetic chemists.[57]

Nevertheless, the approaches to handling the information about organic reactions elaborated during the creation of numerous software products for computer-assisted synthesis and the accumulated experience were not lost in vain — they provided the basis for new-generation software systems actively developed in recent years, which are capable of treating big data covering the information on an enormous number (up to hundreds of millions) of organic reactions. Software products designed for retrosynthetic analysis are, in particular, ARChem (route designer)[58,59] (developed by SimBioSys, Canada), ICSYNTH[60] (developed by InfoChem GmbH, Germany) and CHEMATICA.[57] Mention should also be made of synthesis planners arranged as interfaces to large chemical reaction databases. Using the planners such as Reaxys AutoPlan (https://vimeo.com/63343963) and SciFinder SciPlanner (http://www.cas.org/etrain/scifinder/sciplanner.html), one can interactively construct the trees of synthesis. Most systems of this type are commercial products.

The methods for computer analysis of chemical reaction data, especially computer-assisted synthesis, have always been closely connected with contemporary views on artificial intelligence. Indeed, almost all of the mentioned software products implement the concept of expert systems popular in the 1970–1980s, which included the knowledge bases formed by experienced experts and the inference engine mechanism.[61] By knowledge base is meant the set of principles (rules) describing chemical reactions and inference engine stands for the synthesis trees. However, the 1990s brought disappointment in expert systems of this type, because the manually formed sets of rules are insufficient for solving practically important problems. The modern stage of development of artificial intelligence relies on an alternative paradigm of the formation of knowledge — automatic knowledge extraction from the data by means of machine learning methods. Currently, machine learning methods are actively used in chemoinformatics — a scientific discipline that develops approaches to both the analysis of known chemical data and creation of mathematical models for predicting the properties and methods of synthesis of new compounds.[62–69]

This review is devoted to achievements in analysis of chemical reaction data using machine learning methods. The following large areas that actively use these methods are outlined: computer planning of synthesis, chemical reaction data analysis and visualization, prediction of quantitative characteristics of reactions and finally computer design of catalysts. Prior to analysis of relevant publications, we consider versions of computer coding of chemical reactions that are used in chemoinformatics algorithms.

The limited size of this review does not allow giving detailed information about the popular machine learning techniques such as artificial neural networks, multiple linear regression, k nearest neighbours, partial least squares, support vector machine, decision trees and other mehods that are mentioned in the review. Interested readers can find this information in the literature (*e.g.*, Ref. 69).



B

	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

R

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

E

	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

Figure 1. Dugunji – Ugi representation of hydrobromination of ethylene.[66]
The off-diagonal elements of **B** and **E** matrices correspond to the bond orders, while the diagonal elements correspond to the numbers of lone pair electrons in the atom. For convenience, the zeros in the matrices are omitted; the grey colour marks the ordinal numbers of atoms.

II. Computer coding of chemical reactions

A chemical reaction is a more complex object for computer representation than a chemical compound. It is described by a set of compounds (no fewer than two) corresponding to two types — reactants and products. Chemical transformations are often complex and cover numerous interrelated parallel and successive processes, which depend on the conditions. One and the same reaction equation may refer to different reaction conditions.

Currently, three key approaches to computer representation of chemical reactions are used. The first approach is based on the Dugunji – Ugi matrix formalism,[70] according to which the structures of reactants and products are described by special BE (bond – electron) matrices, while the reaction is described as the difference between these matrices. This approach makes it possible to code chemical reactions to be used in the computer planning of a synthesis; however, it does not provide sufficient opportunities for arranging efficient storage, search and analysis of chemical reactions in databases or for modelling the reactivity.

According to the second approach, a chemical reaction is described as a combination of descriptors † of reaction products and reactants. This representation is the simplest to implement; it is suitable for storing the reactions in databases, for performing some types of search and for calculating descriptors for reactions. The major drawback of this representation is the lack of information about the reaction centre, which precludes the use of reactions in synthesis planning software.

Finally, the third approach is based on the use of a single graph (condensed graph of reaction; CGR), which combines information about the reactant and product structures for a reaction. This approach is universal, as it is applicable to almost any problem concerning the storage, search, analysis and modelling of chemical reactions. In particular, it is possible to calculate and apply descriptors for reactions in the same way as it is done for molecules.

† The numerical values independent of atom numbering characterizing the structural changes during the reaction.

II.1. Dugunji – Ugi matrix formalism

In the mid-1970s, Dugunji and Ugi [70] proposed to describe chemical reactions by so-called **R** matrices. According to this method, all molecules participating in a reaction are represented as BE matrices of reactants (**B**) and products (**E**). This requires that the numbers of atoms incorporated in the reactant and product molecules be equal and each atom in the reactants be matched by an atom in the reaction products, *i.e.*, atom-to-atom mapping should be performed. In this case, the reaction can be described by the **R** matrix defined as the difference between the **E** and **B** matrices (Fig. 1).

Note that the rows or columns of the **R** matrix that correspond to those atoms whose surrounding remains unchanged are filled by zeros. Deleting these rows and columns gives a description of changes taking place in the reaction centre. The Dugunji – Ugi formalism was used to plan organic synthesis and predict reaction pathways,[37, 71] classify reactions according to types,[72] analyze the contents of reaction databases,[73] design new reactions,[74] plan the synthesis [75] and search for the shortest distance between the reactant and product.[76] A drawback of this approach is the absence of a direct method for calculation of reaction descriptors.

II.2. Descriptors of chemical reactions as combinations of descriptors of reactants and products

The goal of modelling of structure – reactivity relationships is to derive an equation relating the property being modelled (rate constant, reaction heat and so on) to the structure of reactants and/or products, which is encoded by a set of parameters — molecular descriptors. It is evident that reaction descriptors should reflect the occurring chemical transformations. There are two options for the calculation of reaction descriptors — either with or without inclusion of information about the reaction centre. The first option is technically more complicated, because identification of the reaction centre requires the atom-to-atom mapping procedure to match the reactant atoms to the product atoms. The set of descriptors describing a molecule or reaction forms a vector of real integers. When binary

numbers (0 or 1) are used as descriptors, they form a bit string called a fingerprint.

A simple method for encoding the structural changes involved in chemical processes is to combine information about the structures of reactants and products, without considering the atom-to-atom mapping or reaction centre. This approach hampers the construction of synthetic and retrosynthetic trees in the computer planning of synthesis; however, it is suitable for computing reaction descriptors by combining the standard descriptors calculated for reactant and product molecules. There are several procedures for transition from reactant and product descriptors to reaction descriptors:

— mere combination of the reactant and product descriptors;

— calculation of the difference between the product and reactant descriptors;

— considering reactants and/or products as mixtures and calculation of descriptors for them.

One of the simplest ways of representing a reaction as the difference between the product and reactant descriptors is the use of reaction difference fingerprints. For example, Hu *et al.*[77] calculated them as the difference between the product and reactant bit strings. In this case, the molecular fingerprints of reactants (products) were derived by means of OR logic operation for combining the bit strings of their molecules. The resulting descriptor strings were used to classify enzymes according to types.

Schneider *et al.*[78] also implemented the idea of reaction difference fingerprints, but used a different procedure: first, they calculated the bit strings for each molecule involved in the reaction, while the descriptor representation of the reaction was found as the difference between the sum of product bit strings and the sum of reactant bit strings.[79]

For taking into account the additional components (low-molecular-mass reagents, additives and catalysts), Schneider *et al.*[78] proposed two procedures. According to the first one, the weigthed sum of the mentioned reaction difference fingerprint and the molecular fingerprints of additives was found (the length of the vector thus formed coincided with the length of each of the initial vectors)

$$\mathrm{RDF} = \omega_{\mathrm{reacton}}(\Sigma_{\mathrm{products}}\mathrm{FP}_i - \Sigma_{\mathrm{reactants}}\mathrm{FP}_i) + \\ + \omega_{\mathrm{additives}}\Sigma_{\mathrm{additives}}\mathrm{FP}_i$$

where FP_i are the bit strings of molecules, $\omega_{\mathrm{reaction}}$ is the reaction weight and $\omega_{\mathrm{additives}}$ is the weight of low-molecular-mass additives.

According to the second procedure, the reaction difference fingerprint was combined with the molecular fingerprints of additives (the vector thus formed contained two times more elements than the initial vectors).

The difference fingerprints obtained by the exclusive OR (XOR) operation were proposed by the Daylight company (http:///www.daylight.com/dayhtml/doc/theory/theory.smirks.htm). The required fingerprints of reactants and products can be derived by combining molecular fingerprints of compounds using the OR logic operation.

One more type of difference descriptors has been proposed by Ridder and Wagener,[80] 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 special type of fragment descriptors, which were called atomic signatures, was applied by Faulon *et al.*[81] The MOLMAP reaction descriptors calculated as the difference between the product and reactant descriptors (see below) have been used for classification of reactions.[82–84]

The difference fingerprints can serve for evaluation of similarity and clusterization of reactions, but they are inapplicable for the similarity search, as they do not reflect the presence of structural fragments that are preserved throughout the reaction. A benefit of this representation is that no atom-to-atom mapping for reactants and products is required, while the reaction centre information is encoded in an implicit form. The principal requirement for the calculation of difference descriptors is the stoichiometric character of the reaction, that is, all of the reactants and products must be indicated.

The authors of some papers approximated the reaction descriptors by reactant descriptors. However, this method is admissible only for a certain type of reactions in which the product structures are unambiguously determined by the reactant structures. An example is mapping of the chemical reaction space, with the reactions being described by descriptor vectors composed of local physicochemical characteristics of particular atoms in reactant molecules (see below).

There are relatively few descriptors that are calculated while taking account of the information on the reaction centre. As an example, consider the fragment descriptors calculated for a special reaction-coding entity called condensed graph of reaction.

II.3. Condensed graph of reaction

The idea of describing a chemical reaction by a single graph that bears information about the structures of all reactants and products appeared back in the 1970s and was initially meant for a formal logic consideration and classification of the structural changes occurring during chemical reactions rather than for computer storage and processing of the reaction data. Indeed, Hendrickson[85] considered electron transfers in the reaction centre by analyzing a family of pericyclic reactions with a cyclic reaction centre with even or odd numbers of atoms. He distinguished the bonds in the reaction centre that formally did not change the order. Arens[86–88] described the changes in the formal bond order in the ring and in the chain of atoms; for this purpose, he developed a line notation for the bonds. The overall type of transformation of the reaction centre was termed as operator. The operator applied to atoms of the reaction centre of a particular molecule made it possible to reveal the structure of products.

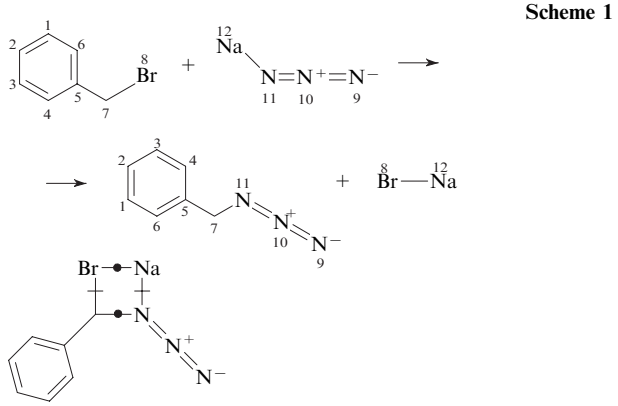
While considering the changes in the types of bonds in the reaction centre in terms of so-called formal logic approach to chemical reactions, Zefirov and Tratch[89–91] proposed in the 1970s–1980s their own classification system for reactions. Within this approach, they developed a procedure based on the strict graph theory for enumerating the types of changes of bond orders in a reaction centre of a particular topology that are allowed by classical organic structure theory.[92,93] This approach was implemented in the SYMBEQ (Ref. 45) and ARGENT (Refs 46–48) programs. It can serve for finding new types of reactions; a new type of transformation has been predicted by SYMBEQ and then confirmed experimentally.[45]

Vladutz and Geivandov [94, 95] defined, for the first time, the reaction centre (called key atoms of the reaction) as a group of atoms in which changes of bonds take place during the reaction and proposed a method for representing the reaction centre in terms of bond changes. According to this approach, the reaction centre atoms of the reactant were to be superimposed onto atoms of the product and changes in the bonds between them during the reaction were to be marked. The graph thus obtained was coined the skeletal scheme of reaction.[94] It is common to reactions of the same type and encodes the changes taking place in the reaction centre. Kiho [96] proposed to construct these compact notations not only for the reaction centre, but also for the whole reaction, including atoms at which no changes of bonds occur. However, this paper passed unnoticed. The same idea was independently stated by Fujita,[97] who called the graph describing the reaction the imaginary transition structure and proposed the terms in-bonds, out-bonds and par-bonds for the bonds that are formed, cleaved and unaffected in the reaction, respectively. Fujita defined the reaction centre as a set of atoms that form the in- and out-bonds. This approach for treating reactions was somewhat upgraded by Jauffret *et al.*[98, 99] and called the condensed graph of reaction. The condensed graph of reaction was used to retrieve information about generalized types of reactions from databases. Scheme 1 ‡ presents an example of the condensed graph of reaction (below) corresponding to nucleophilic substitution (above). The authors proposed to introduce, into the CGR, pseudo-atoms corresponding to the formal charge, stereochemistry and other characteristics of atoms. This enabled explicit encoding of the change in atom characteristics (stereochemistry change and so on) using the concept of dynamic bonds.

Varnek *et al.*[100] paid attention to the fact that CGR can be regarded as a pseudo-molecular graph, for which fragment descriptors can be calculated. This, in turn, opens up the possibility of using many algorithms developed for single molecules for the description of chemical reactions.

The condensed graph of reaction is a versatile tool for representing chemical reactions for computer processing. It can be used to store reactions in databases and to perform various types of search (including the similarity search), automatic classification of reactions, formation of knowledge bases on the reaction pathways, automatic retrieval of rules from reaction databases, calculation of substructure descriptors for reactions, quantitative modelling of reactivity, prediction of the feasibility and kinetic parameters of a reaction, mapping and visualization of reactions and so on.

Scheme 1



‡ In this and other Schemes, the in-bonds are marked by a black dot and the out-bonds are designated by a dash.

III. Preliminary processing of reaction data

A key problem arising in dealing with databases containing information on chemical reactions is the presence of errors [101] caused by the following circumstances:

— problems related to the input data preparation;

— erroneous data annotation;

— incorrect automatic processing of information about reactions.

The first-mentioned issue concerns incorrect identification of the reaction products and the lack of knowledge about some possible products. In some cases, the yield is low and it is unknown what the major product is. In kinetic studies, the type of reaction may be incorrectly determined or ambiguous (*e.g.*, S_N2 or S_N1 mechanism). These errors are exceptionally difficult to trace.

Errors of the second type arise because the data input into the major sources of chemical reaction information, such as Reaxys and CASREACT databases, is manual. This may bring about a variety of problems associated with input errors, misunderstanding or just misprints. Furthermore, errors may be caused by incorrect information presented by the authors of original publications, which is correspondingly reproduced in the database. The situation is aggravated by the problems of structure standardization resulting from the presence of numerous tautomers,[102] resonance structures, methods for depicting the aromaticity and so on.[103] These problems could be solved by inspection of data in order to eliminate faulty, unreliable, doubtful and ambiguous information.[104 – 106] For the processing of large amounts of information about molecules, special automatic tools have been designed in databases, for example, the chemical validation and standardization platform (CVSP) for the ChemSpider database.[107]

In recent years, the manual data annotation for databases has started to be replaced by automatic text mining. Lowe [101] described a protocol for automatic retrieval of information on chemical compounds and reactions from patents. However, this analysis can give rise to errors caused by incorrect determination of the role of some compounds in the reaction. For example, solvents or low-molecular-mass reagents can be erroneously assigned to the left-hand part of the reaction equation, whereas in databases such components should be entered into special designated fields.[108] Even manually inspected databases are not immune from such errors. Schneider *et al.*[109] proposed an algorithm for unambiguous determination of the roles of reaction components based on the difference between reaction fingerprints.

One more source of errors is related to the automatic analysis and annotation of reactions. The key stage of such treatment is identification of the reaction centre, which is accomplished by atom-to-atom mapping (AAM). The essence of this procedure is to establish a one-to-one correspondence between the reactant and product atoms. Methods for the search for reactions by substructure,[110] similarity search for reactions using the hash code for the reaction centre (a multi-digit number unambiguously coding the reaction centre) [111] and application of protocols based on representation of reactions with explicit coding

of the reaction centre (*e.g.*, CGR)[112] require the preliminary AAM. A detailed review of AAM techniques has been published;[113] however, it did not cover several recent publications on this subject.[111, 114–116] The major challenge in the AAM is that this procedure is computationally sophisticated. Currently, there is no algorithm for an unmistakable search for AAM. Muller *et al.*[117] proposed a method for identification of erroneous AAM based on machine learning. The search for erroneous AAM was addressed as a classification problem; for this purpose, a training set containing both correct and wrong mappings was formed. The training set was used for constructing a model in terms of the support vector machine.[118] Using the JRip method[117] trained with this set, simple rules for identification of wrong mappings were created.

It is noteworthy that most of the above problems can be avoided if experimental results are entered into a database directly from electronic laboratory notebooks, which are preferable over the paper versions for the following reasons.

Convenient data access. Electronic laboratory notebooks are suitable for easy search through databases. An opinion survey conducted in the UK has shown that 75% of the respondents encounter difficulties in looking for earlier data obtained in their own laboratory (htpp://www.rsc.org/Membership/Networking/InterestGroups/FaradayDivision/ELNs.asp).

Data structuring. The use of electronic laboratory notebooks requires a certain standardization in presenting the results. Indeed, 90% of the respondents noted that the lack of a clear structure or appropriate quality of data in the paper laboratory notebooks brought about problems hampering the work.

Completeness of data records. The description of an experiment can be accompanied by any digital information such as files containing the spectra, plots or video records.

Easy data exchange. Electronic laboratory notebooks provide easy access for the supervisor to the data of the staff. It is possible to envisage transfer of data to a third party and data import into a database.

Data safety. It is easy to ensure data safety, for example, a copy can be stored in the cloud server. Even a failure or loss of the computer where the record was made would not lead to the loss of data. Meanwhile, destruction of a paper laboratory notebook as a result of fire, loss or just ageing of paper would lead to the permanent loss of data. A study reported in 2013[119] has shown that ~17% of results of research are lost every year because of the use of paper records. Only with timely digitizing, it is possible to avoid huge information and financial loss.

Presence of negative results. The databases containing data from scientific publications store only positive results, that is, how to conduct the reaction, but no negative examples of reactions that the authors failed to conduct. Without this information, chemists may repeat the same unsuccessful experiment over and over again. Furthermore, the presence of negative examples facilitates the use of artificial intelligence techniques with data processing and model design.

Electronic laboratory notebooks have long been used in industrial companies, especially in science intensive fields such as drug development, and are used more and more widely in fundamental research laboratories,[120] where users have noted considerable benefits over paper data storage.[121] From the standpoint of adding data into databases, electronic laboratory notebooks represent a reliable method for data collection, which markedly improves its quality and integrity.

IV. Computer-assisted methods for planning organic synthesis

IV.1. Approaches to computer synthesis

The modern approaches to computer-assisted planning of organic synthesis can be classified into six types (Table 1). The first type comprises methods of retrosynthetic analysis based on the use of rules describing the possible transformations of organic structures. These methods implement the concept of expert systems, which make decisions by emulating the reasoning of a human experts in the subject of interest.[61] As a rule, an expert system includes a knowledge base and an inference engine. The application of expert systems based on the knowledge bases related to various fields of chemistry has been described in detail in a monograph.[122]

In the case of retrosynthetic analysis, knowledge bases are represented by the rules describing chemical transformations, that is, transformations inverse to chemical reactions, which convert products to reactants, and the inference engine is represented by retrosynthetic analysis trees (Fig. 2). The input value is the structure of a chemical compound, while the output is the retrosynthetic analysis tree. The compound to be synthesized is the root of this tree, the possible starting compounds are the leaves, while alternative synthetic pathways form the branches. The number of steps can be either specified manually, or (if a database of available reactants is present) determined automatically. As a rule, the programs include some method for prioritization of particular synthetic pathways. The mechanisms of chemical reactions are usually not considered in the retrosynthetic analysis.

The central position in the computer-assisted retrosynthetic analysis (in the computer synthesis) is occupied by the chemical transformation rules. It is customary to distinguish the empirical approach, in which each rule describes a narrow group of transformations, and the non-empirical approach, in which the rules describe the general reaction principles applicable to a large number of chemical reactions.[9, 10] The empirical approach is used more frequently; the specification of the transformation rules for this approach includes the following:

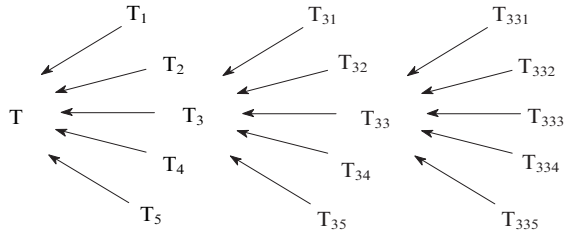


Figure 2. Retrosynthetic analysis tree for a three-step synthesis. Compound T is the final goal of the synthesis; compounds T_i, T_{ij} and T_{ijk} are the possible precursors of compounds T, T_i and T_{ij}, respectively.

Table 1. Various scenarios of computer synthesis

Type	Task	Reactant	Product	Prioritization	Mechanism	Synthesis rules	Reaction plan	Base	Example of program	Ref.
1	Retrosynthesis	out	in	+/–	–	+	out	–	LHASA EROS	6
2	Synthesis planners	out	in	–	–	–	out	–	CHEMATICA	36, 37
3	Generation of mechanisms	in	in	+	out	+	–	–	RAIN	50
4	Forward synthesis	in	out	+	+	+	–	–	CAMEO	34
									EROS	36, 37
5	Forward synthesis for a specified property	out	out	+	–	+	out	–	DOGS	123
6	Reaction prediction with a knowledge graph	in	out	+	–	–	+	+	Knowledge graph	124

Note. The following designations are used: in and out are input and output data, respectively; the plus and minus signs indicate whether or not this parameter has been used in calculations; the plus/minus sign means that prioritization is used only in some retrosynthesis approaches.

— description of one or several substructures that are to be present in the specified system to enable the given transformation;

— description of the conditions usually concerning the local physicochemical characteristics of atoms located in or near the reaction centre that determine the feasibility or priority of the transformation;

— description of structural changes involved in the transformation.

All three components of specification of a transformation are often written as a generalized reaction equation. This can be treated as an analogue of Markush structures containing additional labels that describe the physicochemical properties of atoms and atoms with generalized labels. A classic example of an expert system that implements the empirical approach is the LHASA program.[6] The non-empirical approach is realized in the EROS program,[36, 37] in which all conceivable transformations are described by only five general bond redistribution patterns (so-called reaction generators). The application of programs based on the non-empirical approach is often combined with simple physicochemical calculations for determining the feasibility or priority for a chemical reaction to proceed in the specified structural context. One more specific feature of the systems based on the non-empirical approach is that they can be used to construct both retrosynthetic trees and the trees for the forward synthesis (from reactants to products).

The traditional approach to the formation of rules includes manual coding of the conceivable types of chemical transformations by skilled chemists. Arguments in favour of manual coding are still put forward,[57] despite the fact that the statement about the necessity of automatic retrieval of the rules from chemical reaction databases has become predominant since the 1990s. In this case, each rule is considered as a model that predicts the feasibility of a particular chemical transformation for a compound, while the automatic construction of such models on a computer by generalization of the information contained in databases is a sort of machine learning. The existing approach to automatic generation of transformation rules by analysis of chemical reaction databases is considered below.

The development of databases of chemical reactions has stimulated the appearance of synthesis planners (type 2 in Table 1). Like retrosynthetic analysis programs, synthesis planners construct the retrosynthetic tree for a specified chemical compound — from the compound to be synthesized (root) to the possible starting reactants (leaves). The fundamental difference between retrosynthetic analysis and synthesis planners is that the latter use the reaction equations from the database rather than the coded rules. The application of synthesis planners does not involve the construction of models using machine learning techniques. Examples of synthesis planners are provided by Reaxys AutoPlan (https://vimeo.com/63343963) and SciFinder SciPlanner (http://www.cas.org/etrain/scifinder/sciplanner.html).

Some developed approaches are aimed at elucidation of the mechanisms of complex processes (type 3 in Table 1), *e.g.*, multistep rearrangements. The input data are specifications of the reactants and products and the output is the reaction mechanism as a path in the graph describing transitions between the intermediate compounds. Approaches of this type make use of a set of rules that describe the possible elementary steps of chemical processes and, most often, simple physicochemical calculations for prioritization. The programs for this type of approaches are few in number (RAIN,[50] ICAR [51, 52]) and currently they are not further developed. Meanwhile, active development is observed for the systems for forward synthesis, which often also operate within the mechanistic approach and are quite suitable for generation of various reaction mechanisms.

Forward synthesis programs (type 4 in Table 1) predict the reaction products for specified reactants and conditions. The input data include the structures of the reactants and specified conditions, while the output includes the possible reaction products. A classic example of an expert system operating in this way is CAMEO,[34] which is based on a large number of rules formulated by human chemistry experts and describing the reactivity of organic compounds. The set of rules can specify the presence of both particular substructures and particular calculated physicochemical parameters of the molecules. The successful use of these approaches is largely determined by the quality of the

prioritization method used, which finds the most probable reaction pathway among several alternative pathways. Starting with the 1990s, prioritization has been performed using models obtained by machine learning (see a review [63]) that are based on artificial neural networks.[125–129] It is believed that they implement the so-called connectionist approach to artificial intelligence, by which is usually meant simulation of the cognitive process by modelling functions of brain neurons. The architectures of artificial neural networks and the modes of their use for prioritization of various reaction pathways are considered below.

In recent years, programs for the forward synthesis have been used for compilation of virtual libraries of synthetically readily accessible compounds. These libraries can be used for the search for new biologically active compounds and development of new pharmaceutical agents by means of a virtual screening procedure. An example is provided by the Niklaus' group initiative on the development of 'synthetically accessible virtual inventory' — a database comprising up to a billion of compounds that can be easily prepared in one step from commercially available chemicals, which were found using the knowledge base of transformations (which was developed previously for the expert system LHASA [6] for synthesis planning (see http://cisrg.shef.ac.uk/shef2016/talks/oral2.pdf).

An interesting approach was implemented in the DOGS program [123] (type 5 in Table 1) for finding structures and synthesis pathways for analogues of a particular biologically active compound. In this case, the set of rules includes 58 easily performed organic reactions (click reactions) and the set of possible reactants comprises 25 144 accessible synthetic blocks. As a prioritization criterion, the evaluated measure of similarity with the target compound is used.

An approach for predicting the reaction products and reaction conditions using the knowledge graph was proposed by Segler and Waller.[124] By the knowledge graph is meant a bipartite graph with one group of nodes corresponding to chemical compounds (both reactants and products) and the other group of nodes corresponding to reactions between them. The reaction nodes can bear a label describing the reaction conditions. This strategy is based on the use of a very large knowledge graph containing 14.4×10^6 molecules and 8.2×10^6 reactions between them with indicated conditions as graph nodes. The edges of the knowledge graph specify whether compounds belong to reactants or products. The proposed procedure for predicting new reactions and reaction conditions is reduced to predicting additional edges and nodes of this graph. As an example, Fig. 3 presents a fragment of the knowledge graph consisting of five molecules (dark circles), which serve as reactants for four reactions (diamonds), forming four different products (light circles). In the knowledge graph, there is a path connecting the reactant molecules, with edges being numbered from *1* to *8*. For movement along this

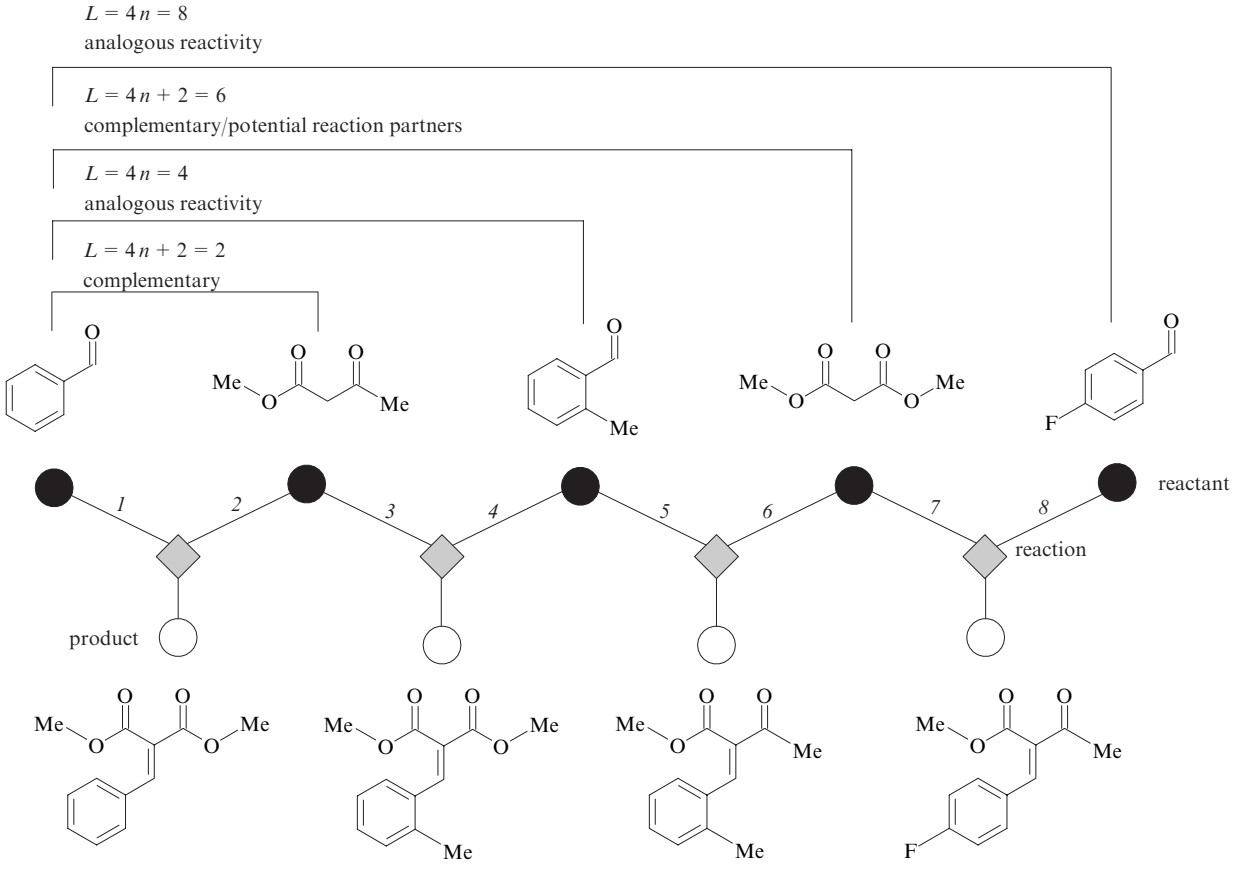


Figure 3. Example of analogues and complementary reactants determined on the path along the knowledge graph. The figure is adapted from Ref. 124;
(*1*–*8*) are the numbers of CGR edges.

path to model the reasoning by analogy, the following conditions should be fulfilled, according to the authors:

— two successively located reaction nodes should describe similar processes;

— two reaction nodes surrounding a reactant node should code the transformations that affect the same atoms in the reactant molecule.

The reaction similarity is evaluated by the Tanimoto index, which compares the corresponding binary fingerprints, which are, in turn, calculated as the difference between the product and reactant fingerprints. In this path, the reactants separated by one reaction node are considered to be complementary, as they are both reactants for this reaction. Meanwhile, the reactant nodes separated by four edges describe compounds that react in a similar way with one and the same compound located in the path between them; therefore, such compounds are coined as analogues. For a path of an arbitrary length, all reactants whose nodes are separated by $4n$ edges (n is any positive integer) are regarded as analogues and those separated by $4n+2$ edges are complementary.

In the considered approach, for predicting the possibility of reaction between two reactants, it is necessary to find a path in the knowledge graph that connects the nodes of the corresponding reactants and meets the two above-listed conditions. If at least one path is found, a reaction between these compounds is predicted, the reaction route being formed depending on the structures. The conditions for conduction of the predicted reaction are forecast taking account of the conditions indicated for reaction nodes in the path between the reactant nodes.

IV.2. Automatic generation of reaction rules

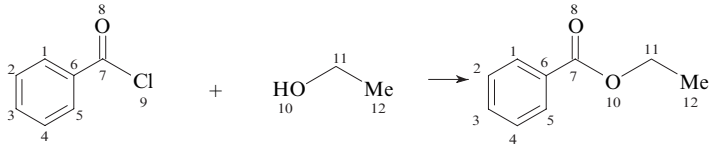
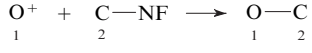
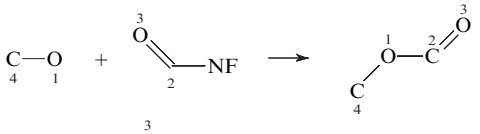
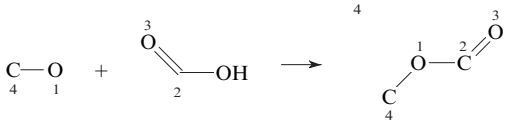
There are three key approaches to automatic generation of the reaction rules as generalized reactions. The first approach developed by Gelernter's research group [130] was formulated in the context of the artificial intelligence subject matter as a machine learning problem. The researchers used the inductive and deductive approaches to generation of chemical transformation rules to be entered into the knowledge base for the SYNCHEM synthesis discovery system [11] by generalizing examples from chemical reaction database. The inductive generalization accomplished by ISOLDA software gave rise to generalized reactions, such that all considered examples of reactions of a particular class present in the database were their specific instances. For all atoms of the reaction centre, the generalized reactions contain specifications of local physicochemical properties of the environment (for example, the presence of neighbouring electron-withdrawing groups) needed for the reaction. The ideas of the inductive approach formed the basis for the HORACE program developed by Gasteiger's research group [131, 132] and later used to form the knowledge bases in the software EROS [36, 37] (for prediction of the reaction products) and WODCA[14] (for retrosynthetic analysis).

The deductive generalization carried out for reactions of the specified type using the TRISTAN program gave rise to generalized reaction pathways, which, on the one hand, contained all atoms participating in at least one step of the hypothetical mechanism and, on the other hand, did not include atoms that do not affect any of the stages.

The second approach was developed by Kaufmann's research group [133] and implemented in the GRAMS software in which the generalized reactions are identified by finding the maximum common subgraph in a set of condensed graphs of reactions.[98, 99, 134] Subsequently this approach was extended to analysis of multistep synthesis pathways *via* the formation of generalized synthesis pathways (COSYMA).[135]

The third approach was implemented in the ARChem route designer system. It includes an extensive system for automatic formation of the reaction rules by analyzing the databases of chemical reactions.[58] For each reaction with a specified atom-to-atom mapping, a reaction centre is identified, comprising only the bonds involved in the reaction. The reaction centres are extended to cover neighbouring atoms, bonds and functional groups that affect the reaction mechanism. Generalization of the types of leaving groups affords the extended reaction centre (Table 2). This is followed by cluster analysis of all reactions from the database corresponding to each of the reaction centres, and the most typical reactions are taken from each cluster to form the final set of reaction rules.

Table 2. Formation of reaction rules by analysis of databases of chemical reactions using the ARChem system.[58]

Stage	Formation of reaction rules	Example
1	Reaction with partial atom-to-atom mapping	
2	Reaction core (NF is nucleofuge)	
3	Generalized extended reaction centre	
4	Example of final reaction rule	

Each of the ~105 000 unique chemical transformations isolated in this way from the database counting $\sim 10.8 \times 10^6$ reactions was subjected to statistical treatment to determine its utility (considering the number of representatives in the database) and possible variations of conditions and accompanying functional groups. All this information is used in the retrosynthetic analysis. The advantage of the automatically generated rules over the rules formed manually is much wider coverage of the organic chemistry data. Unfortunately, no additional data about the reaction mechanisms and principles of the electronic structure of molecules are employed according to this approach. However, presumably, this drawback could be eliminated with upgrading of software systems, updating of databases and implementation of artificial intelligence techniques for their analysis.

IV.3. Prioritization of reaction rules

An important task of computer synthesis systems is to predict which of alternative pathways would predominate for a given reaction. Hence, they perform two closely interrelated functions — prioritization of reaction rules and prediction of the reactivity of organic compounds.

The first studies along this line were carried out in the context of the strategy based on expert system rules. Whereas in the classic CAMEO system developed by Jorgensen's research group[34,35] the rules were formed as a program code simulating the reasoning of an experienced organic chemist (with possible use of the results of simple physicochemical calculations) separately for different classes of organic reactions, in the SOPHIA system developed by Satoh and Funatsu,[39,40] this was done using the knowledge base that was automatically formed from the reaction database.

The first studies utilizing machine learning methods were carried out in the mid-1980s by Gasteiger *et al.*[136] and comprised the construction of classification models (*k* nearest neighbours, linear discriminant analysis and logic regression methods), which predicted the feasibility of the heterolytic bond cleavage on the basis of local physicochemical characteristics of atoms. Subsequently the same sets of descriptors were applied to construct similar models by combining two types of artificial neural networks — multilayer perceptron and Kohonen self-organizing map.[137] To make the sets of descriptors related to the properties of chemical bonds applicable to characterization of molecules of an arbitrary size, it was proposed to use combined descriptors, which were called MOLMAP. It was shown[138] that MOLMAP can serve as an efficient tool for predicting the ability of organic compounds to react with particular chemical reagents. Indeed, the classification models formed by the 'random forest' method were able to predict whether a test molecule would react with Bu^nNH_2 or $Na(CN)BH_3$.

Several studies published in 1990 and 1991 were devoted to the use of artificial neural networks to predict reaction pathways for organic compounds within simple model systems. For example, Elrod *et al.*[139,140] used a neural network to predict the products of electrophilic substitution of monosubstituted benzenes and the regiochemistry of classic Markovnikov, Diels – Alder and Zaitsev reactions. A similar method was applied by Kvasnicka and Pospichal[141] to predict the products of aromatic electrophilic substitution of benzene derivatives. Using simple examples, the authors demonstrated the applicability of artificial neural networks to organic chemistry problems.

Quite a promising approach to prioritization of reaction rules is implemented in the ReactionPredictor system, which utilizes neural networks to predict the products of multistep reactions by considering the reaction mechanisms and comparing the feasibility of the formal electron transfer schemes in the elementary steps.[142] An advanced version of ReactionPredictor is applicable not only to reactions that follow a polar mechanism, but also to reactions occurring by pericyclic and radical mechanisms.[143] A method was also proposed for combining the ReactionPredictor system based on machine learning techniques with quantum chemical calculations.[144] In this case, transition states and activation barriers are calculated for the most favourable reaction steps predicted by the system; the calculation results are entered back into the ReactionPredictor and are

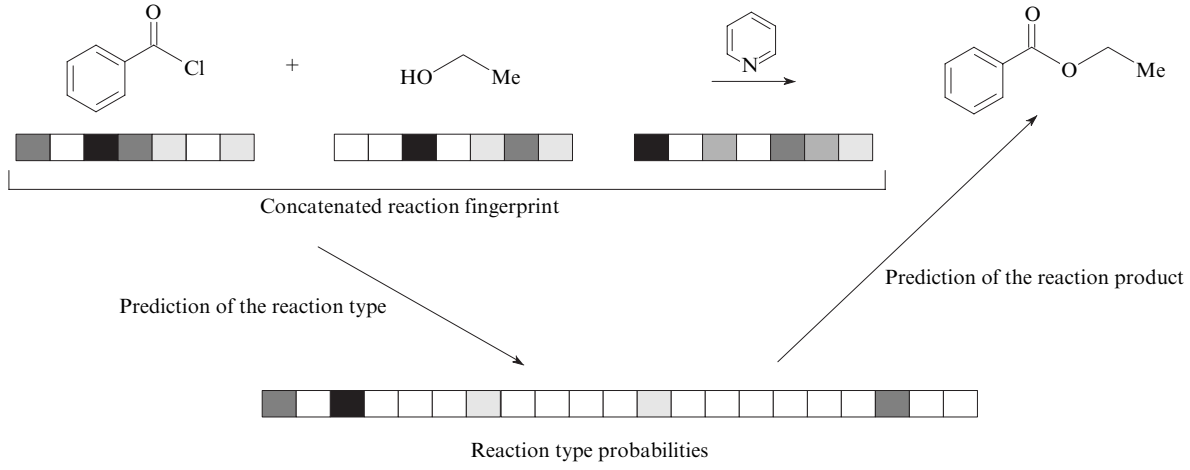


Figure 4. Scheme of the neural network method for prediction of reactions. The figure is adapted from Ref. 145.
The neural network the input of which is the vector formed by concatenation of the vectors describing the reactants and reaction conditions forms the probability vector for alternative reaction pathways at the output and thus enables prediction of the reaction products.

used for additional training of neural networks. This increases the reliability of prediction.

Yet another method of neural network prediction of reaction products was proposed by Aspuru-Guzik's research group.[145] The strategy was tested on a database comprising only 17 types of organic reactions; however, it can quite be extended to a markedly wider range of reactions. The authors used special neural network fingerprints derived from chemical structures by means of a convolutional neural network operating on the deep learning principle.[146] The prediction affords the probability vectors for each of the 17 types of reactions (Fig. 4).

Coley *et al.*[147] developed a general-purpose and more perfect program for predicting the products of organic reactions, which combines the reaction rules and deep learning-based neural networks. The modular neural network underlying the system was trained with the data on chemical reactions taken from a large US patent database ($> 10^6$ reactions). The same database was used for automatic retrieval of the reactivity rules, which were applied to each reaction to generate the possible alternative reaction pathways. This resulted in $> 5 \times 10^6$ generated reactions. They served as 'negative examples' for building of the neural network model, which could predict the probability for each reaction pathway. The products for a specified set of reactants are predicted *via* generation of possible products using the set of reactivity rules and selection of those products that are chosen as most probable by the neural network.

V. Analysis and visualization of reaction data

V.1. Classification of reaction types

The term 'classification of reaction types' is usually referred to several different, but closely related concepts. First of all, this is the regulation of terms used to describe various types of chemical reactions at the conceptual level, which refers to the knowledge background of synthetic organic chemistry. In the field of artificial intelligence, the agreements that define the basic notions of the subject matter (in this case, synthetic organic chemistry) and establish the relationships between them are called ontologies. Usually, ontologies include taxonomies, which are hierarchical classification systems for some objects (in this case, chemical reactions). The reaction classification schemes in synthetic organic chemistry are traditionally based on consideration of formal bond redistribution schemes.

Furthermore, '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.

V.1.a. Classification based on formal bond redistribution schemes

During a reaction, changes usually affect relatively small parts of the reacting molecules. In view of the common natures of these changes, reactions can be grouped together according to the type. Indeed, considering generalized types of bond changes, Thelheimer[148] was able to distinguish four general types of reaction — addition, elimination, rearrangement and exchange (subsequently more often called substitution).

Attempts to elaborate a more specific classification of reactions resulted in the development of a multitude of approaches that considered the changes in the bond orders, valences and formal charges of atoms in the reaction centre. Among them, noteworthy are the approaches of Hendrickson,[85] Arens,[86–88] Zefirov,[89–91] Vladutz,[94,95] Dugunji and Ugi[70] and some other. The approaches of this type to classification of organic reactions have been surveyed in detail.[149,150]

V.1.b. Ontologies of chemical reactions

It was noticed long ago that reactions of similar types can be grouped into more general classes; this allows the construction of a hierarchical classification system — taxonomy of reactions, which forms an important part of their ontology. The attempts at constructing ontologies have been made using two types of approaches. In the first type, which can be called inductive, researchers proceeded 'from the data' and tried to arrange similar types of chemical reactions in groups in accordance with chemical logic. An example of this type of ontology is the three-level classification system proposed by Carey *et al.*[151] and later upgraded by Roughley and Jordan.[152] This system formed the basis for the RXNO reaction ontology (http://github.com/rsc-ontologies/rxno), proposed, used and developed by the UK Royal Society of Chemistry. For reaction classification in accordance with RXNO ontology, the rule-based NameRxn system was devised (http://www.nextmovesoftware. com/namerxn.html);[78] this system was used[153] to investigate the popularity of reactions of various types in medicinal chemistry. The complexity of constructing 'chemistry-based' ontologies is in the fact that one reaction can simultaneously belong to several groups of reactions at a higher generalization level.

It is very difficult to construct an unambiguous and exhaustive hierarchy. The second, rule-controlled approach, which can be considered as deductive, is potentially free from this drawback. According to this method, a set of rules indicating the possibility of electron or bond redistribution during the reaction is formulated based on the consideration of the formal valences of atoms. Representations of reactions with different levels of generalization are thus generated (for example, at the highest generalization level, only the way of electron redistribution is considered, while at a lower level, changes in the bond orders are included, and at the lowest level, symbols of atoms are added), which makes it possible to compose an hierarchical system for the reaction classification. Among the most consistent systems of this type, mention may be made of the Tratch – Zefirov approach.[154] The drawback of this approach is that the results may deviate from chemists' views on classes and designations of reactions.

V.1.c. Problem of the automatic classification of reactions in the databases

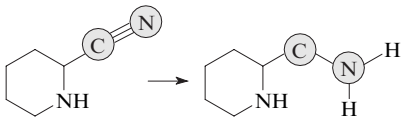
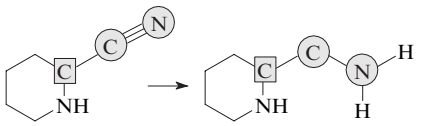
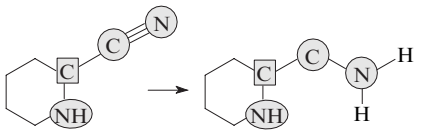
The task of classification of reactions stored in databases according to types includes determination of the type of reaction on the basis of the encoded reaction transformation. This solves four main problems:

— identification of what types of reactions are present in a database and quantitative characterization of these reactions;

— performing search of the database for a specified type of reactions;

— collecting reactions of the same type into groups for the subsequent identification of reaction signatures or retrosynthetic rules;

Table 3. Levels of consideration of the reaction centre used in the ICClassify approach.[111]

Sphere	Level	Example
Zero	Broad (only the reaction centre atoms are included)	
First	Medium (the reaction centre atoms and α-atoms except for hydrogens are included)	
Second	Narrow (the reaction centre atoms and α- and β-atoms except for hydrogens are included)	

— predicting characteristics of a chemical reaction on the basis of known characteristics of other reactions of the same type.

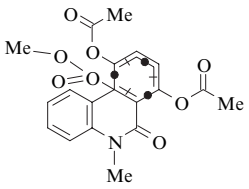
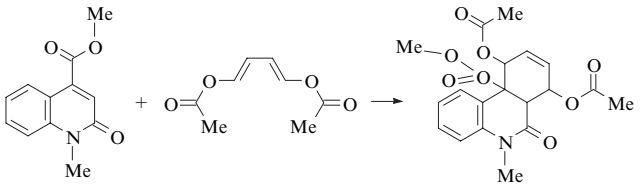
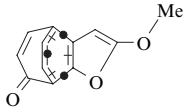
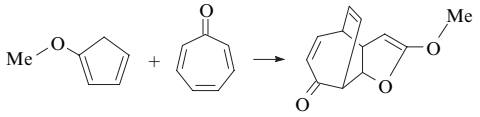
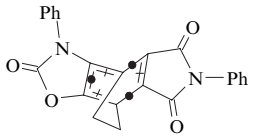
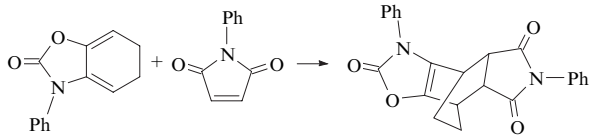
Classification methods based on formal bond redistribution schemes usually do not take into account the chemical environment of the reaction centre. This is typical of data-driven methods in which the classes and hierarchy of reactions arise upon analysis of reaction data. One of the most popular classification systems of this type is ICClassify.[111] For classification, atoms at which bond orders change are identified (Table 3). Reactions of a given type, *e.g.*, reduction of the cyano group presented in Table 3, would always involve similar changes in the reaction centre (the C≡N bond of reactants is converted to the C–N bond of the products). However, in some cases, it is important to consider also the local environment, for example, to find out whether the cyano group is attached to an aliphatic or an aromatic atom. Therefore, several generalization levels for the reaction centre are used, differing in the extent of inclusion of the chemical environment of the reacting group. Altogether, three types of fragments are considered for each reaction: at the broad level, only the reaction centre atoms are considered; at the medium level, the local environment is also included (α-atoms except for hydrogen atoms); and at the most narrow generalization level, the reaction fragment to be considered comprises the reaction centre and the surrounding α- and β-atoms (see Table 3). It is evident that several medium-level fragments can correspond to one broad-level fragment and that several narrow-level fragments can correspond to one medium-level fragment. For better compliance with the chemical logic, the symbols of some chemical elements are replaced by general symbols indicating a certain group of elements (*e.g.*, alkali metals, alkaline earth metals, noble gases, transition metals and so on).

The fragment corresponding to the reaction centre at a certain consideration level is written as the reaction centre hashcode. The hashcode is a concise representation of the reaction centre, which unambiguously describes it, but cannot be converted into a structural view. The reaction hashcodes are calculated by the procedure based on the modified Morgan algorithm[155] with allowance made for types and characteristics of atoms.[111] Thus, reactions designated by the same hashcode belong to the same type at the chosen consideration level. This allows using the ICClassify system for analyzing the diversity of reactions in a database, clusterization of reactions, grouping together of the list of reactions formed by a search system, filtering a database and search for reactions of a desired type. In addition, this algorithm serves for the search for similar reactions in the SciFinder system.[111] In this case, the reactions that have identical hashcodes at the narrow level are considered to be the most similar, while reactions that have coincident broad-level hashcodes are regarded as the least similar. The same algorithm, but involving consideration of the reaction centre environment at six levels (up to the fifth sphere) is used in the Reaxys information system.

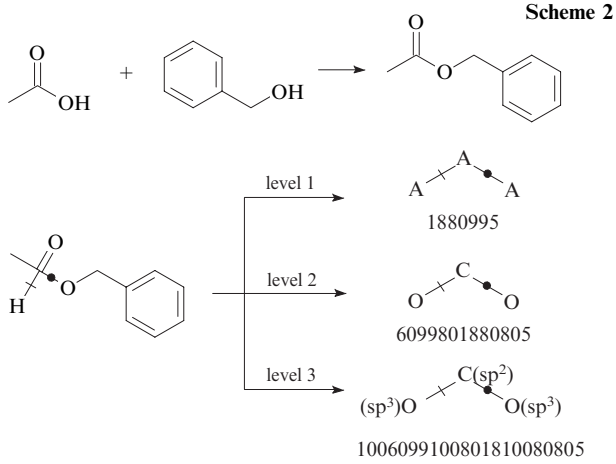
By analyzing the dynamic bonds in the condensed graph of reactions, one can also unambiguously identify the reaction centre, which is unique for each type of reactions. Similarly to early studies of Jauffret and co-workers[98, 99, 134] and de Luca *et al.*,[156] CGR fragments that unambiguously characterize reactions of a given class (reaction signatures) were identified. If reactions are presented in a database as CGR, then reactions of this type can be found by a substructure search using the signatures. Furthermore, CGR can be used for the search for similar reactions on the basis of the 'neighbourhood behaviour' concept.[157] Examples of the signatures of the Diels–Alder reactions found by de Luca *et al.*[156] are presented in Table 4.

Further development of this approach is the CGR-HASH algorithm, which allows representing the reaction centre as a hashcode, the latter being a way of coding of reaction signatures.[158] Three description levels for the reaction centre, substantially differing from the ICClassify classification, have been proposed. The first level takes into account only types of bonds, but types of atoms are neglected. The second level takes into account the types of changing bonds and the symbols of atoms. This level resembles the broad level involved in the ICClassify system,[111] and also it can be considered as the hashcode representation of the 'skeletal schemes' proposed by Vladutz *et al.*[94] The third level considers additional parameters of atoms (hybridization, aromaticity) and types of bond change. The generation of reaction signatures (hashcodes) for the reaction centre is illustrated by Scheme 2, in which the reaction equation is given at the top and the CGR and hashcodes for different levels are given below (A is any

Table 4. Examples of signatures derived from CGR corresponding to Diels – Alder reactions.[156]

Signature	CGR	Chemical reaction
		
		
		

Scheme 2



atom). The atoms of the first, second and other neighbourhoods can be added to the reaction centre, like in the ICClassify system. By adding atoms from all reaction centre neighbourhoods possible for a given reaction, one can represent the whole reaction as a single hashcode, which can be used for looking for reaction duplicates in a database.

Christ *et al.*[159] proposed a 'data-driven' approach to classification, which is based on the formation of rules by means of representing reactions as SMIRKS strings (http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html). Sello and Termini [160, 161] have applied quite an unusual method for hierarchical classification of reactions, which, unlike the techniques mentioned above, is based on the use of physicochemical descriptors calculated from empirical equations rather than on the structural description of reactions.

Machine learning methods are widely used in the classification of reactions by types; most popular are dimensionality reduction methods, such as Kohonen self-organizing maps (see below).

The classification of reactions by types was carried out using machine learning with various modifications of reaction difference descriptors, which appear to be the most natural choice for solving this issue. Schneider *et al.*[78] reported classification of the reactions according to the RXNO ontology using quite a number of machine learning methods and difference descriptors of reactions. This approach provided a successful classification of some reactions that had not been assigned to a particular type in terms of the rule-based NameRxn approach (http://www.nextmovesoftware.com/namerxn.html). Patent claims [79] describe the application of difference fingerprints to identifying reactions of similar type and demonstrate the advantages of this method over Daylight software products (Daylight Chemical Information Systems Inc., http://www.daylight.com). Hu *et al.*[77] performed classification of enzymes according to the Enzyme Commission (EC) rules, using the known types of reactions they catalyze and the nearest neighbour method. The similarity of reaction mechanisms has been used to study the similarity of enzyme-catalyzed reactions.[162] In this approach, the attention is focused on the mechanistic details of the processes.

The condensed graph of reactions is also a convenient tool for structural similarity-based prioritization of reactions. By using the representation of reactions as CGR and rows of fragment descriptors, one can calculate the standard similarity indices between the reactions,[100] which was also reflected in a patent.[163]

V.2. Mapping of chemical reaction space

Out of numerous methods for visualization of chemical data,[164] only two are applied to analyze the chemical reaction space — Kohonen self-organizing maps [165] and generative topographic maps.[166, 167] The Kohonen maps result from calculations involving neural networks of a specific sort (Kohonen networks). These maps project multidimensional objects onto a neural network (map cells) located on a plane in such a way that similar objects are mapped on the same or neighbouring neurons.[168]

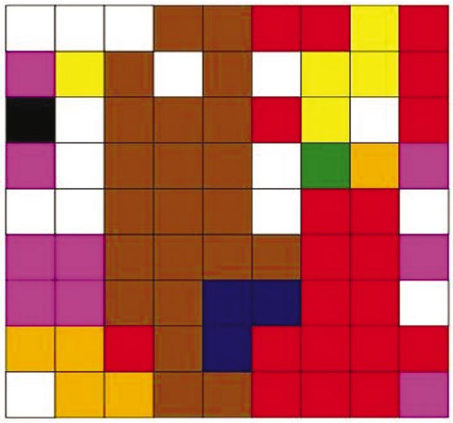


Figure 5. Classification of 8544 reactions using the Kohonen map.[169]
The colours correspond to various types of reactions: Sonogashira (brown), dihydroxylation (yellow), nucleophilic substitution (blue), epoxidation (green), metathesis (orange), domino – Heck – Diels – Alder reactions (black). The red and violet colours mark the Diels – Alder reactions in which the reaction centre comprises carbon atoms and heteroatoms or only carbon atoms, respectively.

An example of analysis of the chemical reaction space using the Kohonen maps for a sample comprising 8544 reactions of various types is shown in Fig. 5. All reactions have been represented as CGR, for which fragment descriptors have been calculated. The Kohonen map cells are filled with a colour corresponding to the type of most of the mapped reactions. It can be seen that reactions of the same type form clusters covering one or several cells of the map.[169] In most cases, cells of the same colour occupy extended areas in the map.

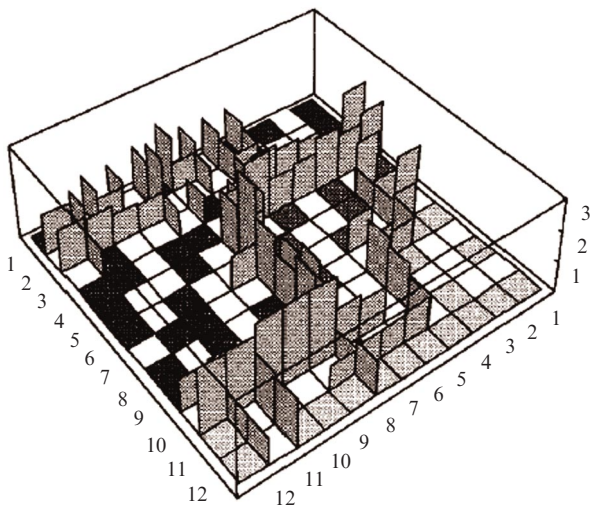


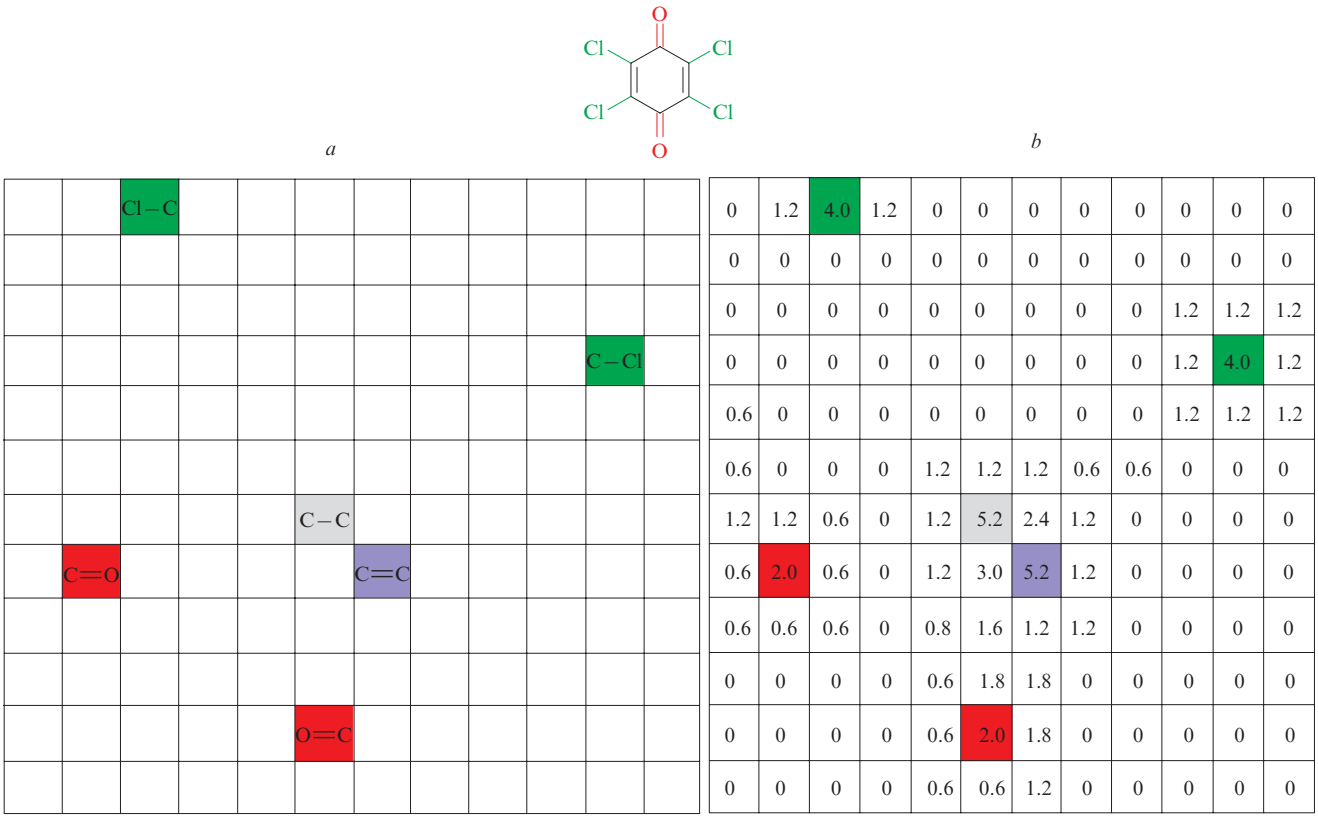
Figure 6. 3D landscape of reactions.[170]
The digits at the axes in the map plane are the row and column numbers in the network determining the positions of nodes, and the vertical scale encodes the distance between the clusters. It can be seen that three 'valleys' separated by 'walls' ('mountain ridges') indicate clearly different reaction types in the Kohonen map.

An interesting method of topological analysis of the distribution of reactions in the Kohonen maps has been reported.[170] The landscapes of reactions, with vertical coordinates encoding the distances between the clusters formed by network nodes, were constructed in the map based on local physicochemical characteristics of atoms. In this case, reactions of the same type form the 'valleys' separated by 'mountain ridges'. An example of such landscape is shown in Fig. 6, where three 'valleys' are seen corresponding to the nucleophilic aliphatic substitution of acyl chlorides (black), acylation of C=C bonds (grey) and acylation of arenes (light grey).

The idea of using local physicochemical descriptors for classification and prediction of reactions by means of the Kohonen maps has been further developed by Satoh *et al.*[171] and Hiroko *et al.*,[172] who calculated the local physicochemical characteristics of atoms as the average energy of their electrostatic or van der Waals interaction with the test atoms [H^+ and $C(sp^3)$, respectively] placed on the open part of the van der Waals surface. The local characteristics of atoms obtained in this way, which were called FRAU (field-characterization for reaction analysis and understanding) descriptors, numerically describe the fields around the reacting molecules. Unlike the molecular field descriptors used in the 3D QSAR methods (*e.g.*, CoMFA [173]), the FRAU descriptors are calculated without the spatial alignment of molecules; the values of these descriptors depend little on the choice of molecular conformations. Therefore, the FRAU descriptors can be applied to structurally different reactant molecules. The Kohonen map formed by means of such descriptors for 39 reactants containing magnesium and boron atoms was used for classification and prediction of the types of reactions involving these reactants.[171, 172]

Machine learning methods can be efficiently applied for the automatic classification of organic reactions. The traditional approach to determination of the type of reaction is based on the consideration of bond redistribution. This, in turn, requires the knowledge of which atoms in the products match particular atoms in the reactants (atom-to-atom mapping). Since this information is often missing in databases, for assignment of the types of chemical reactions, classification problem should be solved in terms of a model that does not require atom-to-atom mapping. This problem was solved for the first time by introducing special MOLMAP reaction descriptors;[174] the descriptors were derived and utilized for the construction of classification models by means of machine learning techniques. The MOLMAP descriptors for molecules are found [175] by constructing the Kohonen map. For this purpose, the Kohonen network is trained using a sample of diverse chemical bonds represented by seven local physicochemical descriptors (electronegativity, atomic charge, total atomic charge, bond polarity, average bond polarizability, resonance stabilization and bond dissociation energy) proposed by Gasteiger and co-workers [137] for modelling the reactivity of organic compounds.

For the calculation of MOLMAP descriptors, the information about each bond in the molecule is put onto a map formed in this way (Fig. 7 *a*), in which the occupied cells are assigned the values corresponding to the numbers of bonds of this type in the molecule. Then the cell occupancies are recalculated to take into account the possibility for a bond to get into neighbouring cells (Fig. 7 *b*). Finally, the vector



0	1.2	4.0	1.2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1.2	1.2	1.2
0	0	0	0	0	0	0	0	0	1.2	4.0	1.2
0.6	0	0	0	0	0	0	0	0	1.2	1.2	1.2
0.6	0	0	0	1.2	1.2	1.2	0.6	0.6	0	0	0
1.2	1.2	0.6	0	1.2	5.2	2.4	1.2	0	0	0	0
0.6	2.0	0.6	0	1.2	3.0	5.2	1.2	0	0	0	0
0.6	0.6	0.6	0	0.8	1.6	1.2	1.2	0	0	0	0
0	0	0	0	0.6	1.8	1.8	0	0	0	0	0
0	0	0	0	0.6	2.0	1.8	0	0	0	0	0
0	0	0	0	0.6	0.6	1.2	0	0	0	0	0

Figure 7. Calculation of the MOLMAP descriptors for a molecule. The figure is adapted from Ref. 176.
(*a*) Mapping of the chemical bonds of the molecule on the Kohonen network cells, (*b*) conversion of the cell occupancy to MOLMAP descriptors.

of MOLMAP descriptors for a molecule is formed as a concatenation of the matrix columns.[176]

The MOLMAP descriptors for reactions are calculated as the difference between the MOLMAP descriptors of the reaction products and reactants.[176] The number of MOLMAP descriptors obtained in this way is equal to the number of Kohonen's cells and does not depend on the number of atoms or bonds in the molecules; therefore, they can be represented as fixed-length vectors and, hence, applied in conjunction with standard machine learning methods for the construction of classification, regression and other types of models and also for data visualization. Zhang *et al.*[176] used the MOLMAP descriptors in conjunction with the random forest method[177] to classify organic reactions. They have also been used to classify metabolic reactions by assigning the EC numbers to enzymatic reactions involved in metabolism[82,83] and for automatic assignment of EC numbers to arbitrary enzymatic reactions.[178] The selection of the random forest method for machine learning in the construction of the classification model is caused by its ability to efficiently operate large numbers of descriptors, although only few of them are relevant. The MOLMAP descriptors were successfully used to generate the Kohonen map suitable for classification of metabolic reactions (Fig. 8).

The automatic encoding of organic reactions accomplished by means of MOLMAP descriptors was also used for mapping and classification of whole metabolic pathways.[179] For calculating the MOLMAP descriptors of the

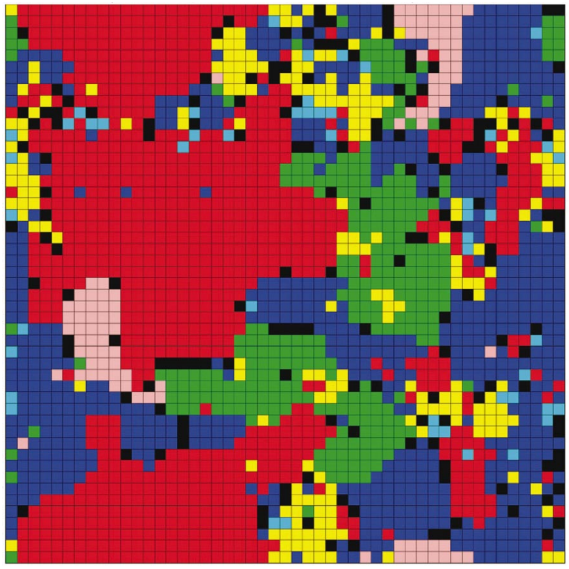


Figure 8. Kohonen map constructed for 7482 enzymatic reactions involved in the metabolism. The figure is adapted from Ref. 82.
Enzymes catalyzing the reactions: oxidoreductases (red), transferases (dark blue), hydrolases (green), isomerases (light blue), ligases (pink); black colour bears no definite correspondence.

next level in which the metabolic pathway is described by a fixed-length vector, all reactions involved in metabolism are placed on the map trained using metabolic reaction databases, and the descriptors are determined from cell occupancies. The descriptors thus obtained can be applied for mapping of various metabolic pathways of various organisms using the next-level Kohonen maps.

Yet another popular approach to visualization of chemical reaction data is generative topographic mapping (GTM), which is a probabilistic analogue of the Kohonen maps. The GTM method was introduced by Bishop *et al.*[166, 167] as a non-linear method for decreasing dimensionality, with data topology being retained. This method approximates the data distribution by a smooth manifold, which can be represented as a flexible sheet of paper located in the data space. The mapping of chemical items (molecules, reactions) on this sheet gives maps describing their relative positions.[164, 180] The GTM method has been used previously for mapping the chemical space of molecules[180 – 186] and for 'activity landscaping', which serves for visualization of quantitative structure – activity relationships.[187]

Structural descriptors are most useful for mapping of the chemical reaction space. The maps provide visual examination of the structural diversity of the set of reactions and identify groups of similar reactions. The GTM approach has been used to form a map describing a large set of hydrogenation reactions (see Fig. 9).[188] The reactions were represented as CGR for which the ISIDA fragment descriptors were calculated.[100, 189] First, using the approach reported by Lin *et al.*,[108] the reactions in which protecting groups are removed were identified. Thus, the reactions were labelled according to the type of protecting group being removed or were designated as 'unclassified' if none of the considered protecting groups was removed in the reaction. The analysis covered > 70 000 reactions from the Reaxys database.[108] Similar reactions formed compact clusters, with reactions with different chemical neighbourhoods being gathered in different areas. For example, zone 1 in Fig. 9 corresponds to removal of benzyl groups at secondary aliphatic carbon atoms, zone 3 corresponds to elimination of one benzyl group in a primary alcohol, and zone 2 reflects complex reactions that include the removal of numerous benzyl protecting groups. Generally, more complex reactions tend to accumulate in the upper part of the map. The reactions that include reduction of double bonds (designated as unclassified) and are accompanied by removal of some protecting groups (zones 4 and 5) got into the left-hand side of the map. Meanwhile, reactions of relatively rare types (< 100 reactions) may form no compact zones.

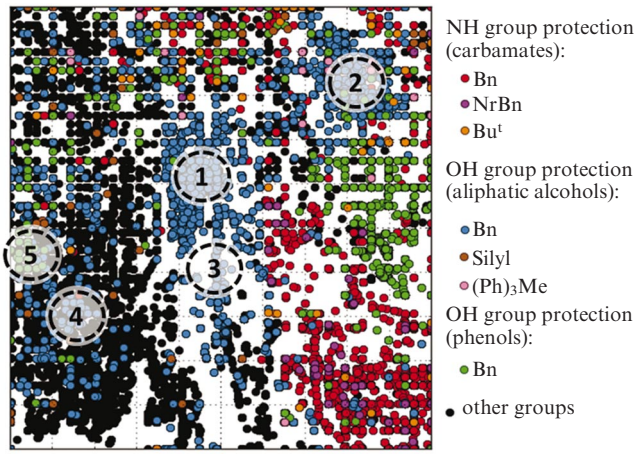


Figure 9. GTM map for hydrogenation reactions.[188]
Protecting groups: Bn is benzyl, NrBn is nitrobenzyl, Bu^t is *tert*-butyl, Silyl is silyl, (Ph)$_3$Me is triphenylmethyl. Less abundant groups are omitted.

VI. Prediction of quantitative characteristics of the reactions

VI.1. Kinetic and thermodynamic parameters

The works on the quantitative prediction of reaction parameters (rate constants or yields) by machine learning methods include both classic approaches based on linear free energy relationships and construction of regression models using advanced statistical analysis and machine learning techniques.

The pioneering studies on the quantitative prediction of rate constants for chemical reactions were based on the linear free energy relationships and were published by Hammett,[190] Taft[191] and Palm[192] in the 1930s – 1960s. These studies underlay the classic physical organic chemistry. In all cases, the quantitative modelling was carried out for very narrow series of compounds that had the same skeleton and reacted under strictly identical conditions and differed only by one substituent. The rate constants were predicted by a linear equation with a single descriptor — the substituent constant § belonging to one of empirical scales. Although these equations have played an important part in the formation of physical organic chemistry and views on reaction mechanisms, they could not find practical use as prediction tools because of a too narrow range of compounds used for modelling and a limited range of substituents with known values of required constants. Nevertheless, with further development of this strategy and wide implementation of computers in chemical research, in the 1970s – 1980s, the series of compounds involved in analysis were extended to include variation of several substituents. In addition, the first procedures for describing the solvent effect were developed. This allowed Palm[193] to write about a quantitative theory of organic reactions.

The foregoing created prerequisites for further development of works dealing with quantitative relationships between the structures and reactivity of chemical compounds. These relationships are designated by the abbreviation QSRR (quantitative structure – reactivity relationship). Currently, by QSRR is meant a strategy according to which the molecules of one or several reactants (or the whole reaction scheme) are described by a set of descriptors; this set is used for quantitative prediction of one reactivity parameter (usually, rate constant for the reaction) by application of a model constructed by machine learning. The QSRR technique was employed for the first time back in 1989 in the study of Mishra *et al.*;[194] however, in full measure it was implemented only 11 years later.

§ One of the constants present in the Hammett, Hammett – Taft and other similar equations, which underlie the linear free energy relationships.

In the 1980s – 1990s, several papers were published addressing the prediction of rate constants for gas-phase reactions of $^{\bullet}OH$ free radicals with organic compounds, considering group contributions [195] and empirical equations based on quantum chemical characteristics.[196] In this case, the sets of compounds comprised hundreds of structures and were not limited to narrow series; however, they still referred only to gas-phase reactions of one type.

In the beginning of the 2000s, Katritzky's group [197, 198] reported two first studies on the QSRR prediction of the kinetic characteristics of reactions. One publication [197] described the prediction of the rate constant for gas-phase homolysis of C – X bonds, while the other paper [198] addressed the rate constant for chain transfer in styrene polymerization, induced by free radicals generated by various organic compounds. In both cases, the descriptors were calculated using the CODESSA program,[199] and multiple linear regression with selection of descriptors was used as the machine learning method. Since the structure of only one reactant was varied in these studies, with the reaction conditions remaining invariable, the QSRR procedure was reduced to the standard QSAR/QSPR (quantitative structure – activity relationships/quantitative structure – property relationships) methodology for the search for quantitative correlations between the structure and property of a chemical compound.[200, 201] The same expedient of reducing QSRR to QSPR was used later for prediction of rate constants for hydrolysis of esters.[202, 203] The same approach was employed to predict the degradation of organic compounds upon oxidation in the upper atmosphere [204 – 208] in which the set of reactions is not always exactly known. The publications on modelling the reactivity of organic compounds to ozone and hydroxyl radical in the atmosphere,[209] to ozone and sulfate radical in water [210] and to nitrate radical depending on temperature [211] can be assigned to the same group of studies.

A publication [212] of 2002 describing the model for predicting the rate constant for acid hydrolysis of arbitrary esters under various conditions (various temperatures, solvents or solvent mixtures with water) is a milestone in the development of QSRR approaches. The set used for solvents included both local quantum chemical descriptors characterizing the properties of atoms near the reaction centre and descriptors of reaction conditions (temperature and proportion of solvent in a mixture with water) as well as four descriptors characterizing the physicochemical properties of the solvent. The machine learning method used in this case represented a multilayer neural network implemented in the NASAWIN program package [213] employed according to the general methodology of structure – conditions – properties relationships.[214] The interpretation of the neural network model by the procedure proposed by Baskin *et al.*[215] demonstrated the exact correspondence to the known principles determining the reactivity for this reactions, thus indicating the ability of neural networks to extract knowledge from data. Subsequently the model was modified by replacing quantum chemical descriptors by labelled fragment descriptors, which improved the predictive power.[216] Thus, fragment descriptors can be used instead of the traditional local quantum chemical descriptors for modelling the reactivity of organic compounds.

The next important step in the development of studies along this line is the construction of a neural network model for predicting the rate constants for nucleophilic substitution by the S_N2 mechanism with variable structures of the nucleophilic reagent and the electrophilic substrate and with variable reaction conditions.[217] The set of descriptors comprised either local quantum chemical or fragment descriptors for both the nucleophile and electrophile molecules and the reaction conditions in accordance with the bimolecular QSPR concept.[218] A neural network model for predicting rate constants for nucleophilic substitution by the S_N1 mechanism was also constructed.[219]

An interesting development of the QSRR strategy is the model for quantitative prediction of the reactivity index in the Michael addition with variable structures of the substrate (24 variants) and the enone (4 variants) and three different catalysts.[220] The models were constructed separately for each of the catalysts using the quantum chemical descriptors calculated for the substrate and the enone, with multiple linear regression and partial least squares being used as the machine learning methods with selection of descriptors using the genetic algorithm.

The approach involving the use of the condensed graph of reaction makes it possible to encode a reaction transformation as a numerical vector in approximately the same way as for a molecule. When fragment descriptors are used, each element of the vector encodes the number of times a particular fragment occurs in a chemical structure. In the case of ISIDA descriptors, these are chains of atoms and/or bonds of a definite length or atoms with the local environment.[100, 189]

Madzhidov *et al.*[221] proposed a CGR-based procedure for predicting the rate constants for chemical reactions; the procedure was considered by the example of bimolecular nucleophilic substitution reaction. Since the rate constant depends not only on the reactant and product structures, but also on the conditions, the descriptors for the solvent and the temperature were added to the vector of descriptors describing the transformation (ISIDA fragment descriptors for CGR). As the solvent descriptors, the Kamlet – Taft and Catalan parameters were used, as well as other parameters depending on the solvent dielectric permittivity and the refractive index, characterizing its polarity and polarizability. In order to describe water – organic mixtures, which are often used, a descriptor encoding the mole fraction of the organic solvent in a mixture with water was added to the solvent descriptors. The temperature was described using the reciprocal of reaction temperature (in Kelvin degrees). The relationship between the logarithm of the rate and the obtained vector of descriptors was modelled by the random forest machine learning method.[222] This method was shown to be able to predict a rate constant to an accuracy close to the error of measurements.

Nugmanov *et al.*[223] continued to develop this approach by applying it to bimolecular nucleophilic substitution reactions giving practically important azides. A specific feature of such reactions is that the nucleophile (azide anion) has a counter-ion (alkali metal, ammonium, alkylammonium or tetraphenylarsonium cation). It was shown that the nature of the counter-ion affects the reaction rate. An attempt was made to find an optimal representation for the counter-ion based on quantum chemical and fragment descriptors. With the latter, models with better predictive power were constructed by the support vector method. The authors performed an independent validation of the model quality on the basis of available kinetic data on the rate of calix[4]arene azide formation. It was shown that in the

absence of steric hindrance, the rate constant can be properly predicted; however, in the presence of steric effects, which are neglected in this model, the prediction of rate constants was expectedly fairly unreliable.

The described strategy of modelling the quantitative characteristics of chemical reactions in water, organic solvents and water–organic mixtures based on the use of CGR technique and physicochemical parameters of the solvents was also applied to predict rate constants for bimolecular elimination[224] and the equilibrium constants of tautomerization reactions.[225]

VI.2. Optimal reaction conditions

Whereas quite a few tools have been developed for planning the synthetic strategy towards a chemical compound, virtually no approaches for predicting the optimal conditions of chemical reactions have existed until recently. The paper of Struebing *et al.*[226] was among the first publications focused on the prediction of the reaction conditions, in particular, the choice of the optimal solvent. The authors chose the optimal solvent for the Menshutkin reaction by a rather complex procedure that included the following stages:

— quantum chemical calculation of the activation barrier of the reaction for a series of solvents;

— construction of the regression model to relate the rate constant of the reaction to solvent parameters;

— choice of the best solvent in terms of the resulting model;

— correction of the obtained regression model by adding information for the reaction in a new solvent.

The procedure was repeated until the solvent of choice was found. This approach to the choice of the optimal solvent for the reaction is largely based on first principle calculations and, therefore, it is computationally rather time-consuming.

One more approach to prediction of the optimal reaction conditions is based on the assumption that the choice of the best conditions such as the catalyst and the solvent can be treated as a classification task. In this case, the reaction itself is considered as a chemical object whose property must be predicted, while the optimal solvent (catalyst) is the property to be predicted. Marcou *et al.*[84] constructed the model for predicting the types of the solvent and the catalyst for the Michael reaction. For this purpose, they collected 222 examples of reactions proceeding under

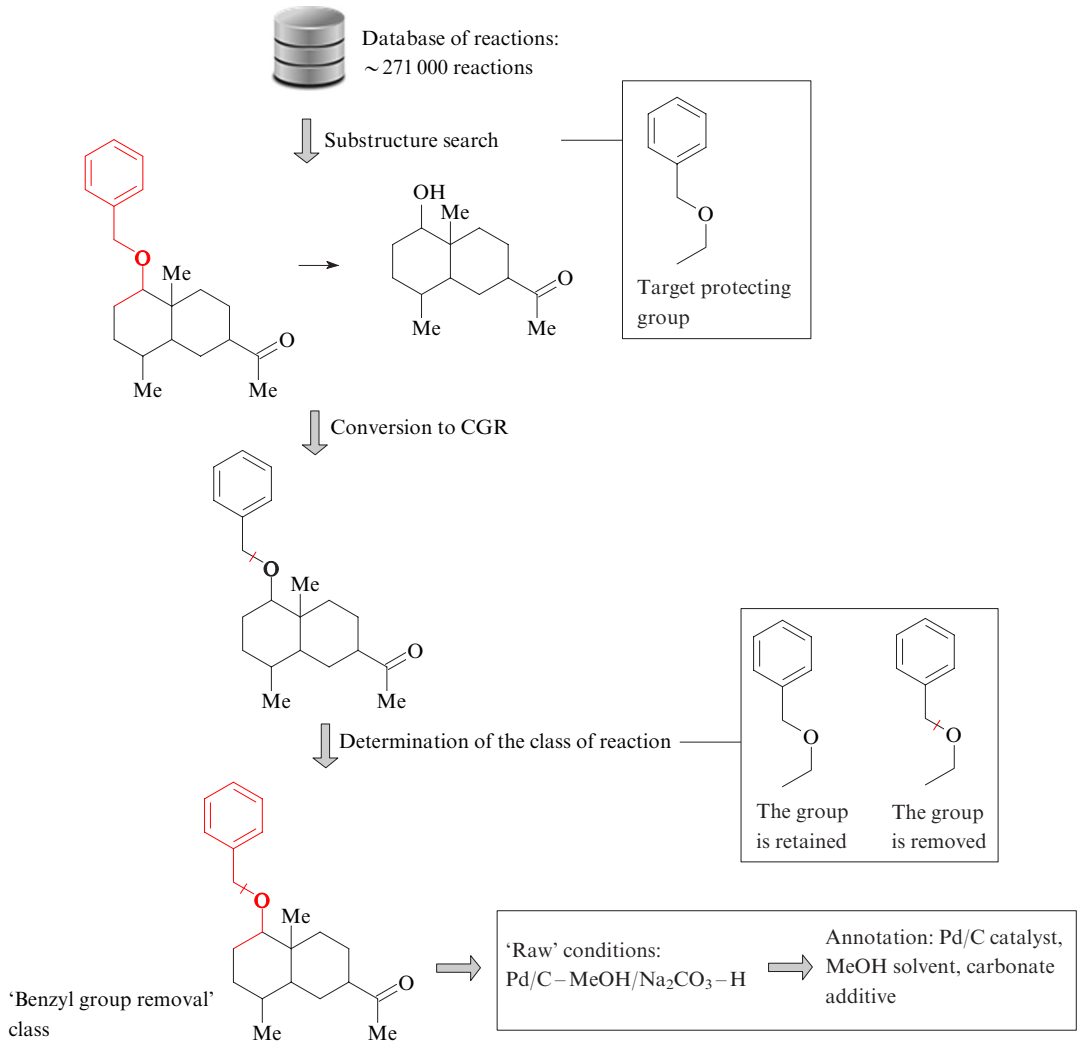


Figure 10. Data processing for collection of statistics on the reactivity of protecting groups.

Table 5. Analysis of deprotection of the phenol OH groups in catalytic hydrogenation reactions.[108]

Protecting group	Raney nickel		Pt, pH 2 – 4 (see [a])		Pd/C		Lindlar		Rh/C or Rh/Al_2O_3	
	A	*CR*	*A*	*CR*	*A*	*CR*	*A*	*CR*	*A*	*CR*
Methyl	L	0	L	0	L	0	L	0	L	0
Methoxymethyl	L	–	M	–	L	1	L	0	L	0
Methoxyethyl	L	–	M	–	L	4	L	–	L	–
Cyclohexyl	L	–	L	–	L	0	L	–	L	–
tert-Butyl	L	–	L	–	L	0	L	0	L	–
Benzyl	H	75	H	17	H	99	L	37	H	28
tert-Butyldimethylsilyl	L	–	H	–	L	1	L	0	L	0
Acetyl	L	0	M	0	L	1	L	0	L	0
Pivaloyl	L	–	L	–	L	6	L	–	L	–
Benzoyl	L	0	L	–	L	50	L	–	L	–
Methoxycarbonyl	L	–	H	–	L	0	L	–	L	–
Benzylcarbamate	H	–	H	–	H	100	L	–	L	–
Mesityl	R	0	L	–	L	8	L	–	L	–

Note. The following designations are used: *A* is reactivity, *CR* is the probability of deprotection under these conditions, % (*CR* = 100% means that this protecting group was removed under the given conditions in 100% of cases; *CR* = 0% means that the group was never removed). The reactivity labels H, M and L were taken from a monograph [227] and stand for high, medium and low reactivity, respectively; R means that no deprotection occurs under the indicated conditions but the protecting group is modified. The colour of the cell encodes the agreement (grey) and discrepancy (dark) between the analysis carried out by Lin *et al.*[108] and monograph data; [227] no colour corresponds to cases where no such reactions were found. The squares indicate the cases where the number of reactions for this group and conditions was <10. [a] It has been assumed [108] that Pt/C served as the catalyst and an acid was present as an additive.

known conditions. The catalysts used were classified according to types (Brønsted acids, Lewis acids, basic catalysts), and the class of reactions proceeding without a catalyst was selected. The solvents were also classified (hydrophobic, aprotic polar and protic) and a class of solvent-free reactions was designated. Thus, using machine learning, it was necessary to construct a model capable of predicting the classes of catalyst and solvent for the reaction. Both MOLMAP descriptors and fragment descriptors generated for CGR were used in combination with various machine learning techniques. This yielded classification models that evaluated the feasibility of the reaction in the indicated solvent with a particular type of catalyst.

Lin *et al.*[108] proposed an original approach to evaluation of the optimal conditions for removal of protecting groups using large bodies of data. First, the researchers developed an automatic procedure for processing the 'raw' information about the reactions from the Reaxys database, which included the following stages:

— standardization of reaction representation (structure standardization);

— atom-to-atom mapping;

— standardization of records about the reaction conditions (conversion of catalyst names to the standard form including identification of the solvents and low-molecular-mass reactants).

Using CGR, they identified the reactions in which the deprotection took place and in which the protecting groups were retained (Fig. 10). Hence it was possible to evaluate how frequently the protecting groups of a particular type were removed with a given catalyst. The results of statistical treatment for catalytic hydrogenation were compared with the results of 'manual' analysis of the literature data on reactivity published in a monograph [227] (see Table 5). It was found that only in 25% of cases, the results of automatic and manual analysis were in full agreement, while in 7% – 10% of cases, contradictions between them were found. In the other cases, the quantity of data was statistically insignificant, so that no substantiated conclusion could be drawn (<10 reactions).

Since a considerable drawback of reactivity tables presented in the mentioned monograph [227] is neglecting the structural environment of protecting groups, a prototype of an expert system free from this drawback was proposed.[108] The proposed strategy is based on the idea that similar reactions proceed under similar conditions. The expert system operation principle is schematically presented in Fig. 11. After structure standardization and conversion to CGR (stages 2 and 3), the program looks for the reaction resembling most closely the query in two folders prepared beforehand, which contain reactions with the protecting group of interest being either removed or retained with the given catalyst (stages 4 and 5). The similarity measure is quantitatively determined by the Tanimoto index, which is varied from 0 (for a pair of absolutely different reactions) to 1 (for a pair of identical reactions). It is taken that the reactivity of a protecting group in the query reaction (either removed or not) with a given catalyst corresponds to the type of reactions with the greatest similarity index (stage 6).

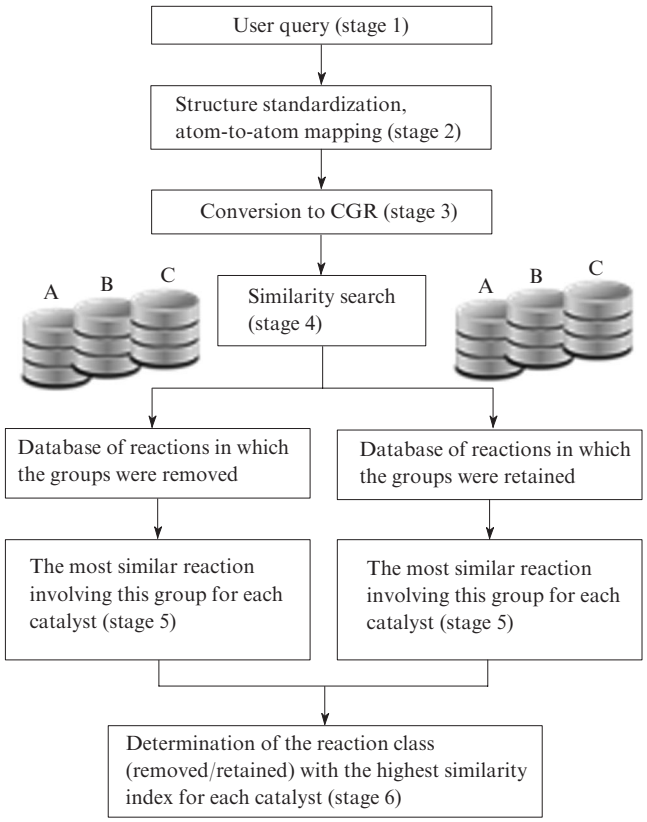


Figure 11. Prediction procedure for the conditions of reactions involving protecting groups. In stage 4, the reactions in the base are grouped depending on the catalyst used — A, B, C, *etc*.

If the behaviour of each protecting group with each catalyst is known, one can easily find a catalyst that is suitable for the selective deprotection.

VI.3. Metabolic transformations

A special type of chemical reactions is represented by metabolic transformations (biotransformations) of xenobiotics taking place in living organisms. Prediction is a very important task associated with new drug development and estimation of drug toxicity and pharmacokinetic properties. As in the case of organic synthesis planning, the first programs for prediction of biotransformations were designed as expert systems based on rules manually encoded by human experts and forming their knowledge base. The most well-known products of this type are MetabolExpert,[228] META[229,230] and METEOR.[231] The sets of rules, even formulated by skilled experts, proved to be insufficient for reliable prediction of metabolic pathways; therefore, now machine learning methods are used predominantly. The prediction is, in this case, accomplished by means of classification or regression models constructed using databases of metabolic transformations, in combination with the descriptor representation of chemical structures within the framework of the general QSAR approach.[201]

Studies dealing with QSAR modelling for predicting the quantitative characteristics of catalytic hydroxylation of *para*-substituted toluenes under the action of purified cytochrome P450 enzyme, which plays an important part in the metabolism of xenobiotics, appeared in the 1980s. The classic Hansch method was used — construction of a linear model relating the predicted rate constant, apparent dissociation constant of the ligand–enzyme complex and Gibbs energy to the substituent constants and hydrophobic constants of the molecule.[232] This model was subsequently improved by the use of physicochemical and quantum chemical descriptors.[233–235] In the later studies, various machine learning methods were used (multiple linear regression, *k* nearest neighbours, partial least squares, support vector machine, decision trees, multilayer artificial neural networks, Kohonen self-organizing maps, *etc*.) in combination with diverse molecular descriptors needed for the design of models for predicting the substrate binding in the cytochrome P450 active site,[236] enzyme inhibition[237] and induction,[238] and substrate specificity to enzyme subtypes.[239] The model designed by the support vector machine was implemented as a web-server predicting which form of cytochrome P450 would be involved in the metabolic transformation of a specified xenobiotic.[240]

In 2003, Poroikov's research group[241] developed the PASS-BioTransfo program for predicting the feasibility of various types of biotransformation for an arbitrary organic compound. The classification models used for this purpose were based on the metabolic transformation data stored in commercial databases, Metabolite (MDL Information System Inc.) and Metabolism (Accelrys), and the methodology that was implemented previously by the same research group[242,243] for predicting the range of biological activity for an organic compound with the PASS program. The chemical structures of xenobiotics were represented as fragment descriptors, which were called MNA (multilevel neighbourhoods of atoms) and describe the multilevel neighbourhood of atoms,[244] and a modified version of the 'naive Bayes' classifier was chosen as the machine learning method.[243] The accuracy of prediction of 1927 biotransformation classes by the model based on the Metabolite database was estimated as 88%, while the prediction accuracy for 178 biotransformation classes by means of the model using the Metabolism database was 85%. Later, this methodology was extended to the prediction of the probable metabolic transformation sites, which was done using special sorts of fragment descriptors called RMNA (reacting multilevel neighborhood of atom)[245] and LMNA (labelled multilevel neighborhood of atom).[246–248] This underlay the MetaTox web-application to predict the structures and toxicity of xenobiotic metabolites.[249] Modern research on computer-assisted prediction of the metabolic transformations of xenobiotics has been covered in detail in a review.[250]

Swamidass' research group developed models for predicting metabolic transformation sites using deep neural networks popular in the artificial intelligence field such as XenoSite P450 Metabolism (for metabolic activation by cytochrome P450),[251] Xenosite Epoxidation (for metabolic activation *via* epoxide formation),[252] Xenosite Quinone Formation (for metabolic activation *via* quinone formation),[253] Xenosite Reactivity (for predicting the toxicity associated with reactions of xenobiotic with nucleophilic centres in nucleic acid and protein macromolecules)[254,255] and XenoSite UGT [for metabolic activation by the uridine diphosphate glucuronosyltransferase (UGT) enzyme].[256]

The metabolic activation of xenobiotics may occur stereospecifically. This cannot be taken into account in the

above-mentioned models constructed by means of 2D descriptors. In order to take into account the spatial structures of molecules, 3D-QSAR models were constructed using the CoMFA [173] and CoMSIA [257] methods. In this case, the descriptor values are the values of molecular fields (electrostatic, steric, *etc.*) calculated at the nodes of an imaginary lattice formed around the training set molecules superimposed in space, with partial least squares being used as the machine learning technique to construct the regression model. The 1A2 (Ref. 258) and 2C19 (Ref. 259) models for cytochrome P450 inhibition were constructed in this way.

One more approach to metabolism prediction taking into account the geometric structures of molecules is pharmacophore analysis. This implies the search for a pharmacophore inherent in all molecules that exhibit a particular type of biological activity. Pharmacophore is defined as a set of generalized functional groups (pharmacophore features) and distances between them. The search for pharmacophores by comparing the chemical and geometric structures of active molecules can be treated as a special sort of machine learning. The pharmacophore approach has been utilized to predict the metabolic transformations of xenobiotics.[260 – 262]

VI.4. Synthetic accessibility of organic compounds

An important area of using artificial intelligence for reactivity analysis is the prediction of synthetic accessibility of organic compounds.[263] The significance of this task is due to the fact that the development of new drugs employs, more and more often, virtual screening of computer-generated chemical structures and *de novo* design of chemical compounds, that is, virtual assembly of a chemical compound within a receptor model. In both cases, it is highly important that the new chemical structures generated on a computer be accessible by laboratory synthesis. The use of only readily proceeding click reactions and libraries of easily accessible chemicals solves the problem only partly, like in the case of the DOGS system,[123] as it sharply restricts the range of eligible molecules.

Currently, there are two fundamentally different approaches to evaluation of the synthetic accessibility of organic compounds. One is based on the following idea: for understanding how accessible the specified compound is, it is necessary to carry out automatic retrosynthetic analysis and propose a pathway for compound synthesis from available reactants. This strategy was implemented in the CAESA commercial expert system (computer assisted estimation of synthetic accessibility [264, 265]). This approach provides a reliable estimate for the synthetic accessibility; however, it is highly time-consuming and, hence, it can be used only at final stages of virtual screening where the number of compounds to be considered is moderate.

One more approach that can be successfully combined with the first one is the use of complexity indices. For calculation of these indices, there is no need to perform retrosynthetic analysis in explicit form; they can be formed both purely heuristically and by machine learning, *i.e.*, by analysis of databases of accessible chemicals and reactions. This approach underlies the commercially available SYLVIA system.[266 – 268]

However, currently the synthetic accessibility problem cannot be considered to be solved, despite the presence of commercially available systems. Solution of this problem may be one of the goals in the development of new, more perfect approaches based on artificial intelligence methods.

VII. Modelling of catalytic activity and computer-aided design of new catalysts

Computer-aided design of new catalysts forms a separate group of studies on the prediction of quantitative characteristics of reactions.[269] In these studies, machine learning methods are used to model the relationship between the reactivity and the structure or composition of the catalyst (see reviews [270 – 273]). This problem is solved in different ways depending on whether catalysis is homogeneous, phase transfer or heterogeneous.

VII.1. Homogeneous catalysis

In the field of homogeneous catalysis, studies of this type address, most often, the quantitative prediction of the catalytic activity and enantioselectivity of organic reactions catalyzed by metal complexes and the use of constructed models for optimization of the structures of organic ligands incorporated in transition metal complexes applied as catalysts. The first studies along this line were based on the 3D-QSAR approach.[274 – 278] Lipkowitz and Pradhan [279] modelled the Diels – Alder reaction in the presence of copper complexes with chiral bisoxazoline and phosphine oxazoline ligands using electrostatic and steric descriptors, standard for the CoMFA method.[173]

In the modelling of acetylene polymerization catalyzed by zirconium-, titanium- and hafnium-based metallocenes, the LUMO and local softness fields calculated by the density functional theory were added to the two standard molecular fields mentioned above.[280, 281] Meanwhile, the 3D-QSAR models for the same reaction catalyzed by zirconocene were constructed using the standard fields.[282] From this, it follows that substituents at the ligands in metallocenes exert, first of all, electronic and steric effects on the reactivity, while quantum chemical descriptors are needed to take account of the differences between the nature of metals in the metallocene catalysts.

The studies of the last several years have mainly dealt with quantitative modelling of the enantioselectivity and reactivity for reactions catalyzed by homogeneous asymmetric metal complexes using the linear free energy relationships, considering variations of the ligands incorporated in the catalyst.[283 – 295] A bunch of studies along this line used the theoretically calculated descriptors, instead of empirical constants, for describing the effect of substituents.[292, 294, 295] In the studies cited, models were constructed within series of organometallic catalysts in which ligands had the same skeleton, but differed in the substituent structures, with the difference between free activation energies being the property to be modelled. Statistical analysis was performed using multiple linear regression, and non-linearity was introduced into the model by using squares and products of the initial descriptors. The constructed models served for optimization of the organometallic catalyst structures.

In 2013, Gonzalez-Diaz *et al.*[296] proposed a general approach to quantitative modelling of the reactivity based on combination of the perturbation theory (PT) with QSRR. The developed PT – QSRR method was applied to construct equations for quantitative prediction of the enantiomeric selectivity (enantiomeric excess) and yields for homogeneous asymmetric catalysis, depending on the reac-

tion conditions, the structure and composition of the catalytic system and the substrate structure. The PT–QSRR method was used to model intramolecular carbometallation,[296] palladium-catalyzed Heck–Heck cascade reaction,[297] Brønsted acid-catalyzed enantioselective α-amidoalkylation,[298] Parham cyclization[299] and other.

VII.2. Phase transfer catalysis

Several papers address the quantitative modelling of stereoselectivity in reactions catalyzed by chiral quaternary amines as phase transfer catalysts.[300–302] In all cases, the CoMFA method implementing the 3D-QSAR concept was used.[173] According to this method, a partial least squares algorithm is used as the machine learning technique to find a correlation between the activity to be modelled (in this case, the enantiomeric excess) and the electrostatic and steric fields of the molecule.

VII.3. Heterogeneous catalysis

Since the end of the 1980s, expert systems have found use in the development of materials for heterogeneous catalysis.[303] The knowledge bases in such systems contain detailed information about catalytic processes, including numerous rules and empirical trends encoded using special knowledge representation languages. The inference engine reproduces reasoning of a chemist guided by this knowledge. As a rule, each of these expert systems is directed towards one or several reactions widely used in chemical industry. The DECADE expert system was the first to be developed (in 1987);[304, 305] it is aimed at the choice of a metal catalyst for carbon monoxide hydrogenation (Fischer–Tropsch process). This was followed by INCAP[306–308] (for oxide-catalyzed oxidative dehydrogenation of ethylbenzene), INCAP-Muse[309] (for butene oxidative dehydrogenation catalyzed by a multicomponent mixture of metal oxides) and ARCADE as its upgraded version,[310] Hu System[311] (for the synthesis of alcohols by carbon monoxide hydrogenation catalyzed by metal oxides), IACES[312] (for the preparation of catalyst supports), EASYCAD[313] (for a group of reactions including CO hydrogenation, alkane dehydrogenation and methane oxidative coupling) and other.

Currently, machine learning methods are actively applied to heterogeneous catalysis.[270, 271, 273] The difficulty of modelling of heterogeneous catalysis is that the catalyst structure cannot be described as simply as it is done for homogeneous or phase transfer catalysts. Moreover, the exact structure of catalytic sites is usually unknown. Often these sites arise at crystal lattice defects with the participation of impurities and also when special metastable microphases are present on the solid surface. Nevertheless, a clear-cut dependence of the catalyst activity on the composition and characteristics of catalyst-forming atoms is often observed and can be studied using machine learning methods. The pioneering studies[314, 315] of 1991–1992, in which the neural network served for predicting the acidity of metal oxide-based catalysts were followed by publications in which this approach was applied to predict the distribution of the products of oxidative dehydrogenation of ethylbenzene catalyzed by promoted SnO_2-based catalysts[316, 317] and the yield of products of butane oxidation catalyzed by lanthanide oxides.[317]

A neural network was used by Huang *et al.*[318, 319] to optimize the composition of the multicomponent catalyst of methane oxidative coupling. If the Na, S, W, P, Zr and Mn mole fractions in the catalyst were taken as the input data, this method predicted the ethane conversion and selectivity to C_2.[319] A similar approach has been used[320] to estimate the product yield in the oxidative dehydrogenation of propane in the presence of a multicomponent catalyst containing B, Fe, Ga, La, Mg, Mn, Mo and V. This formed the basis of a general strategy for the search for the optimal catalyst compositions using neural network modelling in combination with the evolutionary optimization algorithms.[321] This combination also underlies the optimization methods for the composition of solid-state catalysts reported by Serra and co-workers.[322, 323]

The modern approach to the development of heterogeneous catalysts largely relies on experimental high-performance screening of their combinatorial libraries (see a review[324]). This screening can be accelerated to a large extent by using machine learning, in particular, artificial neural networks[325, 326] and the support vector machine.[327] The primary goal of this screening is to specify a combinatorial library of potential catalysts. Since there are infinitely many solid-state catalysts, rational design of a screening procedure requires that a most diverse subset of a finite (and minimum possible) size be chosen. For this purpose, using experimental design techniques, it is necessary to define a small set of descriptors in the space of which similar catalysts would have similar catalytic activity profiles (similarity principle). These descriptors would enable virtual screening of combinatorial libraries of solid-state catalysts and formation of focused libraries. These descriptors are needed for the efficient use of machine learning methods in the development of solid-state heterogeneous catalysts.

A fundamental distinction of the targeted design of solid-state catalysts from the design of low-molecular-mass compounds, for example, drugs, is that the catalytic activity is determined, to a large extent, by not only the catalyst composition, but also the preparation procedure (and, as a consequence, the presence of trace impurities, crystal lattice defects, morphology of the material, the presence of metastable microphases on the surface and a multitude of other factors) and application conditions. For example, catalysis by gold nanoparticles differs substantially from the catalysis by bulk gold, although the composition is the same. Therefore, the attributes describing the composition of the material are not always suitable as descriptors and often do not comply with the above-noted similarity principle.[328]

To solve this problem, Klanner *et al.*[328] proposed the following procedure for the formation of sets of descriptors for the solid-state catalysts:[329, 330]

— forming the primary set of diversified types of solid-state catalysts and high-performance screening of the catalysts against the test catalytic process under various conditions;

— screening gives rise to the maximum diversity of products;

— forming the response vector (*e.g.*, yields of various reaction products at various temperatures), which is used for clusterization of catalysts.

As a result of this procedure, each cluster would comprise catalysts with similar catalytic activity profiles. Then each of the catalysts is described by attributes of three major types, which characterize:

— material properties and preparation methods;

— catalyst composition;

— physicochemical properties of the major elements contained in the catalyst and their simple derivatives (*e.g.*, metal oxides in consideration of oxidation reactions).

Since there is a very large number (thousands) of these attributes, the most valuable ones (providing the best assignment of catalysts to the above-mentioned clusters by using machine learning methods, in particular, neural networks) are selected. The set of attributes generated as a result of this procedure forms the desired set of descriptors for solid-state catalysts. This strategy was successfully tested in the oxidation of propene: the final set of a moderate number of descriptors was formed using a library of 467 solid-state catalysts; initially 3179 different attributes were calculated for each of the catalysts.[329, 330] The set of descriptors thus formed was used for the design of new solid-state catalysts by means of models constructed *via* machine learning techniques.[330] In addition, the formation of a small set of the most valuable descriptors can be regarded as a method for extracting knowledge about the factors determining the efficiency of heterogeneous catalysis.

Corma *et al.*[331] proposed to use the X-ray diffraction spectra of a solid-state catalyst as descriptors. In this case, the initial spectra were subjected to primary processing by dimensionality reduction and clusterization methods (principal component method and Kohonen maps), and the models for predicting the catalytic activity were constructed using machine learning techniques such as logistic regression, multilayer neural networks and decision trees. The use of experimental spectra precludes the application of the constructed models for virtual screening; however, owing to the ease of obtaining the spectra, the models considerably accelerate the high-performance screening of the combinatorial libraries of catalysts in the case of long-term experimental measurements of the catalytic activity. Meanwhile, unlike the descriptors characterizing the catalyst composition and preparation conditions, X-ray diffraction spectra characterize directly the catalyst structure.

A special mention should be made of the studies devoted to modelling of the catalytic activities of zeolites, the crystal structure of which (unlike many other solid-state catalysts) can be determined in an X-ray diffraction experiment. Hence, a set of descriptors characterizing the structures of the zeolite crystal lattice has been formed.[332] An alternative set of physicochemical descriptors characterizing the zeolite structure was used to carry out the virtual screening of libraries of 300 000 zeolite structures in order to search for the best catalyst for CO_2 reduction.[333] The role of zeolite in this case consists in adsorption of H_2 and CO_2 molecules in the pores, thus creating conditions for the reduction reactions. The application of a special sort of Bayesian regularized artificial neural network as a machine learning technique made it possible to construct models that relate the adsorption properties and structures of zeolites. These models were used to design the catalytic system for CO_2 reduction.[333]

VIII. Conclusion

The vigorous development of chemoinformatics gave rise to novel approaches to computer synthesis planning methods, theoretical estimation of the kinetic and thermodynamic characteristics of reactions and prediction of the optimal conditions for a given reaction. Nevertheless, despite the fast progress, the artificial intelligence methods in chemistry have not yet reached the required level and have not yet become an everyday tool of a synthetic chemist. This may be attributed to both the insufficiently advanced methodology of modelling and analysis of chemical reactions and the problems of availability of reliable experimental data. As noted above, reactions are much more complex modelling objects than single compounds. The application of new machine learning techniques[63] with inclusion of parameters obtained in quantum chemical calculations can substantially increase the reliability of predictions of new reactions and their characteristics.

Currently, the CAS and Reaxys databases store information about more than 100×10^6 chemical reactions. Unfortunately, a greater part of this information cannot be used for modelling because of incomplete reaction equations and annotations of experimental conditions (catalyst, solvent, additives, pH, *etc.*) and the lack of information about the product yields or thermodynamic and kinetic reaction parameters. This may account for the moderate number of structure – reactivity models available from the literature — they are all based on rather small sets of data selected and annotated manually. In this connection, we would like to emphasize the importance of development of automatic curation procedures for reaction databases, including not only structure correction and standardization, balancing of atom numbers in reaction equations, correct annotation of reactions, but also atom-to-atom mapping. Our experience has shown that the condensed graph of reaction may prove very useful for the development of these technologies.[100, 156, 334, 335]

Obviously, the creation of open databases of chemical reactions would facilitate the use of approaches of chemoinformatics. These studies are already in progress along several avenues. The first one is manual annotation of the data on chemical reactions. For example, we collected a database of $\sim$15 000 values of kinetic parameters of various chemical reactions, which is planned for open-access publication. However, this approach is limited by the fact that the manual data mining is expensive and available only for large companies. To make it generally available, cooperation of the scientific community and development of a convenient-to-use platform for publication of the data on reactions are needed.

The information on chemical reactions can be retrieved directly from the literature using artificial intelligence methods for text mining[336] and image recognition. For example, computer-assisted text mining was used to retrieve information on more than 10^6 reactions from patents; and this information was used to solve various problems.[78, 101, 153] This approach is restricted by the fact that information is taken from secondary sources (papers, patents, collections and reference books), which gives rise to errors. Gasteiger[337] believes that the only way to solve this issue is to carry out quality checks before the information is deposited in a database, and for this purpose, information should be taken directly from laboratory notebooks. Thus, the most reliable way for collecting data on reactions is to use electronic laboratory notebooks.

One cannot help mentioning the nearly complete absence of open-access databases of chemical reactions,[263] with only small exceptions of websites for synthetic procedures, *e.g.*, OrgSyn (www.orgsyn.org), which cover a minor part of reactions, and the USPTO database (bitbucket.org/

dan2097/patent-reaction-extraction/downloads), which stores reactions taken from US patents. Almost all modern databases are commercial and, hence, accessible only by subscription. These databases include Reaxys (reaxys.elsevier.com), CASREACT (www.cas.org/content/reactions) and SPRESI (www.spresi.com) well-known to synthetic chemists. At some time, the same situation occurred with databases of chemical compounds, but chemoinformatics was revolutionized as open-access databases on chemical compounds appeared. First of all, biological activity data were opened for public access.[338] This markedly accelerated the development of new pharmaceutical drugs, and chemoinformatics became an inherent part of pharmaceutics as an important development tool.

The progress of computer analysis of chemical reactions and its application to 'big data' in chemistry would not only generate tools for predicting characteristics and conditions of any reaction, but also compete with humans in the ability to choose the optimal pathway for the synthesis of compounds.[57, 339] Combining these approaches would finally create an 'intellectual partner' — a device able to provide a chemist with information he is interested in and make decisions what compounds are to be synthesized and how this is to be done. The experimental equipment used in chemistry already has well-developed electronic interfaces, which allow computer control of processes.

It is quite reasonable that artificial intelligence and experiment design methods would be combined with instruments. Indeed, systems for automatic optimization of flow synthesis conditions have been reported.[340, 341] The use of artificial intelligence methods as control systems would give rise to a new type of such systems — an intellectual 'robochemist' [342] able to automate particular procedures of the search for compounds with desired properties, optimize and scale-up the synthesis, and even obtain the desired compound. Unlike the existing systems, the 'robochemist' would be able to make intelligent hypotheses based on 'big data' analysis, conduct an experiment, improve the model on the basis of experiment results and make new, more substantiated and accurate decisions. Although this goal seems to be science fiction and despite the skeptical attitude of some researchers,[340, 342] studies along this line are being carried out and tools of this type have already appeared.

Professor Schneider's group [343] developed the ALOE (automated lead optimization equipment) platform for optimization of the lead compound. The system is controlled by an algorithm based on the use of machine learning techniques. With a set of reactants and a set of synthetic rules at hand, this program predicts the possible products and, using a computer model, predicts their biological activity. Thus it is able to find the potentially most active compound and the method for its synthesis. Using ALOE, it is possible not only to synthesize the proposed compound by means of the microflow synthesis process, but also to experimentally assess its biological activity on a separate chip. These data are entered into a computer control system, which specifies the computational model. Thus, the system learns to give more accurate predictions with time and finds more and more active compounds without human intervention. Currently, ALOE implements click reactions, which do not require sophisticated selection of reaction conditions or separation of products. Nevertheless, this approach is a prototype of a machine that would be able to perform fully automatic structure optimization, synthesis and testing of biologically active compounds. An alternative approach for automatic optimization of the lead compound without synthesis planning elements (compounds are listed as combinatorial library elements), but with more advanced experimental equipment including the isolation of the target compound, has been reported.[344]

The time will come when the progress in the synthesis design approaches associated with prediction of the reaction conditions results in the development of systems for automatic synthesis that would be able to replace humans in the synthesis of chemicals.[342] This challenging goal is already being implemented with the Dial-a-Molecule Grand Challenge project sponsored by the UK Engineering and Physical Sciences Research Council (http://generic.wordpress.soton.ac.uk/dial-a-molecule/). The Dial-a-Molecule staff members plan that in the 20 – 40 years, they would make the preparation of a desired chemical compound as easy as to dial a phone number.[345, 346] Nevertheless, they admit the necessity of essential development of the existing tools. It becomes clear that flow reactors and microreactors, which are gaining popularity, are the major candidates for the future robotized reactor vessels. Owing to their design, they ensure higher rate of synthesis and easy controllability.

Thus, creation of 'synthesis robots' able to solve on their own the problems faced by chemists is a goal that has already passed from being a science fiction fantasy to practical implementation. This gives an optimistic view of the future of artificial intelligence methods in chemistry, which will crucially change the existing approaches to the preparation of new compounds and materials with specified properties.

This review was prepared with financial support of the Russian Science Foundation (Project No. 14-43-00024P).

References

1. I V Tetko, O Engkvist, U Koch, J-L Reymond, H Chen *Mol. Inf.* **35** 615 (2016)
2. D Che, M Safran, Z Peng, in *Database Systems for Advanced Applications: Proceedings of the 18th International Conference, DASFAA 2013, International Workshops: BDMA, SNSM, SeCoP, Wuhan, China, 2013* (Eds B Hong, X Meng, L Chen, W Winiwarter, W Song) (Berlin, Heidelberg: Springer, 2013) p. 1-15
3. A N Averkin, M G Gaaze-Rapoport, D A Pospelov *Tolkovyi Slovar' po Iskusstvennomu Intellektu* (Explanatory Dictionary of Artificial Intelligence) (Moscow: Radio i Svyaz', 1992)
4. E J Corey, W T Wipke *Science* **166** 178 (1969)
5. E J Corey *Pure Appl. Chem.* **14** 19 (1967)
6. E J Corey, W T Wipke, R D Cramer III, W J Howe *J. Am. Chem. Soc.* **94** 421 (1972)
7. W T Wipke, T M Dyott *J. Am. Chem. Soc.* **96** 4825 (1974)
8. W T Wipke, G I Ouchi, S Krishnan *Artif. Intel.* **11** 173 (1978)
9. N S Zefirov, E V Gordeeva *Russ. Chem. Rev.* **56** 1002 (1987)
10. N S Zefirov *Soros. Obrazovat. Zh.* (7) 52 (1997)
11. H Gelernter, A Sanders, D Larsen, K Agarwal, R Boivie, G Spritzer, J Searleman *Science* **197** 1041 (1977)
12. M Takahashi, I Dogane, M Yoshida, H Yamachika, T Takabatake, M Bersohn *J. Chem. Inf. Comput. Sci.* **30** 436 (1990)

13. I Dogane, T Takabatake, M Bersohn *Rec. Trav. Chim. Pays-Bas.* **111** 291 (1992)
14. W D Ihlenfeldt, J Gasteiger *Angew. Chem., Int. Ed.* **34** 2613 (1996)
15. R Doenges, B T Groebel, H Nickelsen, J Sander *J. Chem. Inf. Comput. Sci.* **25** 425 (1985)
16. J B Hendrickson, D L Grier, A G Toczko *J. Am. Chem. Soc.* **107** 5228 (1985)
17. J B Hendrickson, C A Parks *J. Chem. Inf. Comput. Sci.* **32** 209 (1992)
18. L Baumer, G Sala, G Sello *Anal. Chim. Acta* **235** 209 (1990)
19. L Baumer, I Compagnari, G Sala, G Sello *Rec. Trav. Chim. Pays-Bas.* **111** 297 (1992)
20. N S Zefirov, S S Trach *Zh. Org. Khimii* **17** 2465 (1981)
21. E V Gordeeva, D E Lushnikov, N S Zefirov *Tetrahedron Comput. Methodol.* **3** 445 (1990)
22. E V Gordeeva, D E Lushniknov, N S Zefirov *Tetrahedron* **48** 3789 (1992)
23. P Johnson, I Bernstein, J Crary, M Evans, T Wang, in *ACS Symposium Series. Expert System Application in Chemistry* (Washington, DC: American Chemical Society, 1989) Vol. 408, p. 102
24. S Hanessian, J Franco, G Gagnon, D Laramee, B Larouche *J. Chem. Inf. Comput. Sci.* **30** 413 (1990)
25. S Hanessian, M Botta, B Larouche, A Boyaroglu *J. Chem. Inf. Comput. Sci.* **32** 718 (1992)
26. S Hanessian *Pure Appl. Chem.* **65** 1189 (1993)
27. W T Wipke, D Rogers *J. Chem. Inf. Comput. Sci.* **24** 71 (1984)
28. P Azario, R Barone, M Chanon *J. Org. Chem.* **53** 720 (1988)
29. G Mehta, R Barone, P Azario, F Barberis, M Arbelot, M Chanon *Tetrahedron* **48** 8953 (1992)
30. G Mehta, R Barone, M Chanon *Eur. J. Org. Chem.* 1409 (1998)
31. K Funatsu, S-I Sasaki *Tetrahedron Comput. Methodol.* **1** 27 (1988)
32. K Satoh, K Funatsu *J. Chem. Inf. Comput. Sci.* **39** 316 (1999)
33. Q Huang, L-L Li, S-Y Yang *J. Chem. Inf. Model.* **51** 2768 (2011)
34. T D Salatin, W L Jorgensen *J. Org. Chem.* **45** 2043 (1980).
35. W L Jorgensen, E R Laird, A J Gushurst, J M Fleischer, S A Gothe, H E Helson, G D Paderes, S Sinclair *Pure Appl. Chem.* **62** 1921 (1990)
36. J Gasteiger, C Jochum, in *Organic Compounds* (Berlin, Heidelberg: Springer, 1978) p. 93
37. J Gasteiger, M G Hutchings, B Christoph, L Gann, C Hiller, P Löw, M Marsili, H Saller, K Yuki, in *Organic Synthesis. Reactions and Mechanisms* (Berlin, Heidelberg: Springer, 1987) p. 19
38. P Jauffret, C Laurenco, G Kaufmann *Tsi-Techn. Sci. Inf.* **5** 375 (1986)
39. H Satoh, K Funatsu *J. Chem. Inf. Comput. Sci.* **35** 34 (1995)
40. H Satoh, K Funatsu *J. Chem. Inf. Comput. Sci.* **36** 173 (1996)
41. I M Socorro, K Taylor, J M Goodman *Org. Lett.* **7** 3541 (2005)
42. I M Socorro, J M Goodman *J. Chem. Inf. Model.* **46** 606 (2006)
43. J Bauer, R Herges, E Fontain, I Ugi *Chimia* **39** 43 (1985)
44. I Ugi, J Bauer, K Bley, A Dengler, A Dietz, E Fontain, B Gruber, R Herges, M Knauer, K Reitsam, N Stein *Angew. Chem., Int. Ed.* **32** 201 (1993)
45. N S Zefirov, I I Baskin, V A Palyulin *J. Chem. Inf. Comput. Sci.* **34** 994 (1994)
46. N Zefirov, S Tratch, M Molchanova *Match-Commun. Mathem. Comput. Chem.* 253 (2002)
47. S Tratch, M Molchanova, N Zefirov *Match-Commun. Mathem. Comput. Chem.* 275 (2002)
48. M S Molchanova, S S Tratch, N S Zefirov *J. Phys. Org. Chem.* **16** 463 (2003)
49. R Barone, M Arbelot, A Baldy, M Chanon, R Gallo *Rev. Roum. Chim.* **36** 581 (1991)
50. E Fontain, K Reitsam *J. Chem. Inf. Comput. Sci.* **31** 96 (1991)
51. V V Shcherbukhin, N S Zefirov *J. Chem. Inf. Comput. Sci.* **35** 159 (1995)
52. E V Gordeeva, V V Shcherbukhin, N S Zefirov *Tetrahedron Comput. Methodol.* **3** 429 (1990)
53. R Barone, M Chanon, in *Computer Aids to Chemistry* (Eds G Vernin, M Chanon) (Chichester, UK: Ellis Horwood Limited, 1986) p. 19
54. R Barone, M Chanon, in *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes* (Ed. J Gasteiger) (Weinheim: Wiley-VCH, 2003) p. 1428
55. M H Todd *Chem. Soc. Rev.* **34** 247 (2005)
56. A Cook, A P Johnson, J Law, M Mirzazadeh, O Ravitz, A Simon *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2** 79 (2012)
57. S Szymkuć, E P Gajewska, T Klucznik, K Molga, P Dittwald, M Startek, M Bajczyk, B A Grzybowski *Angew. Chem., Int. Ed.* **55** 5904 (2016)
58. J Law, Z Zsoldos, A Simon, D Reid, Y Liu, S Y Khew, A P Johnson, S Major, R A Wade, H Y Ando *J. Chem. Inf. Model.* **49** 593 (2009)
59. O Ravitz *Drug Discov. Today: Technol.* **10** e443 (2013)
60. A Bøgevig, H-J Federsel, F Huerta, M G Hutchings, H Kraut, T Langer, P Löw, C Oppawsky, T Rein, H Saller *Org. Proc. Res. Develop.* **19** 357 (2015)
61. P Jackson *Introduction to Expert Systems* (Reading, MA: Addison-Wesley, 1986)
62. A Varnek, I I Baskin *Mol. Inf.* **30** 20 (2011)
63. A Varnek, I Baskin *J. Chem. Inf. Model.* **52** 1413 (2012)
64. *Chemoinformatics: a Textbook* (Eds J Gasteiger, T Engel) (Weinheim: Wiley-VCH, 2003)
65. *Handbook of Chemoinformatics: from Data to Knowledge in 4 Volumes* (Ed. J Gasteiger) (Weinheim: Wiley-VCH, 2003)
66. T I Madzhidov, I I Baskin, I S Antipin, A A Varnek *Vvedenie v Khemoinformatiku. Komp'yuternoe Predstavlenie Khimicheskikh Struktur. Uchebnoe Posobie* (Introduction to Chemoinformatics. Computer Representation of Chemical Structures. Tutorial (Kazan: Kazan Federal University, 2013)
67. T I Madzhidov, I I Baskin, A A Varnek *Vvedenie v Khemoinformatiku. Uchebnoe Posobie. Ch. 2. Khimicheskie Bazy Dannykh* (Introduction to Chemoinformatics. Tutorial. Pt. 2. Chemical Databases) (Kazan: Kazan Federal University, 2015)
68. I I Baskin, T I Madzhidov, A A Varnek *Vvedenie v Khemoinformatiku. Uchebnoe Posobie. Ch. 3. Modelirovanie 'Struktura – Svoistvo'* (Introduction to Chemoinformatics. Tutorial. Pt. 3. Structure – Property Relationship Modelling) (Kazan: Kazan Federal University, 2015)
69. I I Baskin, T I Madzhidov, A A Varnek *Vvedenie v Khemoinformatiku. Uchebnoe posobie. Ch. 4. Metody Mashinnogo Obucheniya.* (Introduction to Chemoinformatics. Tutorial. Pt. 4. Machine Learning Methods) (Kazan: Kazan Federal University, 2015)
70. J Dugundji, I Ugi, in *Computers in Chemistry* Vol. 39 (Berlin, Heidelberg: Springer, 1973) p. 19

71. J Gasteiger, C Jochum, in *Organic Compounds: Syntheses/Stereochemistry/Reactivity* (Berlin, Heidelberg: Springer, 1978) p. 93
72. J C J Bart, E Garagnni *Z. Naturforsch. B* **32** 455 (1977)
73. E Garagnni, J C J Bart *Z. Naturforsch. B* **32** 465 (1977)
74. J Bauer *Tetrahedron Comput. Methodol.* **2** 269 (1989)
75. R Fick, W-D Ihlenfeldt, J Gasteiger *Heterocycles* **2** 993 (1995)
76. C Jochum, J Gasteiger, I Ugi *Angew. Chem., Int. Ed.* **19** 495 (1980)
77. Q-N Hu, H Zhu, X Li, M Zhang, Z Deng, X Yang, Z Deng *PLoS ONE* **7** e52901 (2012)
78. N Schneider, D M Lowe, R A Sayle, G A Landrum *J. Chem. Inf. Model.* **55** 39 (2015)
79. Patent US 10/367550 (2003)
80. L Ridder, M Wagener *ChemMedChem* **3** 821 (2008)
81. J-L Faulon, M Misra, S Martin, K Sale, R Sapra *Bioinformatics* **24** 225 (2008)
82. D A R S Latino, Q-Y Zhang, J Aires-de-Sousa *Bioinformatics* **24** 2236 (2008)
83. D A R S Latino, J Aires-de-Sousa *Angew. Chem., Int. Ed.* **45** 2066 (2006)
84. G Marcou, J Aires de Sousa, D A R S Latino, A de Luca, D Horvath, V Rietsch, A Varnek *J. Chem. Inf. Model.* **55** 239 (2015)
85. J B Hendrickson *Angew. Chem., Int. Ed.* **13** 47 (1974)
86. J F Arens *Rec. Trav. Chim. Pays-Bas.* **98** 155 (1979)
87. J F Arens *Rec. Trav. Chim. Pays-Bas.* **98** 395 (1979)
88. J F Arens *Rec. Trav. Chim. Pays-Bas.* **98** 471 (1979)
89. N S Zefirov, S S Tratch *Chem. Scr.* **15** 4 (1980)
90. N S Zefirov, S S Tratch *Match (Serbia)* (3) 263 (1977)
91. N S Zefirov *Acc. Chem. Res.* **20** 237 (1987)
92. N S Zefirov, S S Tratch *Anal. Chim. Acta* **235** 115 (1990)
93. S S Tratch, N S Zefirov *J. Chem. Inf. Comput. Sci.* **38** 331 (1998)
94. G E Vladutz, E A Geivandov *Avtomatizirovannye Informatsionnye Sistemy dlya Khimii* (Automated Information Systems for Chemistry) (Moscow: Nauka, 1984)
95. G Vladutz, in *Modern Approaches to Chemical Reaction Searching* (Ed. P Willett) (London: Gower, 1986) p. 202
96. Yu K Kiho, in *Reaktsionnaya Sposobnost' Organicheskikh Soedinenii* (Reactivity of Organic Compounds) Vol. 8 (Tartu: Tartu State Universiy, 1971)
97. S Fujita *J. Chem. Inf. Comput. Sci.* **26** 205 (1986)
98. P Jauffret, T Hanser, C Tonnelier, G Kaufmann *Tetrahedron Comput. Methodol.* **3** 323 (1990)
99. P Jauffret, C Tonnelier, T Hanser, G Kaufmann, R Wolff *Tetrahedron Comput. Methodol.* **3** 335 (1990)
100. A Varnek, D Fourches, F Hoonakker, V Solov'ev *J. Comput.-Aided Mol. Des.* **19** 693 (2005)
101. D M Lowe, PhD Thesis, University of Cambridge, Cambridge, 2012
102. M Sitzmann, W-D Ihlenfeldt, M C Nicklaus *J. Comput.-Aided Mol. Des.* **24** 521 (2010)
103. A J Williams, S Ekins, V Tkachenko *Drug Disc. Today* **17** 685 (2012)
104. D Fourches, E Muratov, A Tropsha *J. Chem. Inf. Model.* **50** 1189 (2010)
105. D Fourches, E Muratov, A Tropsha *J. Chem. Inf. Model.* **56** 1243 (2016)
106. D Fourches, E Muratov, A Tropsha *Nat. Chem. Biol.* **11** 535 (2015)
107. K Karapetyan, C Batchelor, D Sharpe, V Tkachenko, A J Williams *J. Cheminf.* **7** 30 (2015)
108. A I Lin, T I Madzhidov, O Klimchuk, R I Nugmanov, I S Antipin, A Varnek *J. Chem. Inf. Mod.* **56** 2140 (2016)
109. N Schneider, N Stiefl, G A Landrum *J. Chem. Inf. Model.* **56** 2336 (2016)
110. L Chen, J G Nourse, B D Christie, B A Leland, D L Grier *J. Chem. Inf. Comput. Sci.* **42** 1296 (2002)
111. H Kraut, J Eiblmaier, G Grethe, P Löw, H Matuszczyk, H Saller *J. Chem. Inf. Mod.* **53** 2884 (2013)
112. F Hoonakker, N Lachiche, A Varnek, A Wagner *Int. J. Artific. Intelli Tools* **20** 253 (2011)
113. W L Chen, D Z Chen, K T Taylor *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **3** 560 (2013)
114. M Mann, F Nahar, N Schnorr, R Backofen, P F Stadler, C Flamm *Algorithms Mol. Biol.* **9** 23 (2014)
115. D Fooshee, A Andronico, P Baldi *J. Chem. Inf. Model.* **53** 2812 (2013)
116. T I Madzhidov, R I Nugmanov, T R Gimadiev, A I Lin, I S Antipin *Butlerovskie soobshcheniya* **44** 170 (2015)
117. C Muller, G Marcou, D Horvath, J Aires-de-Sousa, A Varnek *J. Chem. Inf. Mod.* **52** 3116 (2012)
118. C Cortes, V Vapnik *Machine Learning* **20** 273 (1995)
119. T H Vines, A Y K Albert, R L Andrew, F Debarre, D G Bock, M T Franklin, K J Gilbert, J-S Moore, S Renaut, D J Rennison *Curr. Biol.* **24** 94 (2014)
120. M Rubacha, A K Rattan, S C Hosselet *J. Association Lab. Automat.* **16** 90 (2011)
121. J Giles *Nature (London)* **481** 430 (2012)
122. P Judson *Knowledge-Based Expert Systems in Chemistry: not Counting on Computers* (London: Royal Society of Chemistry, 2009)
123. M Hartenfeller, H Zettl, M Walter, M Rupp, F Reisen, E Proschak, S Weggen, H Stark, G Schneider *PLoS Comput., Biol.* **8** e1002380 (2012)
124. M H Segler, M P Waller *Chem. – Eur. J.* **23** 6118 (2017)
125. J Gasteiger, J Zupan *Angew. Chem., Int. Ed.* **105** 503 (1993)
126. N M Halberstam, I I Baskin, V A Palyulin, N S Zefirov *Russ. Chem. Rev.* **72** 629 (2003)
127. I I Baskin, V A Palyulin, N S Zefirov *Ross. Khim. Zh.* **50** (2) 86 (2006)
128. I I Baskin, V A Palyulin, N S Zefirov, in *Artificial Neural Networks: Methods and Applications* (Ed. D J Livingstone) (Totowa, NJ: Humana Press, 2009) p. 133
129. I I Baskin, D Winkler, I V Tetko *Exp. Opin. Drug Disc.* **11** 785 (2016)
130. H Gelernter, J R Rose, C Chen *J. Chem. Inf. Comput. Sci.* **30** 492 (1990)
131. J R Rose, J Gasteiger *J. Chem. Inf. Comput. Sci.* **34** 74 (1994)
132. L Chen, J Gasteiger, J R Rose *J. Org. Chem.* **60** 8002 (1995)
133. P Jauffret, C Tonnelier, G Kaufmann *New J. Chem.* **14** 945 (1990)
134. C Tonnelier, P Jauffret, T Hanser, G Kaufmann *Tetrahedron Comput. Methodol.* **3** 351 (1990)
135. P Hamm, P Jauffret, G Kaufmann *Rec. Trav. Chim. Pays-Bas.* **111** 317 (1992)
136. J Gasteiger, H Saller, P Llw *Anal. Chim. Acta* **191** 111 (1986)
137. V Simon, J Gasteiger, J Zupan *J. Am. Chem. Soc.* **115** 9148 (1993)
138. G V S M Carrera, S Gupta, J Aires-de-Sousa *J. Comput.-Aided Mol. Des.* **23** 419 (2009)
139. D W Elrod, G M Maggiora, R G Trenary *J. Chem. Inf. Comput. Sci.* **30** 477 (1990)
140. D W Elrod, G M Maggiora, R G Trenary *Tetrahedron Comput. Methodol.* **3** 163 (1990)

141. V Kvasnicka, J Pospichal *J. Mol. Struct. (THEOCHEM)* **235** 227 (1991)
142. M A Kayala, C-A Azencott, J H Chen, P Baldi *J. Chem. Inf. Model.* **51** 2209 (2011)
143. M A Kayala, P Baldi *J. Chem. Inf. Model.* **52** 2526 (2012)
144. P Sadowski, D Fooshee, N Subrahmanya, P Baldi *J. Chem. Inf. Model.* **56** 2125 (2016)
145. J N Wei, D Duvenaud, A Aspuru-Guzik *ACS Centr. Sci.* **2** 725 (2016)
146. D K Duvenaud, D Maclaurin, J Aguilera-Iparraguirre, R Gomez-Bombarell, T Hirzel, A Aspuru-Guzik, R P Adams *Paper Presented at the Advances in Neural Information Processing Systems* arXiv: 1509.09292v2[cs.LG]3Nov 2015
147. C W Coley, R Barzilay, T S Jaakkola, W H Green, K F Jensen *ACS Centr. Sci.* **3** 434 (2017)
148. W Theilheimer *Theilheimer's Synthetic Methods of Organic Chemistry* (Ed. G Tozer-Hotchkiss) (Basel: S.Karger AG, 2015) p. 1946
149. L Chen, in *Handbook of Chemoinformatics: from Data to Knowledge* Vol. 1 (Ed. J Gasteiger) (Weinheim: Wiley-VCH, 2003) p. 348
150. J B Hendrickson, L Chen, in *Encyclopedia of Computational Chemistry* (Wiley, 2002) p. 320
151. J S Carey, D Laffan, C Thomson, M T Williams *Org. Biomol. Chem.* **4** 2337 (2006)
152. S D Roughley, A M Jordan *J. Med. Chem.* **54** 3451 (2011)
153. N Schneider, D M Lowe, R A Sayle, M A Tarselli, G A Landrum *J. Med. Chem.* **59** 4385 (2016)
154. S S Tratch, N S Zefirov *J. Chem. Inf. Comput. Sci.* **38** 349 (1998)
155. H L Morgan *J. Chem. Doc.* **5** 107 (1965)
156. A de Luca, D Horvath, G Marcou, V Solov'ev, A Varnek *J. Chem. Inf. Model.* **52** 2325 (2012)
157. D Horvath, C Jeandenans *J. Chem. Inf. Comput. Sci.* **43** 691 (2003).
158. T I Madzhidov, I I Baskin, I I Antipin, A Varnek *Russia Chemistry Week Science Innovation (21-26 October 2014) Event 'Cheminformatics: Supporting Innovation through Open Data'* (Royal Society of Chemistry, 2014)
159. C D Christ, M Zentgraf, J M Kriegl *J. Chem. Inf. Model.* **52** 1745 (2012)
160. G Sello, M Termini *Tetrahedron* **53** 14085 (1997)
161. G Sello *Tetrahedron* **54** 5731 (1998)
162. N M O'Boyle, G L Holliday, D E Almonacid, J B O Mitchell *J. Mol. Biol.* **368** 1484 (2007)
163. Patent US 7991730 (2007)
164. H A Gaspar, I I Baskin, A Varnek, in *Frontiers in Molecular Design and Chemical Information Science — Herman Skolnik Award Symposium 2015: Juergen Bajorath* Vol. 1222 (American Chemical Society, 2016) p. 243
165. T Kohonen *Proc. IEEE* **78** 1464 (1990)
166. C M Bishop, M Svensén, C K I Williams *Neural Comput.* **10** 215 (1998)
167. C M Bishop, M Svensén, C K I Williams *Neurocomputing* **21** 203 (1998)
168. T Kohonen *Self-Organizing Maps* (Berlin, Heidelberg: Springer, 2001)
169. F Hoonakker, Doctoral Thesis in Chimie, Louis Pasteur University of Strasbourg, Strasbourg, 2008
170. L Chen, J Gasteiger *J. Am. Chem. Soc.* **119** 4033 (1997)
171. H Satoh, S Itono, K Funatsu, K Takano, T Nakata *J. Chem. Inf. Comput. Sci.* **39** 671 (1999)
172. S Hiroko, F Kimito, T Keiko, N Tadashi *Bull. Chem. Soc. Jpn.* **73** 1955 (2000)
173. R D Cramer, D E Patterson, J D Bunce *J. Am. Chem. Soc.* **110** 5959 (1988)
174. K H Kim, in *Molecular Similarity in Drug Design* (Ed. P M Dean) (Netherlands: Springer, 1995) p. 291
175. S Gupta, S Matthew, P M Abreu, J Aires-de-Sousa *Bioorg. Med. Chem.* **14** 1199 (2006)
176. Q-Y Zhang, J Aires-de-Sousa *J. Chem. Inf. Model.* **45** 1775 (2005)
177. L Breiman *Mach. Learn.* **45** 5 (2001)
178. D A R S Latino, J Aires-de-Sousa *J. Chem. Inf. Model.* **49** 1839 (2009)
179. D A R S Latino, J Aires-de-Sousa *Mol. Inf.* **31** 135 (2012)
180. H A Gaspar, P Sidorov, D Horvath, I I Baskin, G Marcou, A Varnek, in *Frontiers in Molecular Design and Chemical Information Science — Herman Skolnik Award Symposium 2015: Juergen Bajorath* Vol. 1222 (American Chemical Society, 2016) p. 211
181. D M Maniyar, I T Nabney, B S Williams, A Sewing *J. Chem. Inf. Model.* **46** 1806 (2006)
182. J R Owen, I T Nabney, J L Medina-Franco, F Lopez-Vallejo *J. Chem. Inf. Model.* **51** 1552 (2011)
183. N Kireeva, I I Baskin, H A Gaspar, D Horvath, G Marcou, A Varnek *Mol. Inf.* **31** 301 (2012)
184. H A Gaspar, I I Baskin, G Marcou, D Horvath, A Varnek *J. Chem. Inf. Model.* **55** 84 (2015)
185. H A Gaspar, I I Baskin, G Marcou, D Horvath, A Varnek *J. Chem. Inf. Model.* **55** 2403 (2015)
186. H A Gaspar, G Marcou, D Horvath, A Arault, S Lozano, P Vayer, A Varnek *J. Chem. Inf. Model.* **53** 3318 (2013)
187. H A Gaspar, I I Baskin, G Marcou, D Horvath, A Varnek *Mol. Inf.* **34** 348 (2015)
188. P O Sidorov, A I Lin, T I Madzhidov, V A Varnek, in *Symposium 'From Empirical to Predictive Chemistry', Satellite Meeting to the XXth Mendeleev Congress on General and Applied Chemistry, Yekaterinburg, 2016* Vol. 5, p. 54
189. A Varnek, D Fourches, D Horvath, O Klimchuk, C Gaudin, P Vayer, V Solov'ev, F Hoonakker, I V Tetko, G Marcou *Curr. Comput.–Aided Drug Des.* **4** 191 (2008)
190. L P Hammett *Physical Organic Chemistry Reaction Rates, Equilibria, and Mechanisms* (New York, London: McGraw-Hill, 1940)
191. R W Taft *J. Am. Chem. Soc.* **74** 3120 (1952)
192. V A Palm *Russ. Chem. Rev.* **30** 471 (1961)
193. V A Palm *Osnovy Kolichestvennoi Teorii Organicheskikh Reaktsii* (Fundamentals of the Quantitative Theory of Organic Reactions) (2nd Edn) (Leningrad: Khimiya, 1977)
194. B K Mishra, A Sharma, G Behera *Int. J. Chem. Kinet.* **21** 439 (1989)
195. R Atkinson *Int. J. Chem. Kinet.* **19** 799 (1987)
196. A Klamt *Chemosphere* **26** 1273 (1993)
197. R Hiob, M Karelson *J. Chem. Inf. Comput. Sci.* **40** 1062 (2000)
198. F Ignatz-Hoover, R Petrukhin, M Karelson, A R Katritzky *J. Chem. Inf. Comput. Sci.* **41** 295 (2001)
199. M Karelson, U Maran, Y L Wang, A R Katritzky *Collect. Czech. Chem. Commun.* **64** 1551 (1999)
200. J Gasteiger, T Engel *Chemoinformatics: Textbook* (Weinheim: Wiley-VCH, 2003)

201. A Cherkasov, E N Muratov, D Fourches, A Varnek, I I Baskin, M Cronin, J Dearden, P Gramatica, Y C Martin, R Todeschini, V Consonni, V E Kuz'min, R Cramer, R Benigni, C Yang, J Rathman, L Terfloth, J Gasteiger, A Richard, A Tropsha *J. Med. Chem.* **57** 4977 (2015)
202. U A Chaudry, P L A Popelier *J. Phys. Chem. A* **107** 4578 (2003)
203. H Zhang, X Qu, H Ando *J. Mol. Struct. (THEOCHEM)* **725** 31 (2005)
204. P Gramatica, V Consonni, R Todeschini *Chemosphere* **38** 1371 (1999)
205. W M Meylan, P H Howard *Environ. Toxicol. Chem.* **22** 1724 (2003)
206. P Gramatica, P Pilutti, E Papa *J. Chem. Inf. Comput. Sci.* **44** 1794 (2004)
207. M H Fatemi *Anal. Chim. Acta* **556** 355 (2006)
208. Y Ren, H Liu, X Yao, M Liu *Anal. Chim. Acta* **589** 150 (2007)
209. S Gupta, N Basant, D Mohan, K P Singh *Environ. Sci. Pollut. Res.* **23** 14034 (2016)
210. S Gupta, N Basant *RSC Adv.* **6** 108448 (2016)
211. S Gupta, N Basant, D Mohan, K P Singh *SAR QSAR Environ. Res.* **27** 539 (2016)
212. N M Halberstam, I I Baskin, V A Palyulin, N S Zefirov *Mendeleev Commun.* 185 (2002)
213. I I Baskin, N M Halberstam, N V Artemenko, V A Palyulin, N S Zefirov, in *EuroQSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions* (Ed. M Ford) (Blackwell, 2003) p. 260
214. N M Halberstam, I I Baskin, V A Palyulin, N S Zefirov *Dokl. Chem.* **384** 140 (2002)
215. I I Baskin, A O Ait, N M Halberstam, V A Palyulin, N S Zefirov *SAR QSAR Environ. Res.* **13** 35 (2002)
216. N I Zhokhova, I I Baskin, V A Palyulin, A N Zefirov, N S Zefirov *Dokl. Chem.* **417** 282 (2007)
217. A A Kravtsov, P V Karpov, I I Baskin, V A Palyulin, N S Zefirov *Dokl. Chem.* **440** 299 (2011)
218. A A Kravtsov, P V Karpov, I I Baskin, V A Palyulin, N S Zefirov *Dokl. Chem.* **414** 128 (2007)
219. A A Kravtsov, P V Karpov, I I Baskin, V A Palyulin, N S Zefirov *Dokl. Chem.* **441** 314 (2011)
220. B Hemmateenejad, M Sanchooli, A Mehdipour *J. Phys. Org. Chem.* **22** 613 (2009)
221. T I Madzhidov, P G Polishchuk, R I Nugmanov, A V Bodrov, A I Lin, I I Baskin, A A Varnek, I S Antipin *Russ. J. Org. Chem.* **50** 459 (2014)
222. L Breiman *Mach. Learn.* **45** 5 (2001)
223. R I Nugmanov, T I Madzhidov, G R Khaliullina, I I Baskin, I S Antipin, A A Varnek *J. Struct. Chem.* **55** 1026 (2014)
224. T I Madzhidov, A V Bodrov, T R Gimadiev, R I Nugmanov, I S Antipin, A A Varnek *J. Struct. Chem.* **56** 1227 (2015)
225. T R Gimadiev, R I Nugmanov, T I Madzhidov, P G Polishuk, A S Petrovsky, I I Baskin, I S Antipin, A Varnek, in *Proceedings of the Second Kazan Summer School on Chemoinormatics, Kazan, 2015* p. 34
226. H Struebing, Z Ganase, P G Karamertzanis, E Siougkrou, P Haycock, P M Piccione, A Armstrong, A Galindo, C S Adjiman *Nat. Chem.* **5** 952 (2013)
227. P G M Wuts *Greene's Protective Groups in Organic Synthesis.* (Weinheim: Wiley, 2014)
228. F Darvas *J. Mol. Graph. Model.* **6** 80 (1988)
229. G Klopman, M Dimayuga, J Talafous *J. Chem. Inf. Comput. Sci.* **34** 1320 (1994)
230. J Talafous, L M Sayre, J J Mieyal, G Klopman *J. Chem. Inf. Comput. Sci.* **34** 1326 (1994)
231. N Greene, P N Judson, J J Langowski, C A Marchant *SAR QSAR Environ. Res.* **10** 299 (1999)
232. R E whitetypo, M-B McCarthy *Arch. Biochem. Biophys.* **246** 19 (1986)
233. D F V Lewis, C Ioannides, D V Parke *Biochem. Pharmacol.* **50** 619 (1995)
234. D F V Lewis *J. Chem. Technol. Biotechnol.* **76** 237 (2001)
235. D F V Lewis, M Dickins *J. Enzyme Inhibit.* **16** 321 (2001)
236. A Long, J D Walker *Environ. Toxicol. Chem.* **22** 1894 (2003)
237. P Vasanthanathan, O Taboureau, C Oostenbrink, N P E Vermeulen, L Olsen, F S Jorgensen *Drug Metabol. Disp.* **37** 658 (2009)
238. D Schuster, T M Steindl, T Langer *Curr. Top. Med. Chem.* **6** 1627 (2006)
239. L Terfloth, B Bienfait, J Gasteiger *J. Chem. Inf. Model.* **47** 1688 (2007)
240. N K Mishra, S Agarwal, G P Raghava *BMC Pharmacol.* **10** 8 (2010)
241. Y Borodina, A Sadym, D Filimonov, V Blinova, A Dmitriev, V Poroikov *J. Chem. Inf. Comput. Sci.* **43** 1636 (2003)
242. Y V Borodina, D A Filimonov, V V Poroikov *Pharm. Chem. J.* **30** 760 (1996)
243. V V Poroikov, D A Filimonov, Y V Borodina, A A Lagunin, A Kos *J. Chem. Inf. Comput. Sci.* **40** 1349 (2000)
244. D Filimonov, V Poroikov, Y Borodina, T Gloriozova *J. Chem. Inf. Comput. Sci.* **39** 666 (1999)
245. Y Borodina, A Rudik, D Filimonov, N Kharchevnikova, A Dmitriev, V Blinova, V Poroikov *J. Chem. Inf. Comput. Sci.* **44** 1998 (2004)
246. A V Rudik, A V Dmitriev, A A Lagunin, D A Filimonov, V V Poroikov *J. Chem. Inf. Model.* **54** 498 (2014)
247. A Rudik, A Dmitriev, A Lagunin, D Filimonov, V Poroikov *Bioinformatics* **31** 2046 (2015)
248. A V Rudik, A V Dmitriev, A A Lagunin, D A Filimonov, V V Poroikov *J. Cheminf.* **8** 68 (2016)
249. A V Rudik, V M Bezhentsev, A V Dmitriev, D S Druzhilovskiy, A A Lagunin, D A Filimonov, V V Poroikov *J. Chem. Inf. Mod.* **57** 638 (2017)
250. V M Bezhentsev, O A Tarasova, A V Dmitriev, A V Rudik, A A Lagunin, D A Filimonov, V V Poroikov *Russ. Chem. Rev.* **85** 854 (2016)
251. J Zaretzki, M Matlock, S J Swamidass *J. Chem. Inf. Model.* **53** 3373 (2013)
252. T B Hughes, G P Miller, S J Swamidass *ACS Centr. Sci.* **1** 168 (2015)
253. T B Hughes, S J Swamidass *Chem. Res. Toxicol.* **30** 642 (2017)
254. T B Hughes, G P Miller, S J Swamidass *Chem. Res. Toxicol.* **28** 797 (2015)
255. T B Hughes, N L Dang, G P Miller, S J Swamidass *ACS Centr. Sci.* **2** 529 (2016)
256. N L Dang, T B Hughes, V Krishnamurthy, S J Swamidass *Bioinformatics* **32** 3183 (2016)
257. G Klebe, U Abraham, T Mietzner *J. Med. Chem.* **37** 4130 (1994)
258. L E Korhonen, M Rahnasto, N J Mahonen, C Wittekindt, A Poso, R O Juvonen, H Raunio *J. Med. Chem.* **48** 3808 (2005)

259. C W Locuson, H Suzuki, A E Rettie, J P Jones *J. Med. Chem.* **47** 6768 (2004)
260. B C Jones, G Hawksworth, V A Horne, A Newlands, J Morsman, M S Tute, D A Smith *Drug Metabol. Disp.* **24** 260 (1996)
261. D Dong, B J Wu, D Chow, M Hu *Drug Metabol. Rev.* **44** 192 (2012)
262. S L Mo, W F Liu, Y L Chen, H B Luo, L B Sun, X W Chen, Z W Zhou, K B Sneed, C G Li, Y M Du, J Liang, S F Zhou *Combinat. Chem. High Throughput Screen.* **15** 36 (2012)
263. W A Warr *Mol. Inf.* **33** 469 (2014)
264. V J Gillet, G Myatt, Z Zsoldos, A P Johnson *Persp. Drug Discov. Des.* **3** 34 (1995)
265. V J Gillet, W Newell, P Mata, G Myatt, S Sike, Z Zsoldos, A P Johnson *J. Chem. Inf. Comput. Sci.* **34** 207 (1994)
266. K Boda, T Seidel, J Gasteiger *J. Comput.–Aided Mol. Des.* **21** 311 (2007)
267. J Gasteiger *J. Comput.–Aided Mol. Des.* **21** 307 (2007)
268. P Bonnet *Eur. J. Med. Chem.* **54** 679 (2012)
269. C Poree, F Schoenebeck *Acc. Chem. Res.* **50** 605 (2017)
270. G Rothenberg *Catal. Today* **137** 2 (2008)
271. A G Maldonado, G Rothenberg *Chem. Eng. Prog.* **105** 26 (2009)
272. A G Maldonado, G Rothenberg *Chem. Soc. Rev.* **39** 1891 (2010)
273. E-J Ras, G Rothenberg *RSC Adv.* **4** 5963 (2014)
274. *3D QSAR in Drug Design* Vol. 1: Theory Methods and Applications (Three-Dimensional Quantitative Structure Activity Relationships) (Ed. H Kubinyi) (Dordrecht: Kluwer Academic, 2000)
275. *3D QSAR in Drug Design. Ligand-Protein Interactions and Molecular Similarity* (Eds H Kubinyi, G Folkers, Y C Martin) (Dordrecht: Kluwer Academic, 2002) Vol. 2
276. *3D QSAR in Drug Design. Ligand-Protein Interactions and Molecular Similarity* (Eds H Kubinyi, G Folkers, Y C Martin) (Dordrecht: Kluwer Academic, 2002) Vol. 3
277. M Akamatsu *Curr. Top. Med. Chem.* **2** 1381 (2002)
278. J Verma, V M Khedkar, E C Coutinho *Curr. Top. Med. Chem.* **10** 95 (2010)
279. K B Lipkowitz, M Pradhan *J. Org. Chem.* **68** 4648 (2003)
280. V Cruz, J Ramos, A Munoz-Escalona, P Lafuente, B Pena, J Martinez-Salazar *Polymer* **45** 2061 (2004)
281. V L Cruz, J Ramos, S Martinez, A Munoz-Escalona, J Martinez-Salazar *Organometallics* **24** 5095 (2005)
282. S Martinez, V L Cruz, J Ramos, J Martinez-Salazar *Organometallics* **31** 1673 (2012)
283. K C Harper, M S Sigman *Proc. Natl. Acad. Sci.* **108** 2179 (2011)
284. K C Harper, M S Sigman *Science* **333** 1875 (2011)
285. K C Harper, E N Bess, M S Sigman *Nat. Chem.* **4** 366 (2012)
286. K C Harper, M S Sigman *J. Org. Chem.* **78** 2813 (2013)
287. K C Harper, S C Vilardi, M S Sigman *J. Am. Chem. Soc.* **135** 2482 (2013)
288. E N Bess, A J Bischoff, M S Sigman *Proc. Natl. Acad. Sci.* **111** 14698 (2014)
289. A Milo, A J Neel, F D Toste, M S Sigman *Science* **347** 737 (2015)
290. C Zhang, C B Santiago, J M Crawford, M S Sigman *J. Am. Chem. Soc.* **137** 15668 (2015)
291. C Zhang, C B Santiago, L Kou, M S Sigman *J. Am. Chem. Soc.* **137** 7290 (2015)
292. C B Santiago, A Milo, M S Sigman *J. Am. Chem. Soc.* **138** 13424 (2016)
293. M S Sigman, K C Harper, E N Bess, A Milo *Acc. Chem. Res.* **49** 1292 (2016)
294. H Huang, H Zong, G Bian, H Yue, L Song *J. Org. Chem.* **79** 9455 (2014)
295. T Piou, F Romanov-Michailidis, M Romanova-Michaelides, K E Jackson, N Semakul, T D Taggart, B S Newell, C D Rithner, R S Paton, T Rovis *J. Am. Chem. Soc.* **139** 1296 (2017)
296. H Gonzalez-Diaz, S Arrasate, A Gomez-SanJuan, N Sotomayor, E Lete, L Besada-Porto, J M Ruso *Curr. Top. Med. Chem.* **13** 1713 (2013)
297. C Blazquez-Barbadillo, E Aranzamendi, E Coya, E Lete, N Sotomayor, H González-Dìaz *RSC Adv.* **6** 38602 (2016)
298. E Aranzamendi, S Arrasate, N Sotomayor, H González-Dìaz, E Lete *Chem. Open* **5** 540 (2016)
299. L Simon-Vidal, S Arrasate, N Sotomayor, E Lete, H González-Dìaz *Paper Presented at the MOL2NET 2016, International Conference on Multidisciplinary Sciences* (2nd Ed.) (Multidisciplinary Digital Publishing Institute, 2016)
300. J L Melville, B I Andrews, B Lygo, J D Hirst *Chem. Commun.* 1410 (2004)
301. J L Melville, K R J Lovelock, C Wilson, B Allbutt, E K Burke, B Lygo, J D Hirst *J. Chem. Inf. Model.* **45** 971 (2005)
302. S E Denmark, N D Gould, L M Wolf *J. Org. Chem.* **76** 4337 (2011)
303. Y Yoneda *Catal. Today* **23** 305 (1995)
304. R Banãres-Alcántara, A W Westerberg, E I Ko, M D Rychener *Comput. Chem. Eng.* **11** 265 (1987)
305. R Banãres-Alcántara, E I Ko, A W Westerberg, M D Rychener *Comput. Chem. Eng.* **12** 923 (1988)
306. T Hattori, S Kito, Y Murakami *Chem. Lett.* 1269 (1988)
307. T Hattori, H Niwa, A Satsuma, S Kito, Y Murakami *Appl. Catal.* **50** L11 (1989)
308. S Kito, T Hattori, Y Murakami *Appl. Catal.* **48** 107 (1989)
309. S Kito, T Hattori, Y Murakami *Chem. Eng. Sci.* **45** 2661 (1990)
310. T Hattori, S Kito *Catal. Today* **10** 213 (1991)
311. X D Hu, H C Foley, A B Stiles *Industr. Eng. Chem. Res.* **30** 1419 (1991)
312. Y H Sun, Y W Li *Chem. Eng. Sci.* **47** 2799 (1992)
313. E Körting, M Baerns *MRS Proc.* 454 (1996)
314. S Kito, T Hattori, Y Murakami *Anal. Sci.* **7** 761 (1991)
315. S Kito, T Hattori, Y Murakami *Industr. Eng. Chem. Res.* **31** 979 (1992)
316. S Kito, T Hattori, Y Murakami *Appl. Catal., A* **114** L173 (1994)
317. T Hattori, S Kito *Catal. Today* **23** 347 (1995)
318. K Huang, F-Q Chen, D-W Lü *Appl. Catal., A* **219** 61 (2001)
319. K Huang, X-L Zhan, F-Q Chen, D-W Lü *Chem. Eng. Sci.* **58** 81 (2003)
320. M Holeña, M Baerns *Catal. Today* **81** 485 (2003)
321. M Holeña, D Linke, U Rodemerck *Paper Presented at the Asia-Pacific Conference on Simulated Evolution and Learning*) (Berlin, Heidelberg: Springer, 2010)
322. J M Serra, A Corma, S Valero, E Argente, V Botti *Mol. Inf.* **26** 11 (2007)
323. A Corma, J Serra, P Serna, S Valero, E Argente, V Botti *J. Catal.* **229** 513 (2005)
324. D Farrusseng *Surface Sci. Rep.* **63** 487 (2008)
325. L Baumes, D Farrusseng, M Lengliz, C Mirodatos *Mol. Inf.* **23** 767 (2004)

326. A Corma, J M Serra, E Argente, V Botti, S Valero *ChemPhysChem 3* 939 (2002)
327. L Baumes, J Serra, P Serna, A Corma *J. Comb. Chem.* **8** 583 (2006)
328. C Klanner, D Farrusseng, L Baumes, C Mirodatos, F Schuth *QSAR Comb. Sci.* **22** 729 (2003)
329. C Klanner, D Farrusseng, L Baumes, M Lengliz, C Mirodatos, F Schuth *Angew. Chem., Int. Ed.* **43** 5347 (2004)
330. D Farrusseng, C Klanner, L Baumes, M Lengliz, C Mirodatos, F Schuth *QSAR Comb. Sci.* **24** 78 (2005)
331. A Corma, J Serra, P Serna, M Moliner *J. Catal.* **232** 335 (2005)
332. A Rajagopalan, C Suh, X Li, K Rajan *Appl. Catal., A* **254** 147 (2003)
333. A W Thornton, D A Winkler, M S Liu, M Haranczyk, D F Kennedy *RSC Adv.* **5** 44361 (2015)
334. F Hoonakker, N Lachiche, A Varnek, A Wagner, in *Trends in Applied Intelligent Systems. Proceedings. Ser. Lecture Noted in Artificial Intelligence* Vol. 6097 (Berlin, Heidelberg: Springer, 2010) p. 318
335. D Horvath, G Marcou, A Varnek, S Kayastha, A D de Leon, J Bajorath *J. Chem. Inf. Comput. Sci.* **56** 1631 (2016)
336. M Krallinger, O Rabal, A Lourenço, J Oyarzabal, A Valencia *Chem. Rev.* **117** 7673 (2017)
337. J Gasteiger *Mol. Inf.* **33** 454 (2014)
338. G Nicola, T Liu, M K Gilson *J. Med. Chem.* **55** 6987 (2012)
339. M K Yadav *New J. Chem.* **41** 1411 (2017)
340. B J Reizman, Y-M Wang, S L Buchwald, K F Jensen *React. Chem. Eng.* **1** 658 (2016)
341. V Sans, L Porwol, V Dragone, L Cronin *Chem. Sci.* **6** 1258 (2015)
342. M Peplow *Nature (London)* **512** 20 (2014)
343. T Rodrigues, P Schneider, G Schneider *Angew. Chem., Int. Ed.* **53** 5750 (2014)
344. B Desai, K Dixon, E Farrant, Q Feng, K R Gibson, W P van Hoorn, J Mills, T Morgan, D M Parry, M K Ramjee, C N Selway, G J Tarver, G Whitlock, A G Wright *J. Med. Chem.* **56** 3033 (2013)
345. K J Kilpin, R J Whitby *Chem. – Centr. J.* **9** 43 (2015)
346. J Kilpin, J M Goodman, A P Johnson, R J Whitby *Chem. – Centr. J.* **9** 49 (2015)