J Comput Aided Mol Des (2007) 21:33–52
DOI 10.1007/s10822-006-9097-4

ORIGINAL PAPER

Modeling chemical reactions for drug design

Johann Gasteiger

Received: 24 November 2006 / Accepted: 6 December 2006 / Published online: 25 January 2007

Abstract Chemical reactions are involved at many stages of the drug design process. This starts with the analysis of biochemical pathways that are controlled by enzymes that might be downregulated in certain diseases. In the lead discovery and lead optimization process compounds have to be synthesized in order to test them for their biological activity. And finally, the metabolism of a drug has to be established. A better understanding of chemical reactions could strongly help in making the drug design process more efficient. We have developed methods for quantifying the concepts an organic chemist is using in rationalizing reaction mechanisms. These methods allow a comprehensive modeling of chemical reactivity and thus are applicable to a wide variety of chemical reactions, from gas phase reactions to biochemical pathways. They are empirical in nature and therefore allow the rapid processing of large sets of structures and reactions. We will show here how methods have been developed for the prediction of acidity values and of the regioselectivity in organic reactions, for designing the synthesis of organic molecules and of combinatorial libraries, and for furthering our understanding of enzyme-catalyzed reactions and of the metabolism of drugs.

Keywords Physicochemical effects · Metabolism · Chemical reactivity · Synthesis design · Enzymatic reactions · Lead discovery and optimization

J. Gasteiger (✉)
Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, 91052 Erlangen, Germany
e-mail: gasteiger@chemie.uni-erlangen.de
URL: http://www2.chemie.uni-erlangen.de

Introduction

The understanding of the relationships between the chemical structure of a compound and its biological activity has made enormous progress in the last two decades. Two developments have been instrumental in initiating this progress: the production of huge amounts of experimental data and the development of novel methods for analyzing these data.

A variety of novel experimental techniques such as the polymerase chain reaction, gene sequencing, combinatorial chemistry, high-throughput screening, and progresses in X-ray structure analysis have provided large sets of data on genes, on new chemical compounds and on their associated biological activities. This access to new—and large—amounts of data has put pressure on developing methods for analyzing them helping to push the fields of bioinformatics and chemoinformatics into the lime-light.

At each step of the drug design process, bioinformatics and chemoinformatics methods have been involved to make this process more efficient [1, 2]. The analysis of gene sequences by bioinformatics methods is used in the identification of the target proteins that are instrumental in causing diseases. Lead discovery profits from chemoinformatics comprizing both structure-based and the ligand-based methods. Lead optimization benefits from the analysis of large sets of chemical compounds and their associated biological activities and from establishing quantitative structure-activity relationships (QSAR). The prediction of such properties as adsorption, distribution, metabolism and excretion (ADME), important in identifying drug metabolism and pharmacokinetic properties, is largely built on QSAR

methods. The same can be said from modeling the toxicity of chemical compounds.

Much as the field of finding the relationships between chemical structure and biological activity has been developed much less progress has been made in understanding chemical reactions. This deficiency is increasingly felt in the field of metabolism. However, chemical reactions do not only play a role in the metabolism of drugs but chemical reactions are involved in each step of the drug design and development process. Target identification often has to find out how genes regulate enzymes which, for their part, control enzyme catalyzed reactions. In lead discovery and lead optimization, chemical compounds have to be synthesized, both by individual or by parallel synthesis. Furthermore, these compounds are often stored for longer periods of time and thus a knowledge of their stability is important. An understanding of drug metabolism and pharmacokinetics crucially asks for knowledge about metabolic reactions. In many cases also a knowledge of pK_a values—a chemical reaction!—is needed. And, finally, many toxic modes of action are caused by chemical reactions of the compounds considered with nucleic acids or with proteins.

Clearly, many factors influence chemical reactions, the structure of the reactant, the catalyst, which is often an enzyme, reaction conditions such as solvent, pH value, and temperature. The influence of these factors on a reaction is often not yet well understood making the modeling and the prediction of the course and outcome of chemical reactions such a difficult task. Quite some inroads have been made by quantum mechanics to explicitly calculate transition states and thus model a specific chemical reaction by deductive learning methods. However, the many and the large variety of reactions that has to be analyzed still asks for another approach that allows one to process large numbers of diverse reactions. Only then can we answer the many problems faced in understanding chemical reactions at each stage of the drug design process. The only alternative to a purely theoretical, a deductive, approach is an inductive approach, an approach based on the analysis of the many data available about chemical reactions. In order that such an approach can be applied to the broad variety of problems dealing with chemical reactions in drug design, a global approach to dealing with chemical reactivity has to be developed. Only then can the models developed for individual reactions be generalized and thus utilized to gain knowledge on chemical reactions, knowledge that can also be used for new types of problems dealing with chemical reactions.

Methods

In our endeavor to develop a general approach to treating chemical reactivity and chemical reactions we let ourselves be guided by the concepts organic chemists have developed for the discussion of reaction mechanisms. These concepts include notions like partial atomic charges, or inductive, resonance, polarizability and steric effects. Chemists use these concepts mostly in a qualitative or, at best, semi-quantitative manner. It was our firm conviction that if we quantify these concepts we would be put in a position of making quantitative predictions on chemical reactivity. Many of the methods for the quantification of these physicochemical effects were developed 25–30 years ago when computer power was much less developed. Furthermore, we wanted to apply these methods to the calculation of interest to the organic chemist while, at that time ab initio methods could only be applied to molecules with a few atoms. Therefore, simple empirical procedures were designed that allowed rapid calculations of fairly large sets of organic molecules of interest to the chemist [3]. Even now, with the enormous progress in computer technology, these methods still have great value as they allow the treatment of large sets of molecules, comprising millions of structures, presently of interest in combinatorial chemistry and virtual screening.

As these methods have largely been published it suffices to give here the references. Partial atomic charges, q_i, are calculated by the partial equalization of orbital electronegativities (PEOE) method [4]. The PEOE method applies to σ-bonded systems. For the calculation of charges in π-bonded systems initially an approach that generates and weighs all resonance structures was developed [5]. This method has now been replaced by a modified Hückel molecular orbital method [6]. Residual electronegativities, χ_i, obtained in the PEOE method can be taken as a quantitative measure of the inductive effect [7]. The resonance effect in the ground state of molecules is embodied in the method for the calculation of partial charges, q_π, in π-systems [6]. As many reactions are heterolytic in nature, it was found highly useful to have a method that allowed an evaluation how these incipient charges generated by bond breaking are stabilized through delocalization measured by D^+ and D^-, respectively (Marusczyk J, unpublished results). The polarizability effect can be calculated by a damped additivity scheme providing values on the polarizability effect, α_d, exerted on an atom in a molecule [8]. The steric effect can be evaluated based on a simple scheme using atomic radii or local radial distribution functions.

Table 1 Physicochemical effects for which calculation procedures have been developed

Physicochemical effects	Symbol	Reference
Partial charges	q_σ	4
	q_π	5,6
Electronegativities	χ_δ	7
	χ_π	5
	χ_{LP}	5
Effective polarizability	α_d	8
Hyperconjugation	N_{hyp}	9
Delocalization	D^+	
Stabilization of charges	D^-	

Table 1 gives an overview of the physicochemical effects for which methods for their calculation have been developed.

All these physicochemical effects were conceived many decades ago by chemists to explain their observations on chemical reactions; they evade a clear theoretical basis. Therefore, in order to establish that the values calculated by the methods presented above have significance, one has to establish that they model the effects they were coined for, that they are useful. First, this was established by correlations with physical data such as dipole moments and chemical shifts in ESCA or ^{1}H and ^{13}C NMR spectra as here rather exact data were available.

Next, the usefulness of these values calculated for physicochemical effects in correlating chemical reactivity data was investigated. In this endeavour, we intentionally first analyzed and modelled data on gas phase reactions in order to study the inherent reactivity of molecules uncorrupted by solvent effects. Correlations with proton affinities of amines [8], of alcohols and ethers and their thio analogs [9], and of gas phase acidities of alcohols [10] showed the significance of the values calculated for these physicochemical effects and established that indeed they modelled the effects they were initially designed for. As a next step, data of reactions in solution were studied and it was shown that data on aqueous acidity (pK_a values) of alcohols as well as on nucleophilic additions to carbonyl compounds could be quantitatively reproduced by correlation with these physicochemical effects [11].

Recently, we have again taken up the task of predicting pK_a values, now with a larger set of chemical structures [12]. The pK_a values of a wide variety of 1122 substituted aliphatic carboxylic acids could be modelled with a five descriptor equation shown in Eq. 1.

$$\begin{aligned} \mathrm{p}K_a = & 19.10 - 37.54 Q_{\sigma,0} + 12.27\ A_{2D,i} \\ & + 0.11\ x_{\pi,C\alpha} - 1.02_{\alpha O} + 1.89\ I_{\text{amino}} \end{aligned} \quad (1)$$

In this equation, $Q_{\sigma,0}$, is obtained as the sum of partial charges on all atoms taken with decreasing weight the farther the atom is away from the oxygen atom where the proton is taken away. $A_{2D,i}$ is a measure of steric hindrance at the ionization site; $\chi_{\pi,C\alpha}$ is the π-electronegativity on the carbon atom in α-position to the carboxylic group, and α_O is the damped polarizability on the oxygen atom of the ionization site. I_{amino} is an indicator variable used for characterizing amino acids.

The correlation had a correlation coefficient $r^2 = 0.81$ and a standard deviation of 0.42 pK_a units (0.43 in a 5-fold cross-validation). Thus, with five descriptors the pK_a values of such a wide variety of carboxylic acids could be reproduced with an error that about corresponds to the experimental accuracy. Further studies on additional classes of compounds are in progress.

We have again started investigating fundamental reaction types as new, highly interesting data have become available. Extensive kinetic investigations by the group of H. Mayr have established a quantitative scale of nucleophilicity [13].

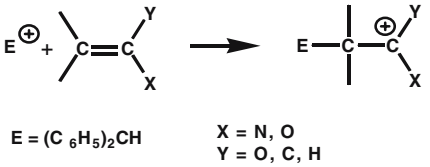
The rate of addition of electrophiles, E^+, such as benzhydryl cations to enamines, enol and phenol ethers and alkenes served as a basis for establishing a nucleophilicity parameter, N, for these substances, spanning 18 orders of magnitude. Figure 1 shows the general reaction equation and some typical nucleophiles together with their nucleophilicity value.

We could show that the nucleophilicity parameter of enamines as well as that of enol and phenol ethers can be calculated by two parameter equations of the form given in Eq. 2 (Vidovic D, unpublished results).

$$N = a[q(X) + q(Y)]\ + bE_{\text{ring}} + c \quad (2)$$

Here, $[q(X) + q(Y)]$ is the sum of the partial charges, calculated by the PEOE method [4–6] of the atoms X and Y (see Fig. 1). E_{ring} is the ring energy (ring strain or resonance energy) if either the double bond or the heteroatom is part of a ring. Figure 2 shows the results obtained by Eq. 2 for a series of enamines as well as of enol and phenol ethers, respectively (Vidovic, D. unpublished results).

It is important to note that these two variables in Eq. 2 can directly be calculated from the constitution of the nucleophiles by simple procedures such as the PEOE method [4–6] and a ring perception routine. Thus, such a fundamental concept of organic chemistry as nucleophilicity could be traced back to values on atomic valence state electronegativity (on which the PEOE



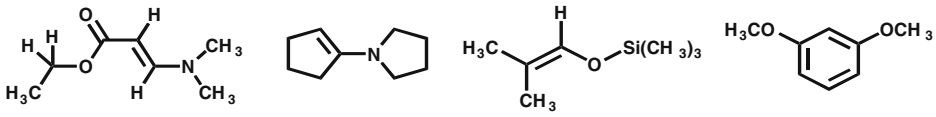


Fig. 1 The equation of the reaction type that was used for establishing the nucleophilicity parameter, *N*, together with some representative substrates and their *N*-values

method is based) and the bond structure of a molecule (inherent in the computation scheme of the PEOE method) and thereby put on a quantitative basis.

Additional support for the significance of the values calculated for the physicochemical effects has come from their extensive use in a variety of problems dealing with establishing relationships between the chemical structure of a compound and its biological activity. To this effect, the values for the physicochemical effects have been combined with information on the geometry of molecules, either the 2D structure, the 3D structure or molecular surfaces. Mathematical transformations provided uniform chemical descriptors applicable to diverse datasets of molecules [14, 15]. These methods for calculating structure descriptors have been integrated into the package ADRIANA.Code [16]. A variety of problems in drug design have been addressed with these methods such as defining the similarity and diversity of chemical libraries, analyzing the overlap of chemical libraries, locating areas of biological activity in chemical spaces, lead discovery and lead optimization, establishing quantitative structure activity relationships (QSAR), analyzing the results of high-throughput screening and modeling ADME-Tox properties [14–17].

Having assembled methods for the rapid calculation of all-important physicochemical effects that influence chemical reactions provided us with a basis for addressing the wide variety of problems associated with chemical reactions in drug design. In the following chapters we will go through the various steps of the drug design process and show where knowledge on chemical reaction is needed and how it can be gained.

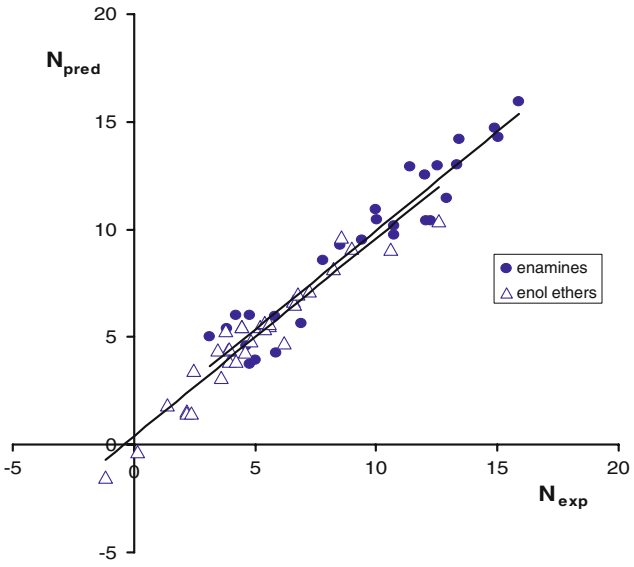


Fig. 2 Correlation between experimental and predicted nucleophilicity values, *N*, for a series of enamines as well as of enol and phenol ethers

Target identification

BioPath database

On first sight, target identification might seem to be the sole domain of bioinformatics methods. Bioinformatics methods have to find the genes that are up- or down-regulated in certain diseases and then point out which proteins are expressed by these genes. These proteins might either be enzymes catalyzing certain metabolic reactions or be proteins involved in signaling pathways.

It is our strong belief that a deeper understanding of the biochemical reactions involved in the endogeneous metabolism will help in shedding light onto the molecular basis of diseases. Small molecules have to be found that bind to enzymes, are inhibitors of enzymes, and thus can block certain metabolic reaction steps. In other situations, reaction pathways have to be searched for that may circumvent reaction steps that have been blocked. Enzymes have to be searched for that can potentially take over the role of underexpressed

enzymes. All this asks for a detailed analysis of the three-dimensional structure of enzymes and their substrates, an analysis of the bonds broken and made in an enzymatic reaction, as well as the modeling of biochemical reaction networks. This is where chemoinformatics methods have to come in.

In order to support the process of shedding more light onto the molecular basis of biochemical reactions we built a database of biochemical reactions based on the poster ''Biochemical Pathways'' distributed by Roche (formerly Boehringer Mannheim) [18] and the associated Atlas [18]. In this database chemical structures were represented with atomic resolution by connection tables giving access to each atom and each bond of the substrates, the products and of co-enzymes [19, 20].

Reactions were stoichiometrically balanced all the way down to even reporting the protons involved in chemical reactions. Furthermore, the reaction site, the bonds broken and made in a reaction, were marked by mapping the atoms of the reactants onto those of the products [19, 20]. Enzymes were identified by their name and EC code.

Having structures represented by connection tables and reactions characterized by reaction centers enables one to invoke the power of the various chemical structure and reaction search methods. We have developed a web-based structure and reaction retrieval system, C@ROL [21], that allows name, structure, substructure and structure property searches as well as reaction center searches that give access to reaction types [19, 20]. Furthermore, the database has been enriched by information on the 3D structure of the metabolites of biochemical reactions by generating a set of diverse conformations with the programs CORINA [22, 23] and ROTATE [24, 25]. This also enables the search for 3D substructures, structures that have certain features such as heteroatoms in common at specified ranges of distances in 3D space.

At the outset of our work there was no database of metabolic reactions available that contained connection tables for the species involved in biochemical reactions. In the meantime other databases on metabolic reactions have appeared that contain structures coded as connection tables, most notably the KEGG database [26] and the BioCyc collection of databases [27]. However, none of these databases has the reaction site marked, a feature of our BioPath database that is so crucial for a deeper understanding of biochemical reactions as will be shown in the next section.

In the following, two applications of the BioPath database are outlined that make specific use of the assignment of reaction centers. The first application concerns the search for inhibitors for a given enzyme even if the 3D structure of this enzyme is not known. The second application presents a novel definition of the similarity of enzymatic reactions that can provide a basis for the search of enzymes that can catalyze a desired transformation even if the original reaction has been blocked or downregulated.

Searching for enzyme inhibitors

Clearly, an enzyme must bind the substrate it is working on, but it must bind even more strongly the transition state of the reaction this enzyme is catalyzing. For, by binding the transition state of a reaction, its activation energy can be decreased and thus lead to the enormous rate enhancements of enzyme-catalyzed reactions covering many orders of magnitude as compared to the non-catalyzed reaction. Taking this idea one step further, an inhibitor of an enzyme should then be a transition state analog as pointed out by Pauling a long time ago [28]. Thus, an analysis of the geometric and electronic structure of the transition state of an enzymatic reaction should provide the structure of a query to search for inhibitors of a given enzyme even if the 3D structure of the enzyme is not known.

We have investigated this hypothesis with a variety of enzyme-catalyzed reactions such as AMP-deamination, (EC 3.5.4.6), triose phosphate isomerisation (EC 5.3.1.1) and arginine hydrolyzation (3.5.3.1) [29].

Figure 3 shows the conversion of adenosine monophosphate (AMP), **1**, to inosine monophosphate (IMP), **2**, a reaction catalyzed by the enzyme AMP-deaminase (EC 3.5.4.6). The unique features of the BioPath database, containing 3D structures and having the reaction site, the bonds broken in this reaction, marked allow the generation of the intermediate of this reaction, **3**, by adding H–OH to the C=N bond. The transition state of this reaction is in all likelihood very similar in structure to this intermediate (late transition state). An inhibitor of this enzyme AMP deaminase, carbocyclic coformycin, **4**, is also shown on Fig. 3.

To probe the transition state analog hypothesis we superimposed the 3D structure of the inhibitor onto the 3D structure of AMP, of IMP, and of the intermediate, **3**. All 3D structures were generated with the program CORINA [22, 23]. The superimposition was performed with the program GAMMA that uses a genetic algorithm to find the maximum number of atoms that can be superimposed in 3D space [30]. In this process, conformational flexibility was allowed to maximize the superimposition. Figure 4 shows the results of these superimpositions.

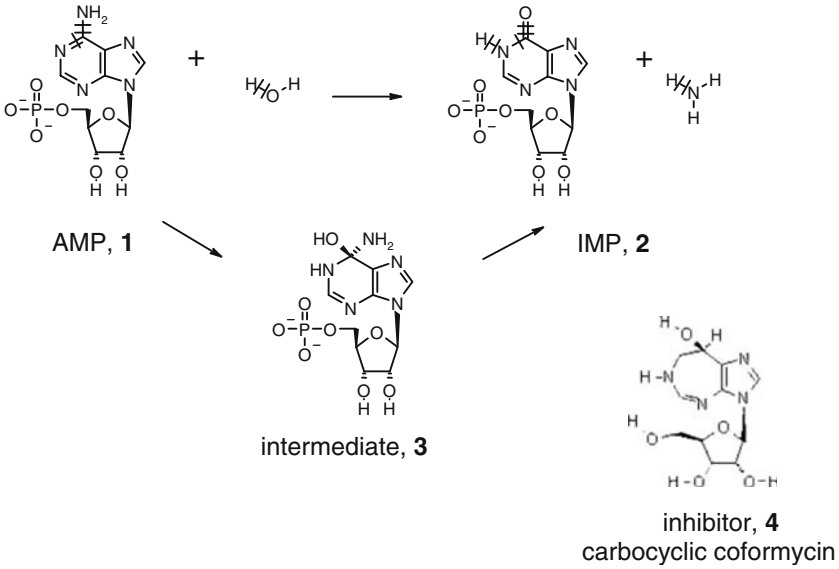


Fig. 3 Hydrolysis of AMP, **1**, to IMP, **2**, catalyzed by AMP deaminase, further showing the intermediate, **3** in this reaction and the inhibitor carbocyclic coformycin, **4**, of AMP deaminase

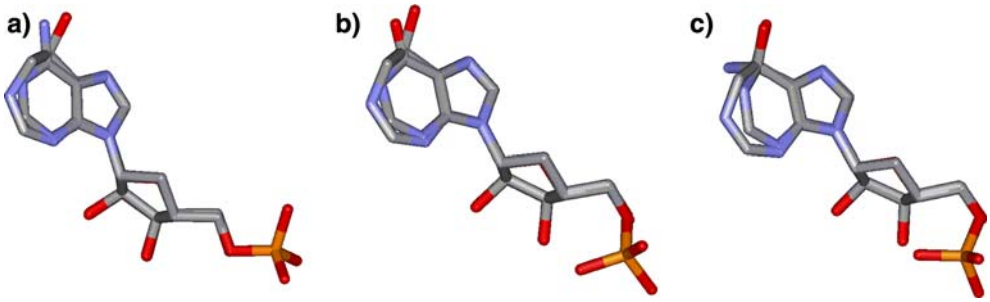


Fig. 4 Superimposition of carbocyclic coformycine, **4**, onto (**a**) AMP, **1**, (**b**) IMP, **2**, and the intermediate, **3**

Clearly, large parts of the structures can nicely be superimposed in all three cases (Fig. 4a–c) because of the high degree of similarity of the structures. The essential differences, however, show up, as expected, at the reaction site where the amino group is converted into an amide group. In the superimposition of the inhibitor with the intermediate of the reaction (Fig. 4c) a perfect alignment of the two OH-groups can be detected. Apparently, this alignment is essential in this reaction with the OH group pointing to a crucial hydrogen-bonding site of the enzyme protein. The superimpositions of the inhibitor on AMP (Fig. 4a) or IMP (Fig. 4b) show remarkable differences at this crucial location of the reaction site.

These results clearly support the idea that an inhibitor should be a transition state analog as the inhibitor of this enzyme-catalyzed reaction best matches the intermediate of this reaction. Thus, the intermediates that can automatically be generated from the reaction site mapping contained in the BioPath database can serve as queries for searching in structure databases for enzyme inhibitors [29].

Searching for similar enzymes

Enzymes are traditionally classified by the Enzyme Commission (EC) code developed and maintained by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) [31].

However, this classification code is based on a variety of criteria such as

- reaction patterns,
- substrates,
- transferred groups, or
- acceptor groups

For example, for oxidoreductases (EC 1) the focus jumps between electron donor and acceptor groups. The transferases (EC 2) are subgrouped by the type of transferred group. The hydrolases (EC3) are subgrouped by the type of bond that is cleaved in the reactions they catalyze, and the lyases (EC 4) are also subdivided by the type of bond broken in their reactions. The isomerases (EC 5) focus on the type of isomerisation, and the ligases (EC 6) are subdivided

by the type of bond formed during the reactions catalyzed by these enzymes. Thus, with the enzymes of class EC 1, EC 2 and EC 5, the classification is based on rather phenomenological criteria that have nothing to do with the mechanism of the enzymatic reactions. In the enzymes of classes EC 3 and EC 4 the bond broken is considered whereas with enzymes of class EC 6 the bond made is taken into account.

When, however, one is searching for an enzyme that can catalyze a certain transformation, e.g., a reaction where the corresponding enzyme has been blocked or down-regulated, then this search certainly must take into account the shape of the enzyme pocket, the arrangement of the binding site in the enzyme pocket (cf. the previous section 'Searching for enzyme inhibitors') and mechanistic details of the reaction to be performed.

In this section we will concentrate on the latter aspect, the physicochemical features influencing enzyme-catalyzed reactions. We will define the similarity of such reactions by physicochemical descriptors for the bonds broken in enzymatic reactions. As a first entry we take a set of reactions where the EC code is indeed based more on mechanistic concepts. This is the case for reactions that are catalyzed by hydrolases (EC 3) because here the EC code focuses on the functional groups that are attached by a water molecule. It is expected that for this set of reactions the EC code will conform with a reaction classification based on physicochemical effects. To better show that fine details of the reaction mechanisms are encoded by the physicochemical effects, we focused the investigation on an even more restricted data set of the class EC 3, the subclass EC 3.1.

The dataset consisted of 52 reactions of the subclass 3.1.c.d comprising the subclasses of hydrolases acting on carboxylic esters (3.1.1.d) with 14 reactions, on thioesters, (3.1.2.d) with four reactions, phosphoric acid monoesters, (3.1.3.d) with 23 reactions, phosphoric acid diesters, (3.1.4.d) with nine reactions, triphosphoric acid monoesters, (3.1.5.d) with one reaction and sulphuric acid monoesters, (3.1.6.d) with one reaction.

The reactions were characterized by the physicochemical effects of the bond broken in the substrate (the other bond involved in the reaction always is a bond in the water molecule and can therefore be neglected) (Reitz M and Gasteiger J, submitted). The following physicochemical effects were used for characterizing the bond broken in the substrates

- difference in total partial atomic charges, Δq_{tot},
- difference in σ-electronegativities, $\Delta\chi_{\sigma}$,
- difference in π-electronegativities, $\Delta\chi_{\pi}$,
- effective bond polarizability, α_b,
- delocalization stabilization of a negative charge, D^-,
- delocalization stabilization of a positive charge, D^+.

It should be recalled that the charge and electronegativity values were obtained by the PEOE method (see 'Methods') and thus are dependent on the atoms involved in the bond and the molecular environment of the bond considered. Each bond is then characterized by six physicochemical descriptors, and therefore, each enzymatic reaction constitutes a point in a six-dimensional spaces spanned by the six physico-chemical descriptors as coordinates. A self-organizing neural network (Kohonen network) [32–34] was used to project this six-dimensional space into a two-dimensional map (Reitz M and Gasteiger J, submitted). Figure 5 shows how the different subclasses are distributed in this map. The subclasses 3.1.1 to 3.1.4 are clearly separated into different areas of this maps. The subclasses 3.1.5 and 3.1.6 each contain only a single reaction which are mapped into spaces (neurons) of other reaction types.

In this case, where the EC code is concentrating on mechanistic criteria, on the type of bond that is broken in a reaction, the classification of reactions by physicochemical effects largely corresponds to the EC code. In other cases, where the EC code is less stringently concentrating on the events on the reaction sites, on the bonds broken and made in an enzymatic reaction, the classification of reactions based on physicochemical effects shows more clearly the events at the reaction site and thus is a better basis for searching for similar enzymes that might catalyze the same reaction types than the EC code.

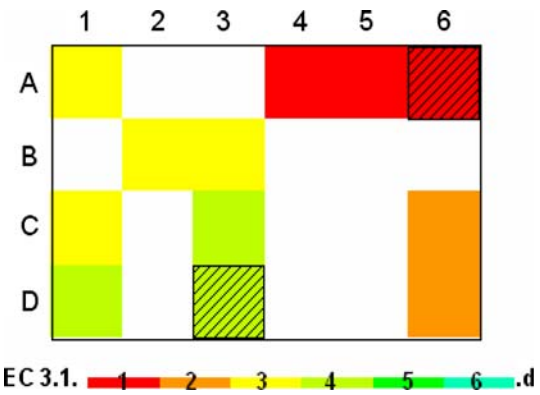


Fig. 5 Self-organizing map of a dataset of 52 reactions of the subgroup EC 3.1 showing the separation into subgroups 3.1.1 (red), 3.1.2 (orange,) 3.1.3 (yellow), and 3.1.4 (green)

Lead discovery and lead optimization

Overview

The process from a hit through a lead to a drug candidate is in practice certainly never a straightforward procedure where each step naturally and logically follows the preceding step. Rather, there will be many feedback-loops, information gathered in the lead optimization process will be used to take a new approach to lead discovery or to refine this step. This certainly also applies to the information gained while running reactions at the various stages of the lead discovery and lead optimization process. Compounds and libraries of compounds have to be synthesized both in the lead discovery and lead optimization steps, knowledge on the scope and limitations of a reaction type will influence whether a certain lead is further pursued or is discarded. If a lead cannot be further optimized or leads to problems with its ADME-Tox properties, another lead structure has to be chosen. In the following, we will therefore not distinguish between lead discovery and lead optimization but focus on the problems encountered in running reactions during these steps and show how chemoinformatics methods can help in solving these problems.

Synthetic accessibility

When the three-dimensional structure of the target protein is known, either docking of small molecular structures into the binding pocket can be performed or one might grow novel molecular structures within the binding pocket by de novo design methods. These structures have then to be synthesized in order to establish whether they can serve as a lead. De novo design methods usually generate a large amount of molecules many of them having quite complicated structures. It therefore becomes imperative to evaluate these compounds as to their ease of synthesis and first concentrate on those compounds that are deemed to be more easily synthesized.

We have developed a method for estimating synthetic accessibility that involves an analysis of the molecular structure such as the complexity of the molecular skeleton, degree of functionality, and number of stereocenters. Furthermore, it compares the query structure as to its similarity with available chemicals and, in addition, searches for reactions from a reaction database that could potentially be used to synthesize that compound [35].

In order to arrive at a numerical scheme for synthetic accessibility, a basis for comparison was needed. To this effect a set of structures was selected from recent publications in the Journal of Medicinal Chemistry and enriched with a few highly complicated structures such as taxol, providing altogether 100 different molecules. These were given to five chemists to assign them a value of ease of synthesis in the range of 1–10, with 1 being a compound that was deemed to be very easy to be synthesized and 10 given to a molecular structure having a high complexity and expected to be accessible only through a synthesis involving many steps and difficult to be performed. Clearly, such evaluations will be influenced by the personal experiences and knowledge of the individual chemists. Accordingly, the values given by the various chemists agreed only to an extent corresponding to correlation coefficients between 0.74 and 0.84. Figure 6a shows the correlation of the values assigned by two different chemists.

Clearly, the correlation leaves something to be desired but, as said, the values given by chemists will reflect their personal insights and thus cannot be expected to be much better. The different criteria used for our computational scheme of estimating synthetic accessibility were combined into a single value of synthetic accessibility by a simple additivity scheme where each criterion was weighted with the weight being determined by scaling it to the average of the values assigned by the five chemists. Figure 7 shows some structures and their assigned values of synthetic accessibility.

Figure 6b shows the correlation of the values assigned by our computational scheme with those of chemist 1 of Fig. 6a.

The two correlations of Fig. 6 are of equal quality. Thus, it can be said that we have developed a scheme that can estimate synthetic accessibility with a quality comparable to the insight of a chemist. The advantage of this computational scheme is that it can perform the assignment of synthetic accessibility rapidly and can therefore be applied to large datasets of molecules. The value of synthetic accessibility can be used in a variety of ways in real life applications. Thus, for some tasks it might suffice to choose a cut-off value for synthetic accessibility and select only those compounds having a value lower than this cut-off.

Design of a synthesis for compounds and compound libraries

Computer-assisted synthesis design has been a challenge from the very beginning of chemoinformatics.

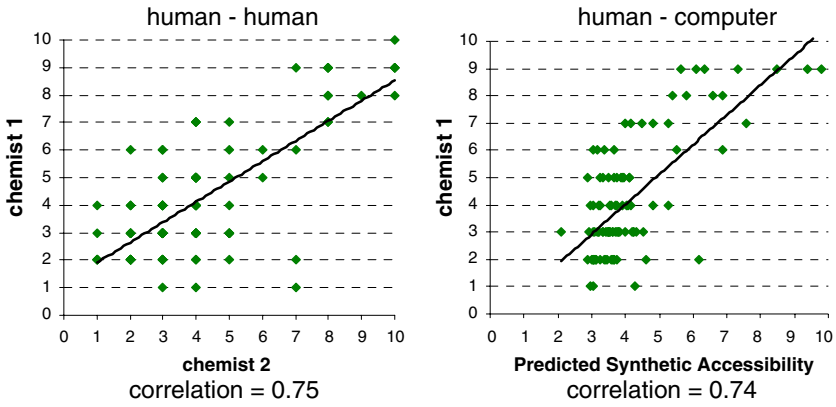


Fig. 6 (**a**) Plot showing the values of synthetic accessibility assigned by two chemists to a set of 100 compounds; (**b**) plot of synthetic accessibility values assigned by chemist 1 of Fig. 5a and of those obtained by our computational scheme for the same set of compounds

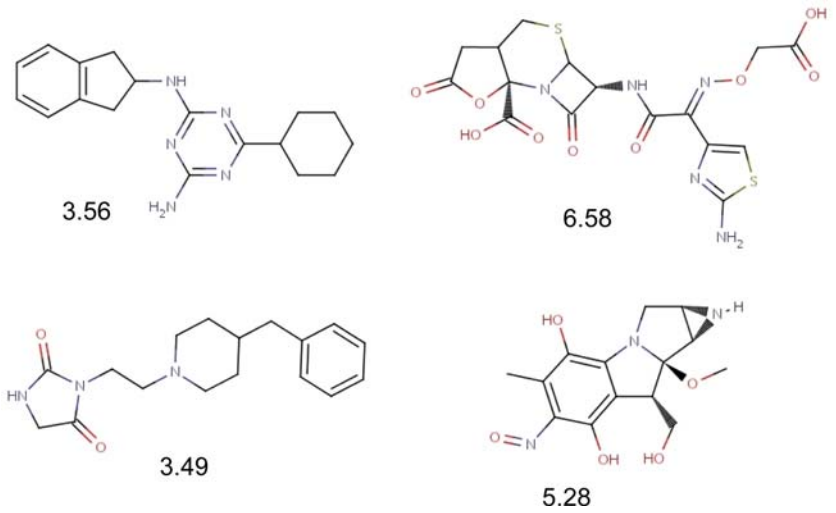


Fig. 7 Four selected structures from the dataset of 100 compounds with their computed synthetic accessibility values

Already in the late sixties and early seventies groups at Harvard (LHASA), Princeton (SECS), Stony Brook (SynChem), Munich (CICLOPS, EROS), and Brandeis (ChemSyn) had started to work on systems for designing the synthesis of complex organic molecules. Our work has progressed through a series of versions arriving at the WODCA system (Workbench for the Organization of Data for Chemical Applications) [36, 37]. As the name implies, WODCA provides a series of tools that the user/chemist can choose from to design a synthesis. Figure 8 gives an overview of the methods contained in the WODCA system.

Here is not the place to discuss all the methods integrated into the WODCA system in detail and to present a range of applications to real life problems. A brief presentation of the methods should suffice. Searches in catalogs of available chemicals for compounds similar to the target structure of a synthesis or to synthesis precursors are used to guide a synthesis plan as rapidly as possible to available starting materials. Various criteria for the definition of similarity are provided that have specifically be designed for the planning of a synthesis, criteria that are either based on common substructures or on generalized reactions. Strategic bonds are defined on the basis of physicochemical effects in order to define bonds that can be made by powerful synthesis reactions. Breaking these strategic bonds thus simplifies the synthesis problem and provides precursor molecules. In this process, queries are generated that allow automatic searches in a reaction database for similar reactions and thus can verify whether the disconnection of a strategic bond indeed leads to a well-known reaction with a broad scope.

Substructure searches allow the generalization of precursor molecules and thus provide a set of starting materials for the synthesis of combinatorial libraries. Figure 9 gives an example of a synthesis for a combinatorial library designed with the WODCA system.

Figure 10 shows the results of substructure searches in the Fluka catalog (16,769 structures) for the three building blocks required for the synthesis shown in Fig. 9.

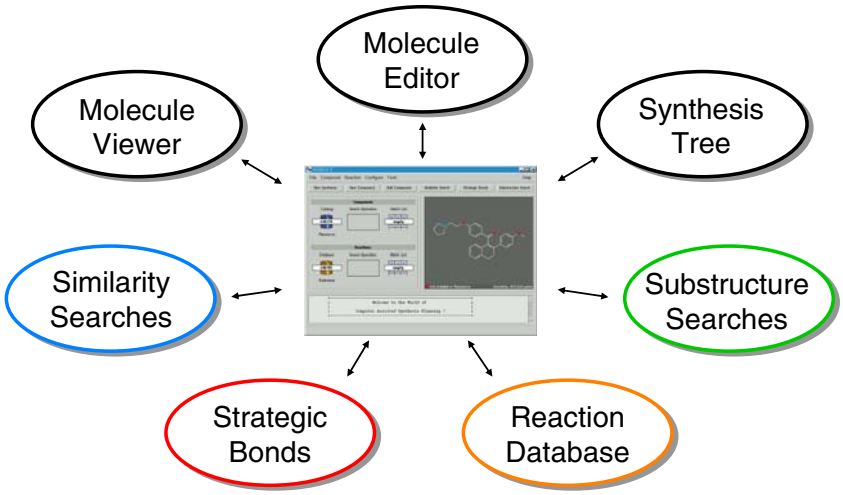


Fig. 8 Methods contained in the WODCA system for the design of organic syntheses

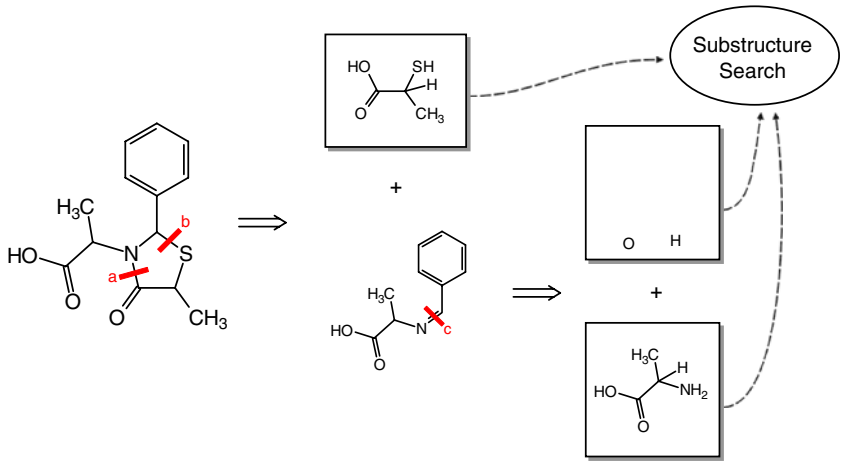


Fig. 9 Design of the synthesis of a combinatorial library of thiazolidones

All these tools, from the input of the target molecule, through similarity searches, the searching for strategic bonds, the generation of synthesis precursors or of precursors for combinatorial libraries and the retrieval of reactions from reaction databases communicate with the user in the language he/she is used to, in the form of structure diagrams and reaction equations.

WODCA is quite a comprehensive system for synthesis design providing methods allowing a chemist to design syntheses of complex organic molecules and of combinatorial libraries. However, we have realized that chemists in the lab in their day-to-day work do not often have the time for a full exploration of all capabilities of the WODCA system. Rather, they need a simple to use system that rapidly provides a synthesis plan that is short and has a high chance of being feasible because of being based on reactions having much precedence.

In other words, the design of highly complex molecules is more the exception than the rule. In most cases, chemists have to synthesize moderately complex molecules as rapidly and efficiently as possible. In order to meet this requirement we have recently developed another approach to synthesis design, a RetroSynthesisBrowser, that is heavily relying on searching in reaction databases and in databases of available starting materials for developing a synthesis plan (Bienfait B, unpublished results).

To this effect a reaction database is converted into a database of retroreactions. The synthesis target is automatically analyzed for substructural features that match the reaction centers in the database of retroreactions. Thus, the target is converted into synthesis precursors on the basis of reactions that are founded on known chemistry. Precursor molecules are automatically checked whether they are commercially available. Those retroreactions that provide available starting

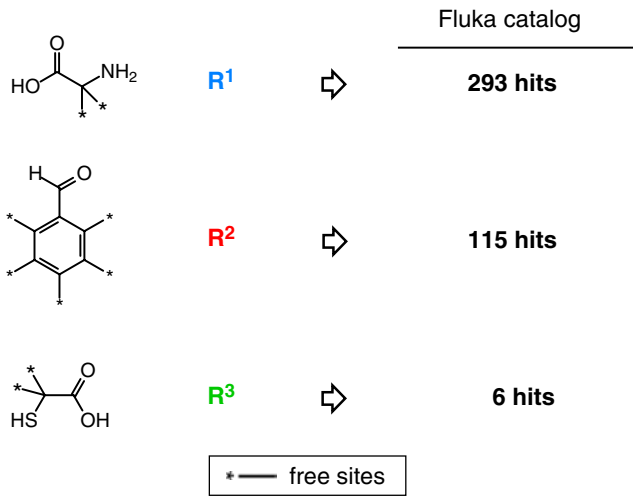


Fig. 10 The results of substructure searches for the building blocks of the synthesis shown in Fig. 9

materials obtain a higher ranking and are presented first. For all retroreactions generated, examples of similar reactions from the reaction database are shown.

Figure 11 shows a target compound, a substituted 1-phenyl-isoquinolidine, for which a seven step synthesis was reported in the literature [38]. Submitting this target structure as query to the Retro Synthesis Browser produced a retrosynthesis step for which one compound, the benzaldehyde was already commercially available. By clicking on the other compound needed for this reaction a second retrosynthesis step was produced that needed a starting material that was commercially available. Thus, the design of the synthesis could be terminated producing the desired target compound in a two-step synthesis from commercially available compounds.

The most similar reaction [39] to the first retrosynthetic step as retrieved from the reaction database is shown in Fig. 12.

Similarly, Fig. 13 shows the reaction [40] retrieved from the database that is most similar to the second retrosynthetic step.

As can be seen, the reactions from the reaction database are quite similar to the suggested synthesis steps and, therefore, it can be expected that the synthesis has a high chance for success.

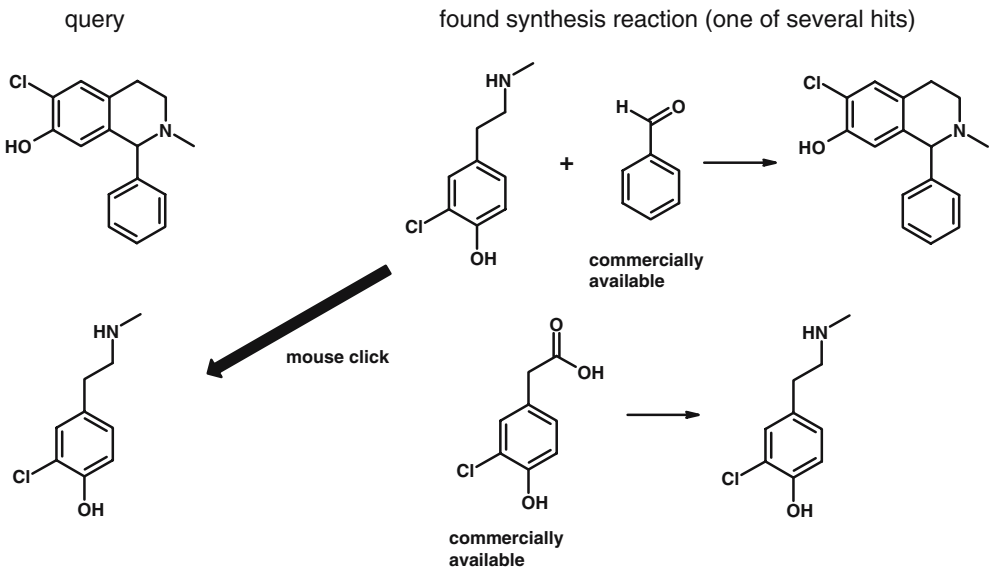
As already said, the synthesis of this compound as reported in the literature comprised a 7-step procedure. Thus, the synthesis scheme developed by using the RetroSynthesisBrowser could be the basis for a major improvement against the literature synthesis.

Knowledge extraction from reaction instances

Much of the knowledge chemists have acquired about chemical reactivity and about the course and products of a chemical reaction has been gained from observations. This learning from experimental observations and from data is called inductive learning and has been the cornerstone for acquiring knowledge in chemistry. As concerns our knowledge about chemical reactivity, the following steps are involved.

A series of related reactions comprising a reaction type is analyzed, a mechanism for such a reaction type is proposed and the physicochemical effects influencing this mechanism are rationalized. On this basis, predictions on the feasibility of other instances of such a reaction type can be made. With more and more information about chemical reactions being stored in electronic form in reaction databases, this process of

Fig. 11 Target structure and the two-step synthesis scheme produced by working with the RetroSynthesisBrowser



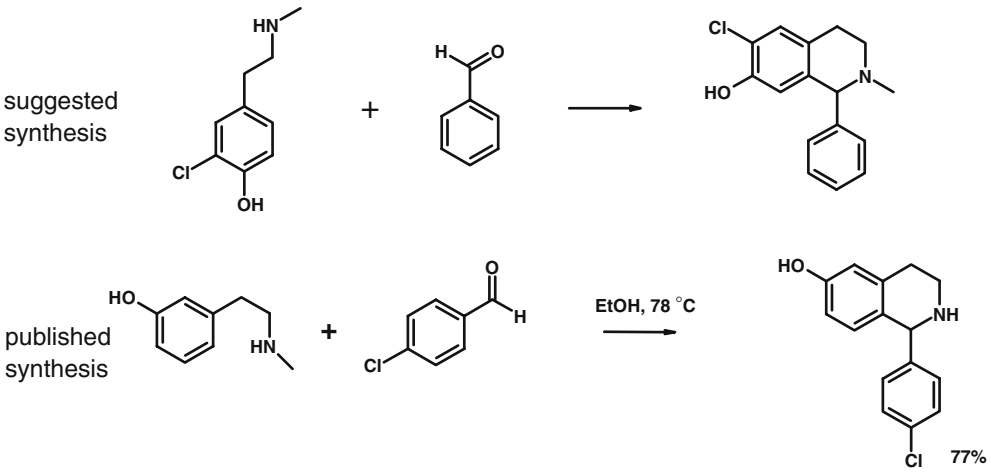


Fig. 12 The first retrosynthetic step and its most similar reaction from the reaction database

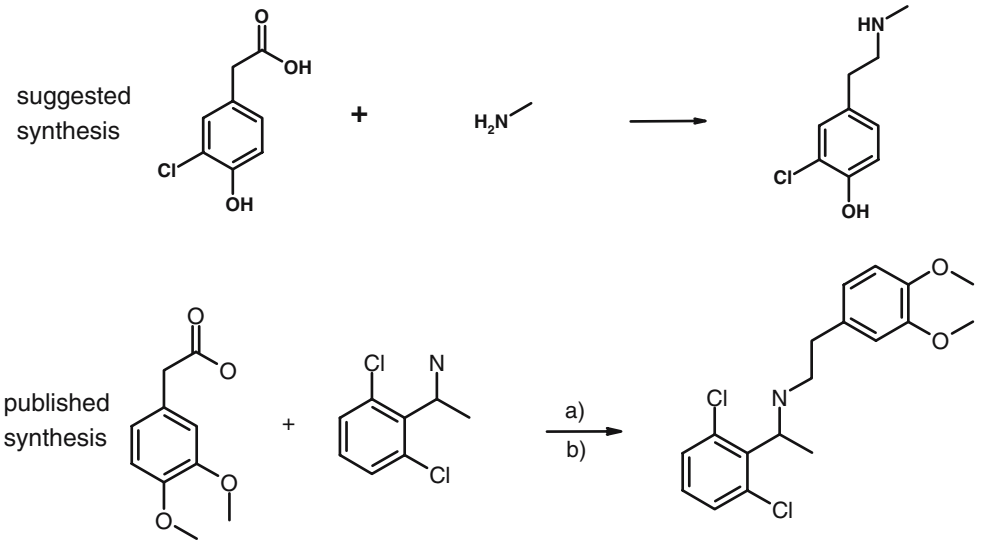


Fig. 13 The second retrosynthetic step and its most similar reaction from the reaction database

acquiring knowledge about chemical reactions can now be pursued in silico by data analysis and data mining methods.

Reaction databases comprising millions of reaction instances have become available and provide a large source of information. However, it has to be realized that this information in reaction databases is often incomplete: Reaction equations are not stoichiometrically balanced, only the major reaction product is specified, not all information on reaction conditions is given, or reaction condition vary from reaction instance to reaction instance rendering a comparison of reactions questionable. However, new light is appearing at the horizon: Firstly, parallel synthesis experiments run a series of reactions of a certain reaction type exactly under the same conditions and thus provide a good basis for comparing these reactions for learning more about this reaction type. Secondly, lab chemists increasingly store their observations on the chemical reactions they perform in electronic laboratory notebooks (eLN). Thus, the person that knows most about a reaction instance because he/she has performed this reaction will be the one that inputs the data into an eLN from where it can directly migrate into a reaction database. Thus, we will see a large surge in the quality of information available on chemical reactions, putting us in a much better position of learning more about chemical reactions.

We will first see how values on the physicochemical effects at the reaction site can indeed be used to define a reaction type and the scope and limitation of such a reaction type. Then, we will explore how information in a database can be used to make predictions on the regioselectivity of a reaction type. Crucial for the

success of such a process of learning from reaction instances is the representation of the information on chemical reactions, foremost the representation of the physicochemical effects influencing chemical reactivity.

Clustering reaction instances into reaction types

The first step in learning about chemical reactions is to group those reactions together that share some common features and can be classified into a reaction type. This process must perceive those features that are necessary for a certain reaction type and it must generalize those features away that can vary from reaction instance to reaction instance.

We have shown that values of the physicochemical effects calculated by the methods presented in the introduction can be used to group or cluster reactions into reaction types. An example should serve to illustrate this: A dataset of 120 reactions involving the breaking of a C–H and of a C=C bond and the making of a C–C and of a C–H bond was selected [41]. This dataset comprised a variety of reaction types such as Michael additions, Friedel-Crafts alkylations by olefins, Nazarov reactions, etc. The bonds broken were characterized by seven physicochemical effects.

In the selection of the physicochemical effects to describe these reaction types we were guided by the mechanism of these reactions. For the carbon atom of the reacting C–H bond the π-charge, q_π, the π-electronegativity, χ_π, the σ-electronegativity, χ_σ, and the atom polarizability, α_d, were chosen. The reacting C=C bond was characterized by the π-charge, q_π, of the atom that will be attacked by the carbon atom of the C–H bond; the other carbon atom was accounted for by its π-electronegativity, χ_π, and its σ-electronegativity, χ_σ. Thus, as each reaction is described by seven physicochemicals descriptors, in effect, a chemical reaction is a point in a 7-dimensional space and a self-organizing neural network (Kohonen network) was used to project the reactions from this 7-dimensional space into a 2-dimensional map [41]. The resulting projection of the reactions into this two-dimensional map was then annotated by letting these reactions be assigned to reaction types by chemists.

Figure 14 shows the distribution of the reaction instances in this map and illustrates that reactions assigned by a chemist to one and the same reaction type are, indeed, projected into coherent areas of the map.

Thus, the physicochemical descriptors do catch the essential effects that define a reaction type. Even further, it was shown that the reactions that are mapped into the outskirts of the area of a reaction type such as the Michael additions correspond to reaction instances that have somehow unusual features and therefore extend a reaction type [41]. In this example, the physicochemical descriptors characterizing a reaction site were intellectually chosen so as to bring in the essential effects deemed operating in these reaction types. However, it was also shown that an automatic selection of descriptors with statistical methods based on the Fisher index led to essentially the same set of physicochemical descriptors. Further work has allowed us to define a general set of descriptors that can be used with advantage for a wide range of reaction types. In the next application, this standard set of descriptors will be used.

Modelling the selectivity of a reaction type

Given two reactants, these can react with each other quite often in a variety of ways, by different reaction channels. Thus, one has to address the problem of selectivity to decide which reaction is indeed pursued. Different types of selectivity have to be addressed:

- chemoselectivity, i.e., selecting between different reaction types
- regioselectivity, i.e., selectivity between different reaction sites
- stereoselectivity, i.e., selectivity providing different stereoisomers

In the example discussed here, we address the problem of regioselectivity. 1,3 diketones react with hydrazines to give pyrazoles. If unsymmetrical 1,3-diketones and mono-substituted hydrazines are used two regioisomers can potentially be performed as shown in Fig. 15.

The decision which regioisomer will be formed has to consider a variety of factors such as electrophilicity of the carbonyl carbon atoms, nucleophilicity of the nitrogen atom and the influence of steric factors. Thus, it will not be easy to decide which of the two reaction pathways will indeed be followed. We investigated how information contained in databases can be utilized to decide on the regioselectivity of this reaction type (Sacher O, unpublished results).

Searching in the Beilstein database for reactions of unsymmetrical 1,3-diketones yielding more than 50% of a regioisomer provided 313 reactions. Learning from data also needs information on the limitations or the failure of the system being investigated. Unfortunately, no information was contained in the Beilstein database whether the other regioisomer was also formed albeit in some lower yield. In order to provide information on the limitations of this reaction type we made the

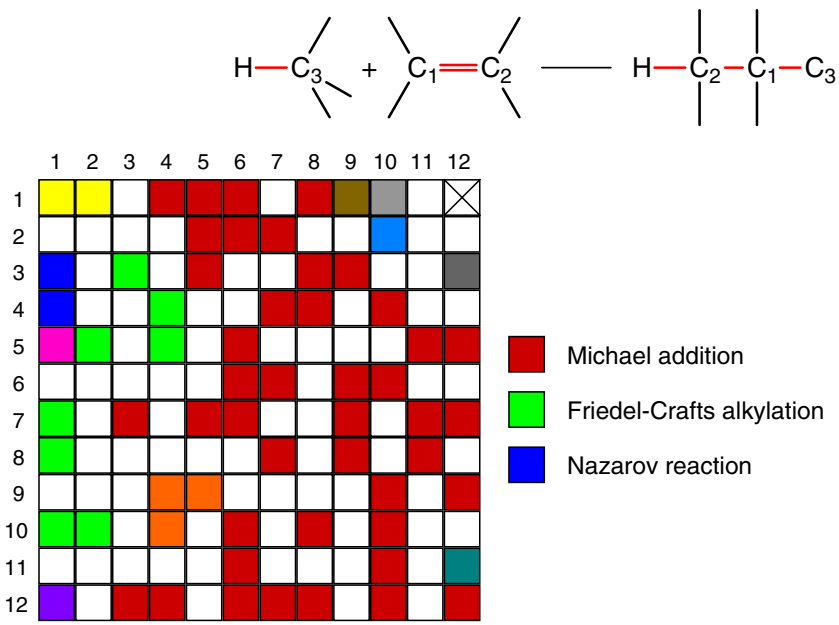


Fig. 14 Projection of a dataset of 120 reactions involving the addition of a C–H bond to a C=C bond into a self-organizing map. Different reaction types are identified by different colors. Only the assignment as of Michael additions, Friedel-Crafts alkylations by obfins and Nazasor reactions are explicitly indicated

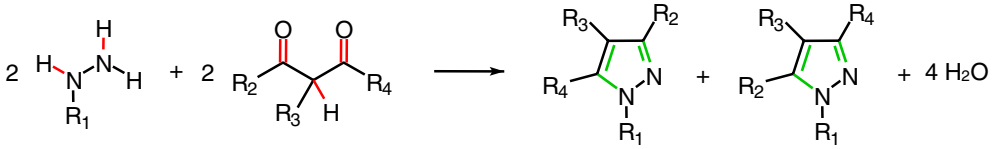


Fig. 15 The reaction of a mono-substituted hydrazine with an unsymmetrical 1,3-diketone can give either one of two regioisomers

assumption that the other regioisomer is never produced, an assumption that is certainly wrong in quite a few cases. Thus, another 313 reactions to the other regioisomer were generated and characterized as not occurring.

In order to characterize the reaction, the five bonds of the reactants more or less directly involved in the reaction process were chosen. Each bond was characterized by our standard set of six physicochemical effects:

- bond order, b_o,
- difference in partial charges, Δq_{tot},
- difference in σ-electronegativity, $\Delta\chi_\sigma$,
- difference in π-electronegativity, $\Delta\chi_\pi$,
- delocalization stabilization of a positive charge, D^+,
- delocalization stabilization of a negative charge, D^-.

Each reaction is then described by 30 descriptors (6 physicochemical effects for each of the 5 bonds of the reaction center). The dataset consists of 313 reactions to the correct regioisomer and 313 reactions to the wrong regioisomer.

The question is then, can the physicochemical descriptors separate the observed reactions from the reactions to the wrong regioisomer? A variety of data analysis methods could be used for answering this question. For the sake of easy visualization we have again used a self-organizing (Kohonen) neural network to project the 30-dimensional space into a two-dimensional map. Figure 16 shows the map thus obtained (Sacher O, unpublished results).

As can be seen, the reactions to the observed regioisomers (in green) clearly separate from the reactions to the non-observed regioisomers (in red); areas where no decision can be made are labelled in gray. Thus, this result shows that the chosen physicochemical effects model the effects that govern the regioselectivity in these reactions. The Kohonen network in effect stores the relationship between the physicochemical descriptors and the regioselectivity in the formation of pyrazoles. Thus, it can be used for making predictions on which regioisomer is formed for novel situations. As an example, the 30 physicochemical descriptors were calculated for the formation of both regioisomers in the reaction of 4,4,4-trifluoromethyl-*p*-fluorophenylbutadione-1,3 with 4-methylsulfonylphenylhydrazine.

One of the reactions was mapped into the area of the wrong regioisomers (red) whereas the other reaction was mapped into the area of the observed

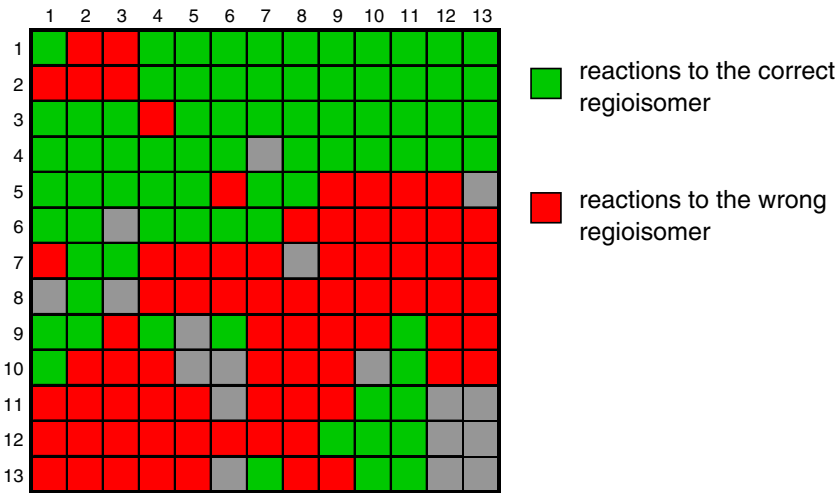


Fig. 16 Self-organizing map for the reaction of a monosubstituted hydrazine with an unsymmetrical 1,3-diketone. Observed regioisomers are projected into the area colored green, not-observed regioisomers are projected into the area colored red

regioisomers (Fig. 17). Thus, on this basis the prediction is made that the isomer that is mapped into neuron (2,6) (green) shown on the left-hand side is produced. This, indeed is the observed reaction leading to a COX-2 inhibitor.

This example, showed how observations on the products of a set of reactions can be used to build a model for predicting the outcome of a reaction. The available information was rather crude indeed, only reporting the yield of one regioisomer. No information was available whether the other regioisomer was also formed albeit in some lower yield. Nevertheless, a reasonable model for predicting correct regioisomers could be made.

With the advent of electronic laboratory notebooks more and more detailed information on the course and outcome of chemical reactions will be stored by the primary investigators. This will greatly increase the quality of information in reaction databases and thus provide a much better basis for deriving knowledge on chemical reactions.

Stability of compounds

Many compound libraries are stored in dimethylsulfoxide (DMSO) solution at low temperature. However, even at low temperatures compounds might decompose over time, particularly as DMSO nearly always contains a certain amount of water. In order to systematically investigate the stability of compounds in DMSO and in DMSO/water mixtures, a project was initiated by a consortium of eight pharmaceutical companies under the direction of Specs company. About 12,810 compounds were kept in DMSO and DMSO/20% H_2O at temperatures of 50°, 20°, 5° and –20 °C over periods of 0, 14, 35 and 135 days. At each time point the percentage purity of these compounds was measured by LC/MS.

These data allowed the classification of compounds into stable ones (if the purity was higher than 80%) and unstable ones (if the purity was lower than 80%). These data were then used to build a model for the prediction of the stability of compounds.

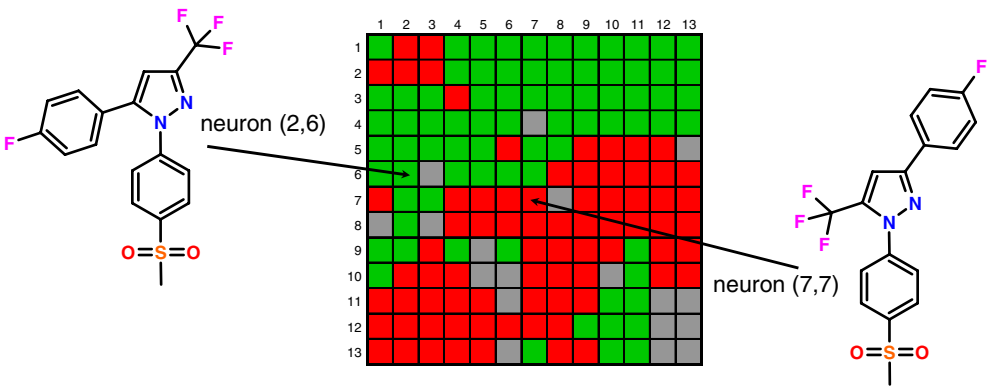


Fig. 17 Mapping of the reactions to the two potential isomers of a selected example into the self-organizing map shown in Fig. 13

First, data in DMSO/20% H_2O obtained after keeping the compounds for 105 days at 50 °C were investigated.

Many of the compounds contained several functional groups that all could potentially undergo a reaction. In order to investigate a certain reaction type in isolation, not disguised by the reaction of another functional group, single substructure datasets were first generated containing only one functional group that might undergo a reaction. Among these was a dataset that contained 34 compounds having only ester groups; 21 of these compounds were found to be stable, 13 unstable.

Figure 18 shows the mechanism of ester hydrolysis and the physicochemical descriptors that were used to describe the reaction center:

- for the carbonyl C-atom, the partial atom charge, q_{tot} ,
- for the carbonyl C = O bond, the charge differences, Δq_{tot}, the potential for stabilizing a positive charge on the carbonyl carbon atom through delocalization, D^+, and the bond polarizability, α_b,
- for the C–O bond, the bond polarity, Δq_{tot}, and the potential of stabilizing the incipient charges on heterolysis of this bond, $D^{+/-}$.

Furthermore, two indicator variables were used to distinguish whether the ester is part of a ring (lactone) or whether the ester group is attached to an aromatic system (Hu QN, unpublished results).

With these 8 descriptors a classification model was established by PLS that had a classification accuracy of 88.2% and of 82.4% in a 5-fold crossvalidation experiment.

In a similar fashion, analogous reactivity models were obtained for other reaction types. These models were then applied in conjunction to predict the stability of multifunctional compounds.

NF-κB activity

Quite a few biological activities are caused by covalent binding of a compound to a protein or to a nucleic acid. Sesquiterpene lactones (STLs) are a large group of secondary plant metabolites comprising more than 4,000 compounds of various skeletal types. A variety of STLs such as compounds **5** and **6** (Fig. 19) possess considerable anti-inflammatory activity. Several studies have shown that these STLs exert this effect in part by inhibiting the activation of the transcription factor NF-κB. A dataset of 103 structurally quite diverse STLs was investigated to develop a quantitative structure–activity relationship (QSAR) model [42].

A variety of different structure descriptors were studied either alone or in combination and statistical methods were used to reduce the number of descriptors. The relationship between structure and biological activity was modelled by a counterpropagation neural network. The investigations clearly pointed out the importance of using descriptors based on a 3D molecular model and it further emphasized the significance of π-effects. In particular, values on π-electronegativity, χ_π, encoded by radial distribution functions (RDF) were found to be well suited for developing a QSAR model. The components in these representations remaining after descriptor elimination allowed a direct mechanistic interpretation. For example, the combinations of χ_π in RDF encoding corresponding to a distance of 2.6 Å and 3.5 Å were found to be of particular importance. These descriptors correspond to the distances shown in Fig. 20.

The QSAR model developed in this approach nicely corresponds with the present knowledge of NF-κB activity of STLs. According to this, the α-methylene-γ-lactone unit of the sesquiterpene lactones reacts with cysteine-38 of the p65 subunit of NF-κB in a Michael

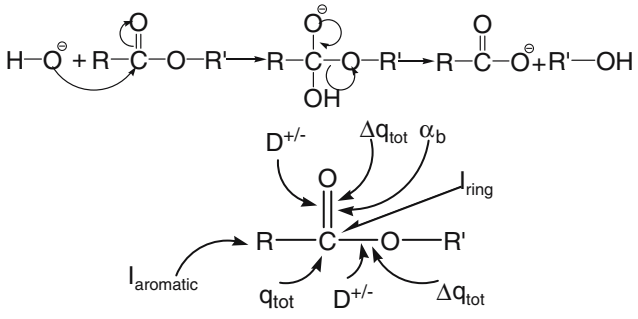


Fig. 18 Mechanism of ester hydrolysis and assignment of physicochemical descriptors used in the modelling approach

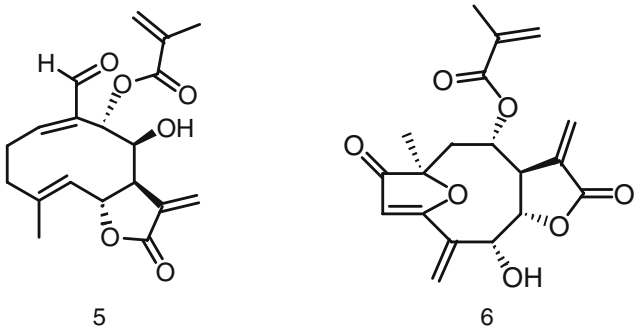


Fig. 19 Two sesquiterpene lactones (STLs) inhibiting the activation of the transcription factor NF-κB

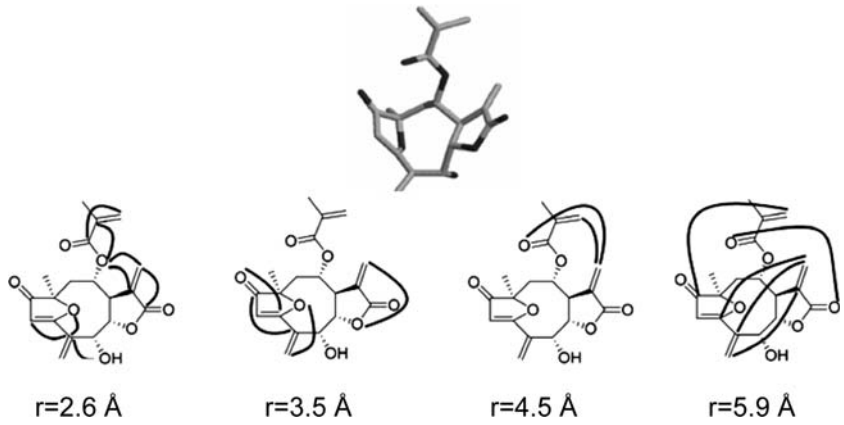


Fig. 20 Distances that were found to be important for model NF-KB activity in a radial distribution function encoding scheme

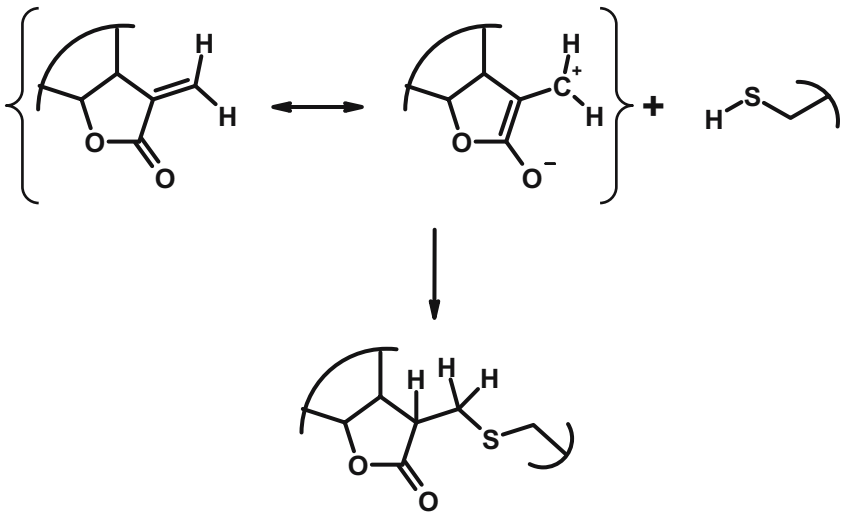


Fig. 21 Mechanism for the covalent binding of a cysteine-38 of the p65 submit of NF-κB to the α-methylene-γ-lactone of an STL

addition to covalently bind the STLs to this protein (Fig. 21)

Thus, this investigation highlighted several points: (1) The descriptors developed for treating chemical reactivity are also of value for modelling biological activities that are based on covalently binding an inhibitor. (2) A QSAR model derived with these descriptors provides deeper insights into the mechanism of such a covalent binding. Vice versa, mechanistic insights into the nature of covalent binding could be used to more directly and efficiently select descriptors for modelling biological activity.

ADME-Tox properties

Different selectivities

The term ADME collects such properties as adsorption, distribution, metabolism, and excretion that are governed by quite a variety of physical, chemical, and biological mechanisms. It is our strong belief that the development of predictive models for these individual properties should take account of these effects, should consider the underlying mechanisms that govern these properties, in order to arrive at a prediction method with a broad validity.

Clearly then, the prediction of the metabolism of a drug has to be concerned with modeling chemical reactions. A drug is a xenobiotic and can, in most cases, not directly be channelled into the endogenous metabolism of biochemical pathways, discussed in Chapter 3. Rather, nature has developed specific mechanisms, specific enzymes, that work on xenobiotics. Their primary task is to make these compounds water-soluble and then excrete them. One of the most important enzymes for phase I of this metabolism are Cytochrome P450 enzymes (CYP450) coming in a variety of isoforms.

Any explicit modeling of phase I metabolism of xenobiotics has to deal with three different types of selectivities

– selectivity between different CYP450 isoforms
– selectivity between different reaction types of these CYP's (chemoselectivity)
– selectivity between different reaction sites (regioselectivity)

An understanding of the action of CYP's and the modeling of these different selectivities has to balance steric and shape effects that govern the approximation of the substrate to the reaction site of a CYP, an iron-porphyrin system, against considerations of chemical reactivity. We will briefly present two studies made to model these different types of selectivities.

Selectivity between different CYP isoforms

A dataset of 146 drugs metabolized by the cytochrome P450 (CYP450) isoforms 3A4, 2D6 and 2C9, respectively, was analyzed. These compounds had already been studied in a previous publication [43]. Figure 22 shows some typical structures and gives the CYPs that metabolize them.

In a build-up process, involving four stages, a wide variety of descriptors based on topological autocorrelation, 3D autocorrelation, shape/size related descriptors and substructure counts was taken to represent the chemical structures. The total number of descriptor components was started with 128 and ended up with 303 values. A model for the classification of a drug into the CYP isoform that metabolizes it was obtained by a support vector machine together with automatic variable selection within the Weka package. The final model used 12 descriptors comprising topological autocorrelation of σ and π-charges, a value of spatial autocorrelation at a distance of 5.8–5.9 Å, as well as the number of acid groups, the number of nitrogen atoms, the number of aliphatic amino groups and a value calculated for the molecular diameter (Terfloth L, Bienfait B and Gasteiger J, submitted). This model was able to correctly classify 89% of the compounds (in a leave-one-out cross-validation. Furthermore, for an external dataset of 233 compounds a classification accuracy of 83% was obtained (Terfloth L, Bienfait B and Gasteiger J, submitted).

Such a percentage of correct classification is probably the best one can get as quite a few drugs are metabolized by several CYPs. In fact, application of a model built with 3D descriptors based on partial charges by a counter-propagation network to dextromethorphan resulted in the prediction that this compound is metabolized with about equal probability by CYP 3A4 and 2D6 (Fig. 23).

Indeed, dextromethorphan is metabolized by CYP3A4 by a *N*-demethylation reaction and by CYP2D6 by O-demethylation.

Selectivity in *O*- and *N*-dealkylation

In order to dig deeper into understanding the reactions initiated by CYPs we investigated the significance of radical stabilities and steric effects on *O*- and *N*-dealkylation of dextromethorphan. The bond dissociation energy for breaking a C–H bond in α-position to an *O*- and to an *N*-atom was calculated to be 418 and 364 kJ/mol, respectively.

Based on this, one would expect preferential cleavage of a C–H bond in α-position to the N-atom. On the other hand, the methyl group at the O-atom is certainly less sterically hindered than the methyl group at the N-atom. In order to investigate the effect of steric hindrance of *O*- and *N*-dealkylation a quantitative measure for steric hindrance was developed.

In fact, for *N*-dealkylation, there are three different C–H bonds in α-position to the nitrogen atom in dextromethorphan which offers three potential pathways for dealkylation. Breaking the C–H bond at the bridgehead carbon atom would require the lowest bond dissociation energy as it leads to a tertiary carbon radical. However, this reaction is not observed; rather only *N*-demethylation occurs. The preference of *N*-demethylation against the other two possibilities for *N*-dealkylation is certainly caused by steric effects. In order to quantify steric hindrance for *O*-dealkylation and for *N*-dealkylation at the three different sites at

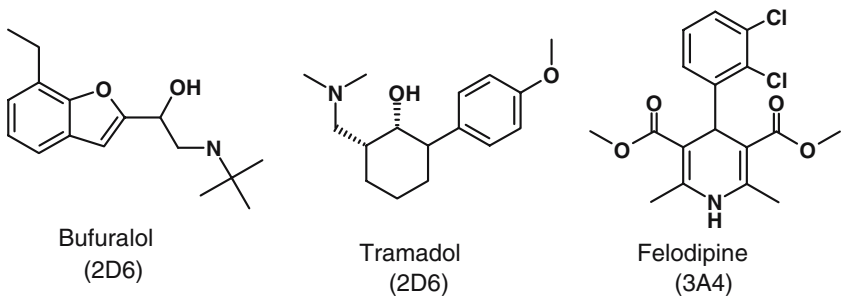


Fig. 22 Three typical examples metabolized by different CYP4506

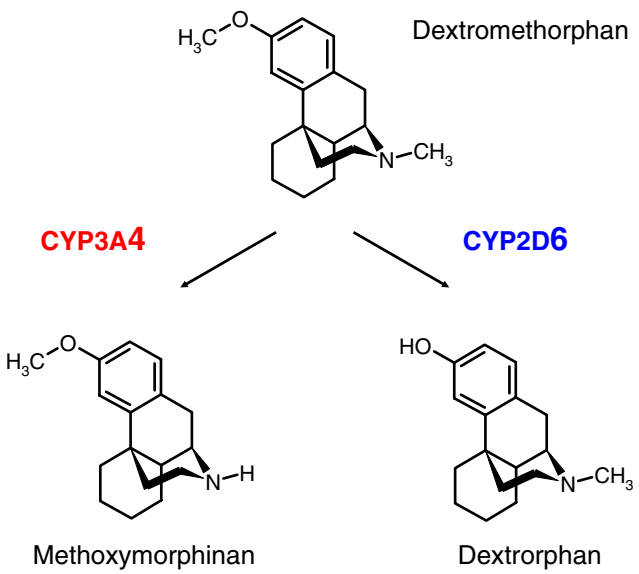


Fig. 23 Metabolism of dextromethophan by CYP3A4 and 2D6

each position a local radial distribution function [44] was calculated. Integration of these radial distribution functions in the range of 1.5–3.5 Å provided values of 0.59 for *O*-demethylation and of 1.33, 2.83 and 3.45, respectively, for the three sites for *N*-dealkylation (Fig. 24).

These results show several things; First, although the bond dissociation energy of a C–H bond in α-position to an O-atom is higher than in α-position to an N-atom, *O*-demethylation competes with *N*-demethylation because of the lower steric hindrance of attack at this position. Secondly, the lower value for steric hindrance for *N*-demethylation compared to the values for the other two α-positions to the N-atom rationalizes that only *N*-demethylation is observed. This example shows that chemo- and regioselectivity in the metabolism of drugs is governed by a subtle balance of thermochemical and steric effects. The methods presented here allow the quantification of these effects.

Conclusion

We had embarked many years ago on trying to shed light onto chemical reactivity by relying on the concepts that organic chemists had developed for rationalizing reaction mechanisms. These concepts could be put on a quantitative basis by simple empirical methods that allowed the rapid calculation of these physicochemical effects for large sets of molecules.

These physicochemical descriptors proved to be of merit in modeling a wide range of chemical reactions. In particular, this has been established for different types of reactions that have to be analyzed in the drug design and development process. This ranges from the many different types of reactions that are run to produce the compounds that are tested for their biological activity all the way to modeling and predicting both the endogenous metabolism and the metabolism of xenobiotics.

The clear interpretability of the physicochemical descriptors could always be taken as a guidance in the development of models for predicting chemical reactivity and it also allowed an understanding of the effects that govern individual reactions.

Clearly, we are only at the beginning of better understanding chemical reactions. The development of a model for a reaction type has often been hampered by a lack of data of high quality. It is our hope that in the future chemists will catch their observations on chemical reactions in all aspects and in great detail in electronic form. This can then provide the foundation

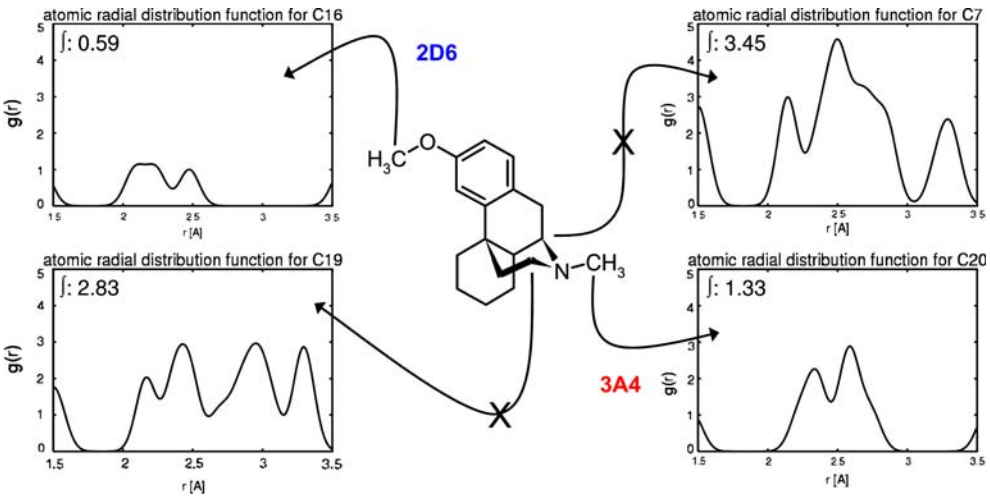


Fig. 24 Estimation of steric hindrance of *O*-and *N*-dealkylation by interpretation of radial distribution functions

for greatly deepening our knowledge on chemical reactions.

Acknowledgements Able coworkers, mentioned in the references, have embarked into the unknown domain of modeling chemical reactivity and chemical reactions. I owe much gratitude to their dedication. Our work has been funded by the Bundesministerium für Forschung und Technologie (BMFT), the Bundesministerium für Bildung und Forschung (BMBF), the Deutsche Forschungsgemeinschaft (DFG), ICI plc, UK, Pfizer, Groton, CT, USA, and Pfizer, Sandwich, UK. To all these institutions I am deeply indebted. I also thank Elsevier MDL, San Ramon, CA, USA for making their databases available to us. Over many years I have been inspired by Yvonne Martin's great scientific interest and her exploring questions. As a case in point: More than 15 years ago she asked me to come up with some methods for estimating synthetic accessibility. Unfortunately, it took us quite some time to find the time and the right people to achieve this goal.

References

1. Gasteiger J (ed) (2003) Handbook of chemoinformatics—from data to knowledge. Wiley-VCH, Weinheim, pp 1870, ISBN 3-527-3068
2. Gasteiger J, Engel T (eds) (2003) Chemoinformatics—a textbook. Wiley-VCH, Weinheim, pp 650, ISBN 3-527-30681
3. Gasteiger J (2003) In: Gasteiger J, Engel T (eds) Chemoinformatics—a textbook. Wiley-VCH, Weinheim, pp 169–202
4. Gasteiger J, Marsili M (1980) Tetrahedron 36:3219
5. Gasteiger J, Saller H (1985) Angew Chem 97:699; (1985) Angew Chem Int Ed Engl 24:687
6. Kleinöder T (2005) Ph.D. thesis, University of Erlangen-Nuremberg
7. Hutchings MG, Gasteiger J (1983) Tetrahedron Lett 24:2541
8. Gasteiger J, Hutchings MG (1984) J Chem Soc Perkin 2:559
9. Gasteiger J, Hutchings MG (1984) J Am Chem Soc 106:6489
10. Hutchings MG, Gasteiger J (1986) J Chem Soc Perkin 2:447
11. Hutchings MG, Gasteiger J (1986) J Chem Soc Perkin 2:455
12. Zhang J, Kleinöder T, Gasteiger J (2006) J Chem Inf Model 46:2256
13. (a) Mayr H, Patz M (1994) Angew Chem 106:990; (1994) Angew Chem Int Ed Engl 33:938 (b) Minegishi S, Mayr H (2003) J Am Chem Soc 125:286
14. Gasteiger J (2003) Mini Rev Med Chem 3:789
15. Gasteiger J (2006) J Med Chem 49:6429
16. ADRIANA.Code; Molecular networks. GmbH, Erlangen Germany; info@molecular-networks.com; http://www.molecular-networks.com (accessed Nov 2006)
17. An extensive list of references on the application of structure-coding methods to problems in drug design can be obtained at http://www2.chemie.uni-erlangen.de/publications/
18. (a) Michal G (1993) Biochemical pathways wall chart. Boehringer Mannheim (now Roche), Mannheim, Germany, www.expasy.org/tools/pathways (accessed Apr 2006) (b) Michal G (1999) Biochemical pathways biochemistry atlas. Spektrum Akademischer Verlag, Heidelberg, Germany
19. Reitz M, Sacher O, Tarkhov A, Trümbach D, Gasteiger J (2004) Org Biomol Chem 2:3226
20. Biopath can be accessed at http://www.molecular-network.com/biopath
21. C@ROL. Molecular networks. GmbH, Erlangen, Germany. info@molecular-networks.com, http://www.molecular-networks.com (accessed Nov. 2006)
22. Sadowski J, Gasteiger J, Klebe G (1994) J Chem Inf Comput Sci 34:1000
23. CORINA. Molecular networks. GmbH: Erlangen, Germany. info@molecular-networks.com, http://www.molecular-networks.com. CORINA can be tested on the internet at http://www2.chemie.uni-erlangen.de/software/corina/free-struct.html (accessed Nov. 2006)
24. Renner S, Schwab CH, Gasteiger J, Schneider G (2006) J Chem Inf Model 46:2324
25. ROTATE. Molecular networks. GmbH, Erlangen Germany, info@molecular-networks.com; http://www.molecular-networks.com (accessed Nov. 2006)
26. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M (2002) Nucl Acids Res 30:402
27. BioCyc Database collection. http://biocyc.org (accessed Apr 2006)
28. (a) Pauling L (1946) Molecular architecture and biological reactions. Chem Eng News 24:1375; (b) Pauling L (1948) The nature of forces between large molecules of biological interest. Nature 161:707
29. Reitz M, von Homeyer A, Gasteiger J (2006) J Chem Inf Model 46:2330
30. Handschuh S, Gasteiger J (2000) J Mol Model 6:358
31. http://www.chem.qmul.ac.uk/inbmb/enzyme/
32. Kohonen T (1989) Self-organization and associative memory, 3rd edn. Springer, Berlin
33. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design, 2nd edn. Wiley-VCH, Weinheim, pp 380, ISBN 3-527-29778-2
34. SONNIA. Molecular networks. GmbH, Erlangen, Germany, info@molecular-networks.com (accessed Nov 2006)
35. Boda K, Seidel T, Gasteiger J (2006) J Comput-Aided Mol Design (in print)
36. Ihlenfeldt WD, Gasteiger J (1995) Angew Chem 107:2807; (1995) Angew Chem Int Ed Engl 34:2613
37. Pförtner M, Sitzmann M (2003) In: Gasteiger J (ed) Handbook of Chemoinformatics—From data to knowledge. Wiley-VCH, Weinheim, pp 1457–1507
38. Hoffman B, Cho SJ, Zheng W, Wyrick S, Nichols DE, Mailman RB, Tropsha A, (1999) J Med Chem 26(42):3217
39. Kamatani T, Kigasawa K, Hiiragi M, Ishimaru H (1971) J Chem Soc C:2632
40. Richard P, Polniaszek, Craig R Kaufman J (1989) Am Chem Soc 111:4859
41. Chen L, Gasteiger J (1997) J Am Chem Soc 119:4033
42. Wagner S, Hoffmann A, Siedle B, Terfloth L, Merfort I, Gasteiger J (2006) J Med Chem 49:2241
43. Manga N, Duffy JC, Rowe PH, Cronin MTD (2005) SAR & QSAR Environ Res 16:43
44. Hemmer MC, Steinhauer V, Gasteiger J (1999) Vibrat Spectrosc 19:151