

*“Form follows function.”*

(Louis Henry Sullivan, Architect, 1896)

# 1

## Molecular Objects and Design Objectives

An ultimate goal of computer-aided molecular design (CAMD) is to propose novel substances that exhibit desired properties, for example a particular biological activity profile including selective binding to a single target or desired activity modulation of multiple targets simultaneously. This design certainly must include proper physicochemical as well as ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of the novel compounds. An attempt to create molecules from scratch is called “*de novo* design”.

There are several possible strategies that can be followed in order to succeed in this game. Important concepts and prominent examples of the actual design process will be presented and discussed. In any case, the molecule designer must become familiar with the basic building blocks and properties of “druglike” molecules (Chapter 1), and the principles of ligand–receptor interaction (Chapter 2).

### 1.1

#### What is a Molecule?

Irrespective of the chosen design approach, one must have a profound understanding of the structure–activity relationship (SAR) that guides the receptor–ligand interaction process. This in turn requires an adequate representation of molecular structures and their physicochemical properties to allow the extraction of molecular features that are responsible for a certain compound property or pharmacological behavior. Ideally, we need to understand the various molecular interactions of a particular molecule in different environments over time, which can be addressed by molecular dynamics.

Consequent physical treatment of molecule dynamics can, in principle, be achieved based on solutions of the Schrödinger equation (Eq. 1.1):

$$\hat{H}\Psi = E\Psi \quad (1.1)$$

where  $\hat{H}$  is the Hamilton operator defining the operations that need to be performed with the set of wavefunctions  $\Psi$  (psi) of the electrons of a molecular system, and  $E$  is the system’s potential energy;  $\Psi^2$  represents the electron density.

Each  $\Psi$  and its corresponding energy relate to a single electron. They are called “one-electron orbitals” denoted as 1s, 2s, 2p, etc.

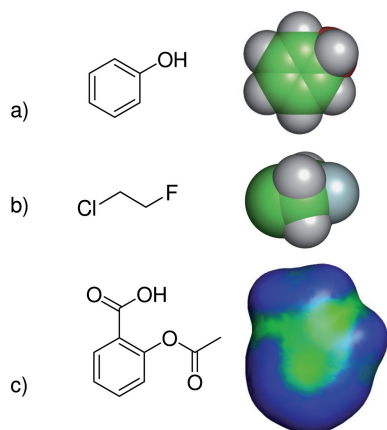
The Schrödinger equation provides a theoretical foundation for *ab initio* quantum chemical and quantum mechanical (QM) calculations. The term “*ab initio*” indicates that no empirical values are required and a solid physical and mathematical framework can be applied. QM calculations represent the formally most accurate way of calculating energies of molecular systems, allowing an assessment of conformational stability, chemical reactivity, etc. The problem is, however, that exact solutions of the Schrödinger equation cannot be obtained for molecules that are more complex than  $\text{H}_2^+$ . For other molecules, approximations are required. For example, the Born–Oppenheimer approximation treats atom nuclei as fixed, and only the movement of electrons is taken into consideration. A further approximation is the Hartree–Fock method that is grounded on solving the Schrödinger equation for each electron of the molecular system individually, thereby leading to single-electron wavefunctions (orbitals). Finally, semi-empirical approximations lead to the Hückel theory of molecular orbitals. Molecular orbital (MO) techniques result in a number of important molecular properties, for example partial atomic charges and the electrostatic potential, which can be used to describe molecule–molecule interactions.

QM methods are precise but rather slow, meaning they are computationally too expensive to use for calculating the potential energy of “large systems” containing more than approximately 100 atoms. However, “hybrid potentials”, that is, combinations of methods that treat different parts of a system at different levels of precision, permit QM calculations even for larger molecules.

## 1.2 Simplistic Molecular Representations

For drug design purposes, simplification and abstraction from the rigorous physically motivated conception of molecules goes even further. One reason is the still comparatively long computing time needed to treat a single molecule. Long computation times are acceptable if a calculation has to be done only once or for single or only a few molecules. In early phases of drug design, many thousands or millions of molecules need to be analyzed and “virtually screened” – here QC/QM methods do not find their application domain (yet). Instead, appreciably simpler molecular models and representations are employed.

A drastic simplification is in fact to neglect time-dependent behavior: Typically, molecules are treated as static two-dimensional (2D) molecular graphs or as three-dimensional (3D) space-filling rigid bodies with a defined surface (Fig. 1.1). While such models facilitate rational molecular design it is important to keep in mind that they represent only crude approximations of the “true nature” of molecules. This is one reason why an appropriate molecule description (choice of molecular “descriptors”) is essential for successful SAR modeling. As a consequence, the particular molecular representation that allowed for success-



**Fig. 1.1** Simplified representations of molecules. Left: two-dimensional (2D) sketch with “implicit” (suppressed) hydrogen atoms. Right: surface representations; (a) and (b) show three-dimensional (3D) van-der-Waals surface models with “explicit” hydrogens (CPK model according to Corey, Pauling and Koltun). In the CPK models, carbon and chlorine atoms are shown in green, hydrogens

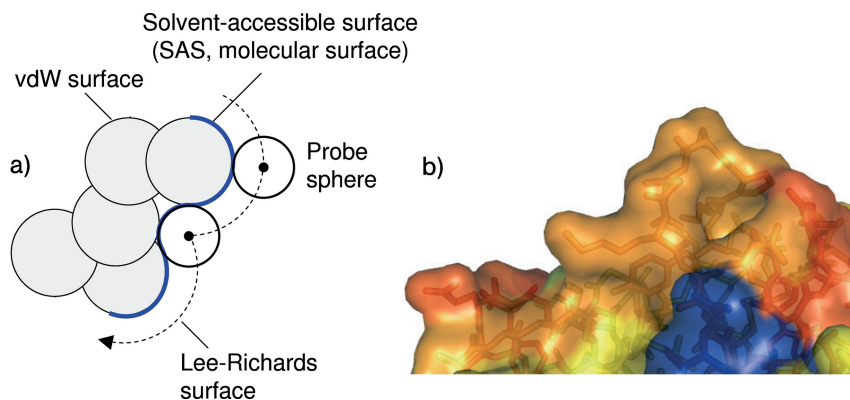
in grey, and the fluorine atom in turquoise. In (c) a surface representation of acetylsalicylic acid (trade name Aspirin®) is depicted, where the coloring indicates the curvature of the surface (blue: convex, green: concave). This surface model was calculated using Gaussian curve fitting and results in an analytically treatable surface model. Such surface models are used for comparison of molecular shapes.

ful SAR modeling in one drug design project is not necessarily generally applicable. Rather it should always be considered as a *context-dependent* model with a local validity domain only.

### 1.3

#### The Molecular Surface

QM teaches us that atoms and molecules possess a “soft” hull based on electron density. A useful – though not strictly accurate – simplification for molecular design purposes is to treat atoms and molecules as objects with a “hard” surface. Important concepts are molecular surfaces and surface properties, as molecules are thought to interact with each other via their surfaces. Calculation of surface properties, in particular the electrostatic surface potential, has received broad attention by computational chemists. Surfaces define an “inside” and an “outside” of a molecule and greatly facilitate modeling of molecular objects as 3D bodies with a finite shape and volume. Grounded on the work of Lee and Richards (1971) the “Solvent Accessible Surface” (SAS) is the most frequently used surface representation in drug design (Fig. 1.2). This is related to the aqueous environment in which pharmacokinetic and pharmacodynamic processes like ligand–receptor interaction, membrane permeation or drug metabolism take



**Fig. 1.2** Definition of molecular surfaces. For calculation of the solvent-accessible molecular surface (SAS) several concepts exist. According to Richmond (1984), the SAS can be obtained simply through an extended vdW surface. For most current small-molecule design applications Connolly's definition is employed. The Connolly algorithm uses a virtual solvent molecule represented as a

probe sphere that is rolled over the van-der-Waals (vdW) surface of the molecule. Usually, the radius of the sphere is chosen to be 1.4 Å, which is the effective radius of a water molecule. The resulting trace defines the SAS, which consists of parts of the vdW surface and the "smoothing" trace of the probe sphere (a). In (b) parts of the SAS of a protein structure are shown.

place. Molecular representations and the choice of appropriate descriptors must take this into account. The Connolly algorithm is the most frequently used method for SAS calculation (Fig. 1.2a). It leads to a smoothed representation of the van der Waals (vdW) surface which can be obtained by simply representing each atom by a ball with a fixed radius, its vdW radius (Table 1.1). These radii

**Table 1.1** Van der Waals radii of the most abundant atoms in druglike molecules.

Element symbol	vdW radius/Å	Approximate abundance in drug molecules (%) <sup>a)</sup>
H	1.20	
C	1.85	37.2
O	1.40	7.6
N	1.54	4.8
P	1.90	0.7
S	1.85	0.6
F	1.35	0.5
Cl	1.81	0.5

a) Values were calculated from molecules contained in DrugBank v. 10/2005 (URL: <http://redpoll.pharmacy.ualberta.ca/drugbank/>).

**Table 1.2** Average covalent bond lengths and their average dissociation energy ("bond energy"). Note that the energy values may differ depending on the substitution pattern.

Bond type	Bond length/Å	Energy/kJ mol <sup>-1</sup>
C–C	1.54	348
C=C	1.34	614
C≡C	1.20	839
C–H (sp <sup>3</sup> -H)	1.11	415
N–N	1.46	170
C–N	1.47	293
SO <sub>2</sub> –N	1.68	308
C–S	1.82	272
C–O	1.43	358
C=O	1.23	799
C(sp <sup>3</sup> )–H	1.09	413
N(sp <sup>3</sup> )–H	1.01	391
O–H	0.96	366
C–F	1.35	485
C–Cl	1.77	328
C–Br	1.94	276
C–I	2.14	240

can be experimentally determined from the distances of nonbonded atoms in crystal lattices.

Average bond lengths and strengths of prominent covalent bonds in druglike molecules are listed in Table 1.2. A schematic of two atoms approaching each other and the corresponding potential energy (Morse potential) of covalent bond formation is shown in Fig. 1.3.

Most drug–receptor interactions are reversible and dominated by non-covalent interactions. We will discuss non-covalent interactions in detail later, as the rational design of non-covalent interaction patterns is one of the most important parts in molecular design. Note that among the molecular design community atomic distances are usually given in Ångström units ( $1 \text{ Å} = 10^{-10} \text{ m} = 0.1 \text{ nm} = 100 \text{ pm}$ ).

Selected bond angles are given in Table 1.3. Such data provide the basis for molecular modeling and the atom-by-atom construction of new molecules. They are contained and tabulated in respective software packages, some of which allow for manual editing which can be essential if non-standard atom or bond types occur in a molecule. Prominent examples of molecular modeling packages that contain methods for molecular dynamics simulations and are often used as a reference for atom type definitions include AMBER ("Assisted Model Building with Energy Refinement", by P. Kollman et al.), SYBYL (Tripos Inc.), and CHARMM ("Chemistry at HARvard Macromolecular Mechanics", by M. Karplus et al.).

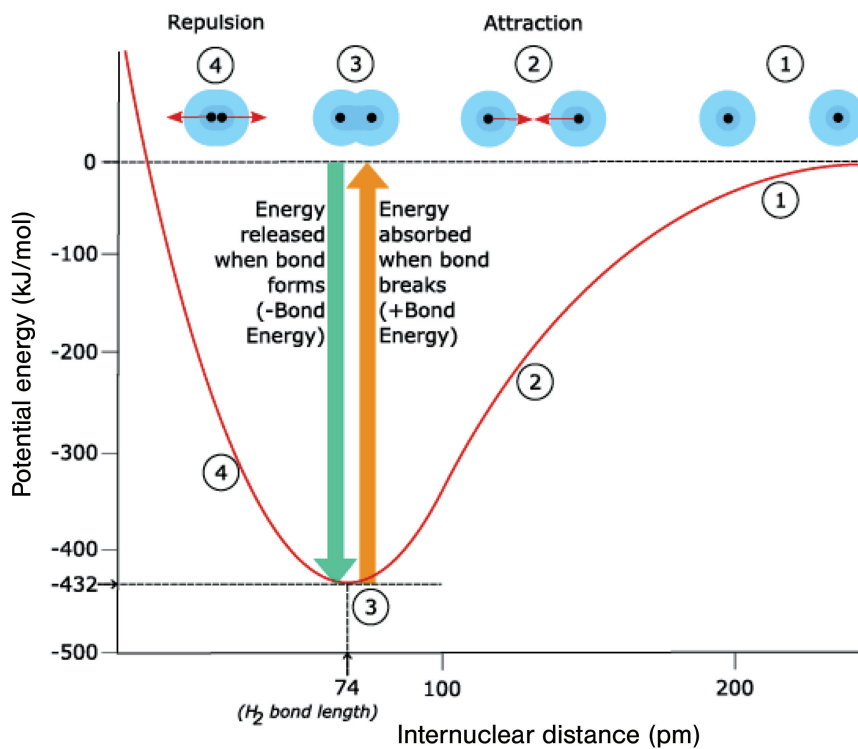
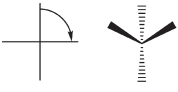


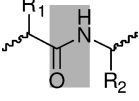
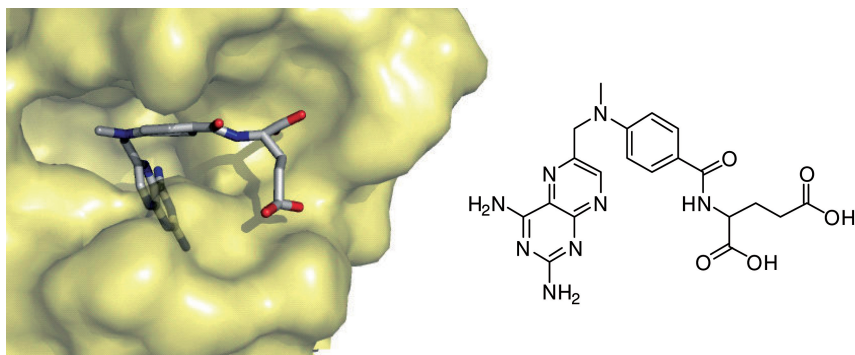


Fig. 1.3 Covalent bond formation between two hydrogen atoms (original figure by Prof. C. Chieh, adapted with kind permission).

Table 1.3 Bond geometry.

Bond type	Bond angle
 ( $sp^3$ )	109.5°
 ( $sp^2$ )	120°
 ( $sp$ )	180°
 $R_1$ $O$ $H$ $N$ $R_2$	CCN: 122° CNC: 116° peptide bond: planar



**Fig. 1.4** Cocystal structure of DHFR from the bacterium *Lactobacillus casei* complexed with Methotrexate (PDB identifier 3DFR) at a resolution of 1.7 Å. Parts of the binding pocket are shown with the protein surface. The inhibitor is drawn as a stick model. The 2D structure of Methotrexate is shown on the right.

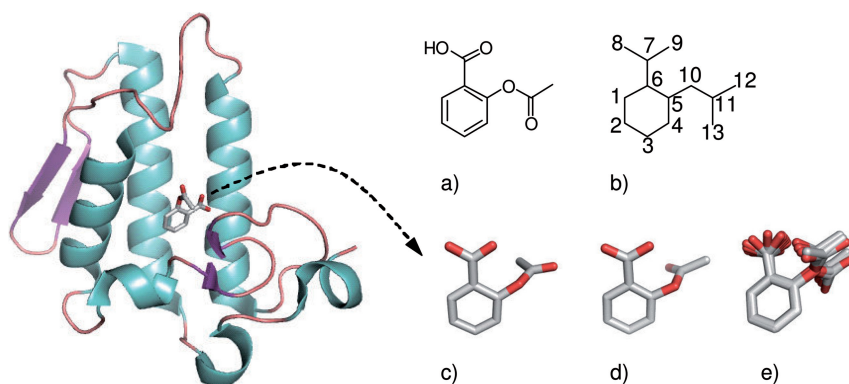
## 1.4 Molecular Shape

One important prerequisite for tight ligand–receptor interaction – and consequently ligand design – is *shape complementarity*. Surface representations of both the receptor and the ligand thus reflect a level of abstraction from the atomic structure that plays a pivotal role in molecular design. Molecules interact with each other via their surfaces and their surface properties. Figure 1.4 shows the Connolly surface of the active site of the enzyme dihydrofolate reductase (DHFR) together with the inhibitor Methotrexate. Methotrexate’s anti-tumor activity is a result of DHFR inhibition, as DHFR inhibition leads to inhibition of DNA synthesis and hence to the arrest of cellular replication. The active site of the enzyme is filled by the inhibitor, Methotrexate fits perfectly into the binding pocket. As with many drugs, this receptor–ligand interaction is probably not strictly selective, since Methotrexate exhibits other pharmacological effects, for example activity against rheumatoid arthritis via an unknown mechanism. Shape complementarity between a ligand and a binding pocket is a necessary though not sufficient attribute of a selective interaction. Proteins that have similar biological function often have similar binding pockets or active sites (in the case of enzymatic function). This similarity can be modeled by the analysis of complementary surface patches of known receptor–ligand pairs. The most widely used public sources of information about such interactions are the Protein Data Base (PDB) and Relibase.

## 1.5

## The Topological Molecular Graph

A molecule can be represented in the form of a graph drawing. The molecular graph and atom types represent principal concepts for empirical molecule representation and molecule design. Atoms are the vertices of the graph, and chemical bonds are the edges (Fig. 1.5). From the molecular graph matrix representations of molecular structure and topology can be derived. The connection table is a squared matrix that contains information about the type of bond connecting atoms that are neighbors in the graph representation. It is completely defined by the hydrogen-depleted, topological molecular graph. From this table many secondary matrices can be derived, which provide relevant information for automatic molecule design. Two such matrices are particularly important for molecule design, the adjacency matrix  $\mathbf{M}$  and the distance matrix  $\mathbf{D}$ . Both are squared symmetric matrices. The adjacency matrix contains the value 1 at positions representing adjacent nodes in the molecular graph, and 0 at all other positions. The *topological* distance matrix contains the numbers of bonds along the shortest path connecting the two respective atoms that define a position in the matrix. A distance matrix can also be used to represent a 3D conformation of a molecule; only this time *topographical* (spatial) distance values are given in Ångström units (Å). Such matrix representations of molecular features are ideally suited for computer processing. The adjacency matrix is defined by the constitution of a molecule, the distance matrix by its conformation. An example is given for the structure



**Fig. 1.5** (a) 2D-structure of acetylsalicylic acid (Aspirin®), a potent inhibitor of cyclooxygenases and phospholipase  $A_2$ ; (b) its molecular graph representation with atom numbers; (c) the phospholipase  $A_2$ -bound conformation (from PDB structure 1OXR at 1.93 Å resolution, shown as a secondary structure cartoon); (d) the conformation in the acetylsalicylic acid crystal

(from Cambridge Structure Database, CSD); (e) multiple superimposed conformations obtained from a semiempirical calculation. Note that none of these calculated conformations correspond to the annotated protein-bound conformation. Keep in mind that atom type assignment on the basis of an electron density map obtained by X-ray crystallography can be ambiguous and result in errors.

of acetylsalicylic acid (Fig. 1.5). Canonical atom numbers (Fig. 1.5b) are obtained by sophisticated algorithms, many of which are based on the Morgan algorithm (see cheminformatics texts for details).

The **connection table** of acetylsalicylic acid.<sup>a)</sup>

C	1	2	3	4	5	6	7	8	9	10	11	12	13
1	C	2	0	0	0	1	0	0	0	0	0	0	0
2		C	1	0	0	0	0	0	0	0	0	0	0
3			C	2	0	0	0	0	0	0	0	0	0
4				C	1	0	0	0	0	0	0	0	0
5					C	2	0	0	0	1	0	0	0
6						C	1	0	0	0	0	0	0
7							C	1	2	0	0	0	0
8								O	0	0	0	0	0
9									O	0	0	0	0
10										O	1	0	0
11											C	1	2
12												C	0
13													O

a) Contains element symbols on the diagonal; bond order indices on the off-diagonals (0: no bond, 1: single bond, 2: double bond)

The **adjacency matrix** of acetylsalicylic acid.<sup>a)</sup>

M	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	0	0	0	1	0	0	0	0	0	0	0
2		0	1	0	0	0	0	0	0	0	0	0	0
3			0	1	0	0	0	0	0	0	0	0	0
4				0	1	0	0	0	0	0	0	0	0
5					0	1	0	0	0	1	0	0	0
6						0	1	0	0	0	0	0	0
7							0	1	1	0	0	0	0
8								0	0	0	0	0	0
9									0	0	0	0	0
10										0	1	0	0
11											0	1	1
12												0	0
13													0

a) Contains only values of 0 (non-adjacent atoms) and 1 (adjacent atoms)

The **topological distance matrix** (shortest path in bonds) of acetylsalicylic acid.

D <sup>2D</sup>	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	2	3	2	1	2	3	3	3	4	5	5
2		0	1	2	3	2	3	4	4	4	5	6	6
3			0	1	2	3	4	5	5	3	4	5	5
4				0	1	2	3	4	4	2	3	4	4
5					0	1	2	3	3	1	2	3	3
6						0	1	2	2	2	3	4	4
7							0	1	1	3	4	5	5
8								0	2	4	5	6	6
9									0	4	5	6	6
10										0	1	2	2
11											0	1	1
12												0	2
13													0

A **3D distance matrix** (distances in Å) of acetylsalicylic acid (cf. Fig. 1.5c).

D <sup>3D</sup>	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1.38	2.40	2.78	2.42	1.40	2.49	2.85	3.59	3.66	4.80	4.94	5.93
2		0	1.39	2.40	2.78	2.40	3.75	4.23	4.73	4.13	5.26	5.39	6.36
3			0	1.38	2.40	2.77	4.25	5.06	4.97	3.64	4.70	4.97	5.70
4				0	1.39	2.41	3.76	4.86	4.18	2.38	3.45	3.96	4.37
5					0	1.40	2.49	3.71	2.79	1.36	2.58	3.22	3.60
6						0	1.47	2.44	2.33	2.39	3.62	3.84	4.62
7							0	1.35	1.21	2.83	3.73	4.03	4.69
8								0	2.22	4.17	4.98	5.06	5.93
9									0	2.50	3.18	3.65	3.93
10										0	1.48	2.62	2.31
11											0	1.51	1.22
12												0	2.36
13													0

The Structure Data Format (SDF) was developed by MDL Information Systems and represents a text-based file format that allows standardized exchange of molecular structure information. Two- and three-dimensional atom coordinates, the connection table, and molecular property information can be included. The SDF of the 2D structure of Aspirin is shown in Fig. 1.6 (note that in this ex-

First line: compound name (optional)  
 Aspirin  
 ChemDraw06050618212D

Element symbol

Number of atoms → 13 13 0 0 0 0 0 0 0 0 0999 v2000

Number of bonds

Atom coordinates (here: 2D)

-1.7862	-0.2062	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-1.7862	-1.0313	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-1.0717	-1.4438	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.3572	-1.0313	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.3572	-0.2062	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-1.0717	0.2062	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-1.0717	1.0313	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-1.7862	1.4438	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.3572	1.4438	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.3572	0.2062	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0717	-0.2062	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.7862	0.2063	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0717	-1.0312	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Conection table  
 The first two numbers are atom numbers,  
 the third number gives the type of connecting bond

1	2	2	0
2	3	1	0
3	4	2	0
4	5	1	0
5	6	2	0
6	1	1	0
6	7	1	0
7	8	1	0
7	9	2	0
5	10	1	0
10	11	1	0
11	12	1	0
11	13	2	0

M END  
 \$\$\$

Fig. 1.6 The structure data format of the 2D structure of Aspirin.

ample the hydrogen atoms are suppressed. Such a molecule representation is referred to as an “implicit hydrogen” representation).

## 1.6

### Molecular Properties and Graph Invariants

Any given molecule may be drawn in several equivalent ways, e.g. rotated and translated on a piece of paper, as different mesomeric forms of aromatic systems, or as tautomers. The human eye can often see if two drawings represent the same molecule, that is, when two graphs are *isomorphic*. However, millions of molecules in drug design require a fast computational differentiation and identification of *molecular graphs*. For this, we must represent any given molecule in a unique form. Structure normalization is therefore essential prior to computational analysis.

Topological indices are instances of *invariant* graph properties (graph “invariants”) determined from the graph of a molecule. Basic invariants are, for exam-

ple, the number of vertices and the number of edges. They have received wide attention in molecular modeling and design. Many topological indices include or are grounded on the term (Eq. 1.2)

$$\sum_{i,j} x_i x_j \quad (1.2)$$

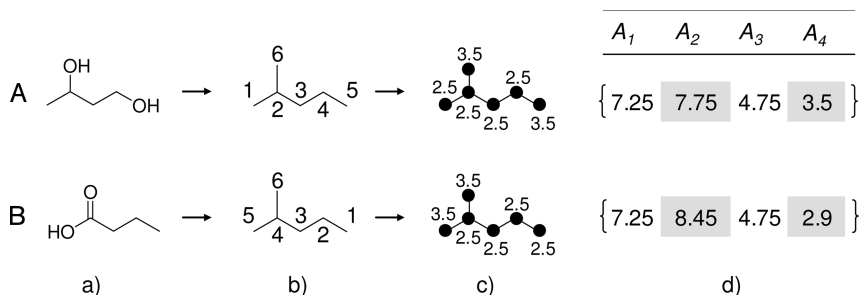
where  $i$  and  $j$  are vertices (atoms), and  $x$  is some vertex property, for example a physicochemical property or atom type (*vide infra*). Topological autocorrelation is such a descriptor (Eq. 1.3). The autocorrelation concept was introduced to the field of molecular modeling and design by Moreau and Broto (1980) almost 30 years ago. Several molecular descriptors employ this idea, for example the GRIND descriptors for 3D autocorrelation by Pastor and Cruciani (2000). We will discuss additional applications of correlation vector coding of molecular features in Chapter 2.

$$A_d = \frac{1}{L} \sum_{\left\{ \begin{array}{c} \{i,j\} \\ \text{for which} \\ D_{i,j}=d \end{array} \right\}} x_i x_j \quad (1.3)$$

where  $A_d$  is a number that expresses the correlation of a property  $x$  at a distance  $d$ , scaled by the number of non-hydrogen atoms  $L$ . There are other scaling options, for example division by the number of summations. In this case,  $A_d$  becomes the average autocorrelation  $\langle A_d \rangle$ .

Calculation of  $A_d$  is grounded on the distance matrix  $\mathbf{D}$ . This index describes the distribution of a property over the molecular graph or a 3D conformation if the spatial autocorrelation index is calculated. Autocorrelation is one way to denote a molecule in a compact form as a numerical vector, for example as  $\mathbf{A} = \{A_1, A_2, A_3, A_4\}$  which contains the autocorrelation values obtained for atom pairs spaced one ( $d=1$ ) to four bonds ( $d=4$ ) apart (Fig. 1.7). This particular molecular representation has a further desirable property: it is independent of the orientation of the molecule – in other words, it represents a *graph invariant*. This property provides an important concept for any comparison of molecules: if we want to compare molecules that do not share a common structure or if we are not able to find a meaningful alignment of the two structures, we can employ *alignment-free descriptors* like topological autocorrelation.

This is certainly of high interest because there are many ways to define an alignment of molecules, for example by structure, shape or property distribution. Different alignment techniques most often result in varying superpositioning of molecules. Obviously, a single “correct” alignment of molecules does not exist. Alignment-free descriptors circumvent this issue. However, they do not allow for an unambiguous conversion of the descriptor vector back to the original molecule; they are *non-bijective*.



**Fig. 1.7** Topological autocorrelation vector representation of two molecules A and B according to Eq. (1.3). The molecular structure (a) is converted to the molecular graph with atom numbers indicated (b), and a vertex property is assigned (c). Here, electronegativity values were considered as a property of the vertices in the molecular graph (values according to Allred-Rochow, note: for molec-

ular electrostatics calculations the electronegativity values according to R. S. Mulliken are typically employed). Despite identical structure graphs (b) of the compounds their resulting autocorrelation vectors differ for the distances of two ( $A_2$ ) and four ( $A_4$ ) bonds (highlighted values) (d). The values of  $A_d$  are scaled by the number of non-hydrogen atoms ( $L = 6$ ).

The concept of substructure elements as generic molecular building blocks can be used as a basis for calculation of molecular properties like “topological polar surface area” (TPSA) or the n-octanol–water partition coefficient  $P$  (Eq. 1.4). The logarithm of the latter,  $\log P$ , is often referred to as the “lipophilicity” or “hydrophobicity” of a molecule. It can be computed using a large look-up table of predefined molecular fragments and their experimentally measured  $\log P$  contribution to the lipophilicity of a molecule. Currently, approximately 5000 such data records are available. The calculated  $\log P$  estimation, “clog $P$ ”, employs a weighted sum of the individual fragment contributions present in a given molecular structure and associated correction factors accounting for potential interactions between the fragments, which result in modified fragment contributions to the total lipophilicity of the molecule (Eq. 1.5).

$$\log P = \log \frac{[C]_{\text{org}}}{[C]_{\text{aq}}} = \log [C]_{\text{org}} - \log [C]_{\text{aq}} \quad (1.4)$$

where  $[C]_{\text{org}}$  and  $[C]_{\text{aq}}$  are the concentrations of a compound in the organic and aqueous phases of a mixture, usually octanol and water.

$$\text{clog} P = \sum_{i=1}^n a_i f_i + \sum_{j=1}^m b_j F_j \quad (1.5)$$

where there are  $a_i$  fragments of type  $i$  with their lipophilicity contribution  $f_i$ , and  $b_j$  occurrences of correction factor  $j$  with  $F_j$  being the correction factor.

Several variations of this scheme exist, and different software implementations vary in their accuracy. For molecules containing known fragments and their contribution factors the accuracy of such  $\text{clog}P$  methods is remarkably high (squared correlation coefficient  $R^2 > 0.97$ ). However, these methods are not perfect and likely fail for novel molecules.

Until today, far more than one thousand QM and empirical descriptors have been devised over the past decades, approximating four main aspects of molecular structure and molecular recognition:

- molecular distribution
- molecular shape
- directed interactions
- non-directed interactions

## 1.7

### The Drug-likeness Concept

The term “druglike” describes various empirically found features of molecular agents that are associated with pharmacological activity. It is not strictly defined but provides a general concept of what makes a drug a drug. Drug-likeness may be considered as a complex property representing various physicochemical and structural features. Several drug-likeness indices have been proposed, most of which are grounded on two-dimensional molecular structure and calculated properties. We will present several such indices throughout this text. They can be used as guidelines for the rational design of lead structures that can be further optimized to become a drug. Typically, one considers the following basic attributes for the construction of such a *structure–property relationship* (SPR):

- substructure elements
- lipophilicity (hydrophobicity)
- electronic distribution
- hydrogen-bonding characteristics
- molecule size and flexibility
- pharmacophore features

*Lipinski’s rule-of-five* (sometimes called “Pfizer rules”, because its inventor, Lipinski, worked at the pharmaceutical company Pfizer) is probably the best known guideline that helps to raise awareness about properties and structural features that make molecules more or less drug-like. In pharmaceutical drug discovery it can help to avoid costly late-stage preclinical and clinical failures. The guidelines predict that poor passive absorption or permeation of an orally administered compound is more likely if the compound meets at least two of the following criteria:

- molecular weight greater than 500 Da
- high lipophilicity (expressed as  $\text{clog}P > 5$ )

- more than 5 hydrogen-bond donors (expressed as the sum of OHs and NHs)
- more than 10 hydrogen-bond acceptors (expressed as the sum of Os and Ns)

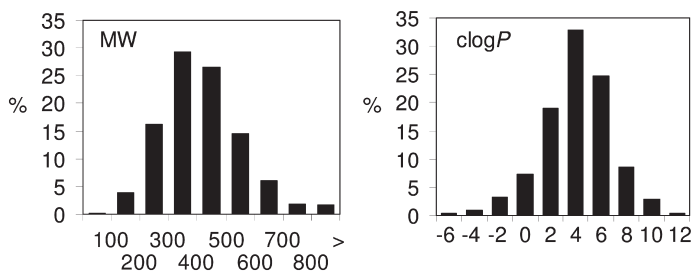
Figure 1.8 shows the distributions of molecular weight and lipophilicity of 4500 druglike molecules. From these histograms it becomes evident that interpreting Lipinski's "rules" as strict threshold values is not appropriate. They represent empirically found guidelines for the design of orally bioavailable druglike molecules. Several additional guidelines have been proposed addressing different attributes of drugs. For example, a druglike molecule *should not*

- contain more than 5 rotatable bonds to limit its conformational freedom (this recommendation is sometimes referred to as the fifth Lipinski rule);
- possess a polar surface area exceeding  $120 \text{ \AA}^2$  to avoid potential bioavailability problems of a compound;
- have predicted aqueous solubility ( $\log S$ ) below  $-4$  (solubility in mol/l).

These criteria are particularly suited for assessing averaged property values of sets of molecules ("compound libraries"), rather than individual agents. For a particular indication however, the receptor tissue location (for example, in the brain or periphery) and other criteria must be taken into account, leading to adjusted (target-related) rules.

With the introduction of high throughput screening (HTS) in the 1970s and combinatorial chemistry in the early 1990s, a new era in drug discovery emerged. The synthesis and biological testing of thousands of compounds at comparatively low costs became feasible. Despite their appeal these techniques have had surprisingly small impact on the derivation of novel drugs and druglike candidates for lead optimization. This low impact is presumably caused by limited structural diversity and lack of "pharmacological relevance" of the underlying structures of many of the combinatorial and screening libraries.

In contrast, natural products represent the richest source of inspiration for the identification of novel scaffold structures that can serve as the basis for rational drug design. Among the FDA-approved New Chemical Entities (NCEs) that were



**Fig. 1.8** Distribution of calculated molecular weight (MW) and lipophilicity ( $\text{clog}P$ ), expressed as a logarithm of the calculated octanol/water partition coefficient  $P$ , for a collection of 4500 selected druglike molecules containing marketed drugs and drug candidates.

introduced between 1981 and 2002, 49% were of natural product origin or were derived from natural products. As Costantino and Barlocco (2006) stated, “... *natural products are considered to contain scaffolds with the potentiality to be privileged structures because in many cases they are synthesized by biological systems to specifically interact with protein targets.*” Natural products could thus serve as biologically validated starting points for combinatorial variation, and combinatorial libraries built around these “privileged structures” should facilitate hit and lead finding.

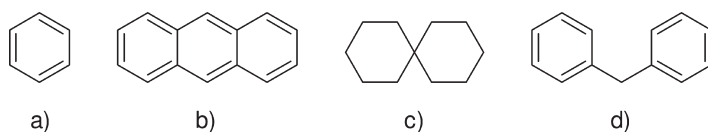
## 1.8

### Scaffolds, Linkers, and Side-chains

Generally, systematic investigations of molecular scaffolds are used as a way of measuring the diversity of a compound library. Computer-based analysis revealed that natural products exhibit a remarkable structural diversity of molecular frameworks and scaffolds that could be systematically exploited for combinatorial synthesis. Natural products offer a rich pool of unique molecular frameworks that complement the known “drug space”. They often possess desirable druglike properties, rendering them ideal starting points for molecular design considerations. Certainly, not all natural product-derived scaffolds will be directly accessible to synthesis due to their inherent structural complexity or substitution pattern. Still, systematic scaffold analysis can provide medicinal chemists with ideas about chemotype variations.

In 1996 Bemis and Murcko introduced a formalistic approach for the dissection of molecules into “side-chains” and “frameworks”. According to this concept, a molecule can be segmented into four units: ring systems, linkers, side-chains, and frameworks (Fig. 1.9). Formally, we can define these structural units:

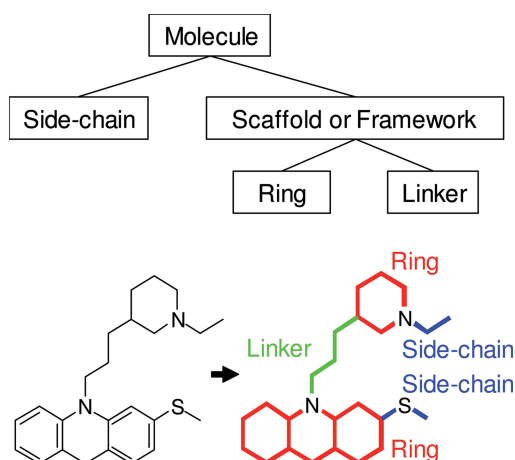
- *Ring systems* are cycles within the molecular graph (rings) or rings sharing an edge (a bond between two atoms) or vertex (atom) in the molecular graph. A set of rings with fused or spiro connections is defined as a single-ring system. For example, benzene, anthracene and spiro[5.5]undecane (Fig. 1.9) are single-ring systems; benzene is a monocyclic, anthracene a tricyclic, and spiro[5.5]undecane a bicyclic ring system.
- *Linkers* are defined as vertices (atoms) and/or edges (bonds) on the path connecting two different ring systems. Diphenylmethane (Fig. 1.9d) contains a one-atom linker between two different ring systems.
- *Side-chains* are those atoms that are not classified as a ring system or linker atoms.
- *Frameworks* are defined as ring systems (if no linker exists) and ring systems connected by linkers, that is, everything that remains after removing the side-chains. Acyclic molecules do not have any framework, by definition. The framework of the molecule shown in Fig. 1.10 has two different ring systems (one monocyclic and one tricyclic) connected by a three-atom linker.



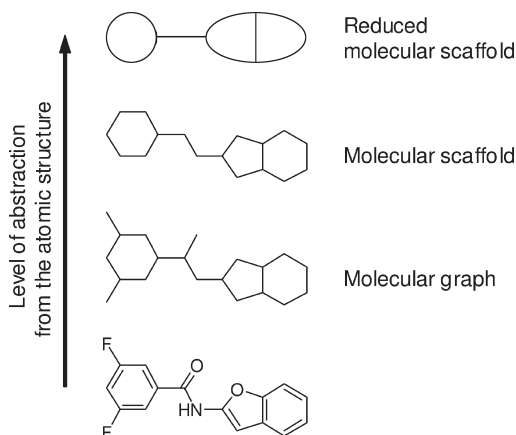
**Fig. 1.9** Benzene (a), anthracene (b), and spiro[5.5]undecane (c) are three single-ring systems, diphenylmethane (d) contains two ring systems and a one-atom linker.

Recently, Xu and Johnson introduced the concept of the “reduced molecular scaffold” (Fig. 1.11). At this level of abstraction from the atomic molecular structure only the numbers of ring systems and the existence of linkers are considered. The reduced scaffold representation is suited for the comparison of different chemotypes present in a compound collection.

Manipulating living systems at the molecular level requires profound knowledge of the variability of small molecule effectors that provoke a particular cellular response. Medicinal chemistry therefore relies on libraries of molecular probes that can be rationally designed to contain a desired degree of chemotype diversity. Despite great advances in the field of virtual screening and rational compound library design, “scaffold-hopping” remains a challenging goal. The concept of scaffold-hopping aims at finding molecules that possess different scaffolds but exhibit identical or very similar pharmacological activity. Ideal screening methods that perform successful scaffold-hops would not only find a maximum



**Fig. 1.10** Molecules can be formally divided into scaffolds or frameworks, linkers, and side-chains which are attached to the ring systems. Automated framework analyses following this concept were pioneered by Bemis and Murcko at Vertex Pharmaceuticals.

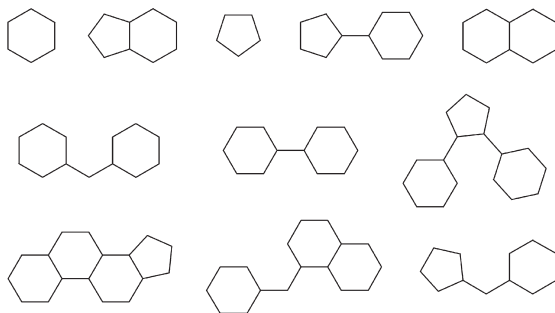


**Fig. 1.11** Levels of abstraction from the two-dimensional atomic structure of a molecule.

number but also a maximally diverse set of active compounds from a given chemical subspace (cf. Chapter 4).

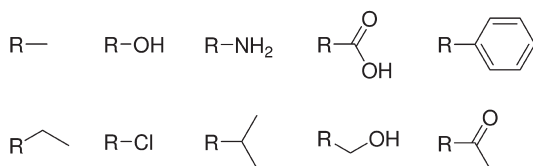
There are several reasons for seeking a set of diverse scaffolds. Different chemotypes offer a choice in terms of chemical accessibility and prospects for lead optimization. Multiple lead structures for a particular biological target lower the chance of drug development attrition in the case of undesirable ADMET properties. Scaffold-hopping can also be applied to move from natural substrates to more druglike or synthetically tractable chemotypes. Furthermore, the creation of new intellectual property is facilitated when multiple novel bioactive agents are available.

In Fig. 1.12 examples of prominent molecular drug scaffolds are shown. It is evident that the aromatic six-membered ring represents the dominant building block of known synthetic drugs. Chemotype diversity is often obtained only by linker variation and different substitution patterns.

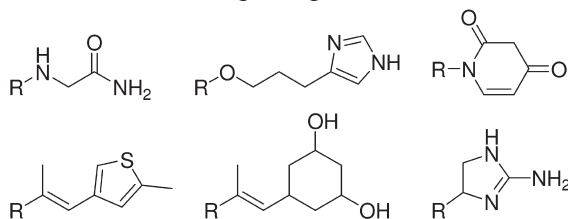


**Fig. 1.12** The ten most frequent molecular scaffolds ("frameworks") found in a collection of drugs and lead structures.

## The most common side-chains



## Side-chains with a high drug-likeness index



**Fig. 1.13** Examples of substituents of known pharmacologically active molecules from the World Drug Index (WDI), and selected examples of druglike substituents.

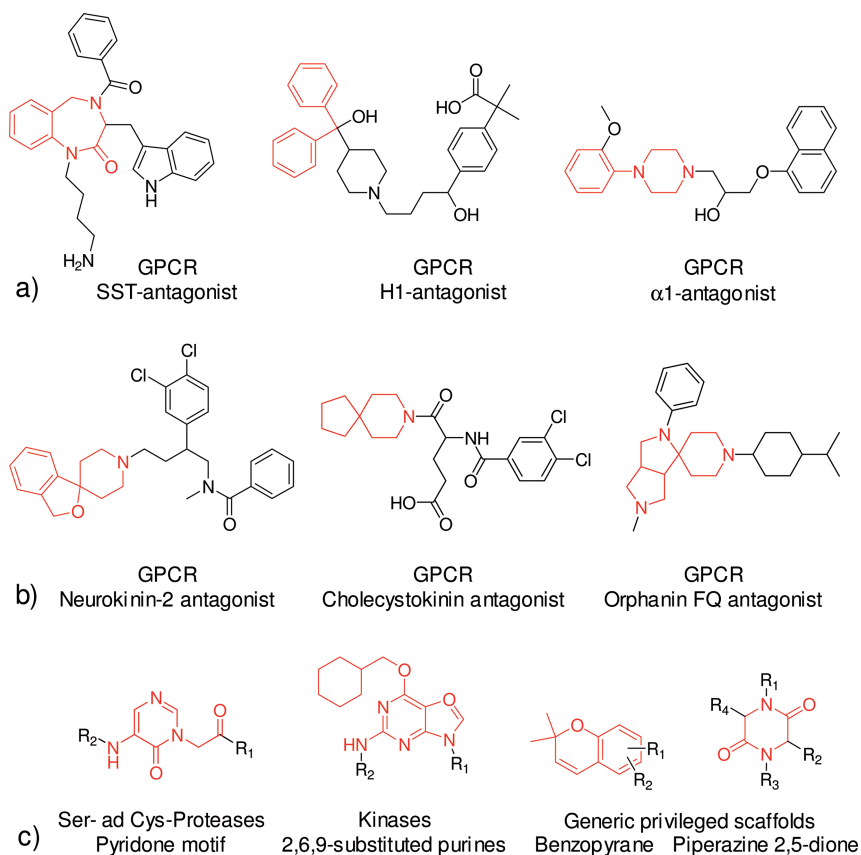
Ertl (2003) from Novartis analyzed the Derwent World Drug Index (WDI), a large collection of known pharmaceutically active agents, against frequently occurring substituents. Figure 1.13 shows examples of druglike substituents. Such substructures are not referred to as *privileged motifs*, as they are abundant in all kinds of drugs and usually not restricted to one receptor or receptor family. We will come back to this important topic later in this book when we discuss *de novo* design strategies (cf. Chapter 3).

## 1.9

## Substructure Similarity and “Privileged Motifs”

Molecules that exhibit similar pharmacological behavior, e.g. binding to the same receptor, often look similar. In fact, molecular design is grounded on the *similarity principle* found by Johnson and Maggiora (1990): “*Similar (structurally related) compounds exhibit similar biological activities*”. Please keep in mind that multiple exceptions to this rule are known.

It is sometimes possible to define mutual structural elements, *privileged motifs*, which facilitate ligand binding to a particular receptor or receptor family. Figure 1.14 shows some of these motifs. For example, benzodiazepinones, arylpiperazines, and certain biphenyl motifs have been successfully employed for designing ligands that bind to G-protein coupled receptors (GPCR). The spiroperidine motif is an example of a target-family preferring motif, because it frequently oc-

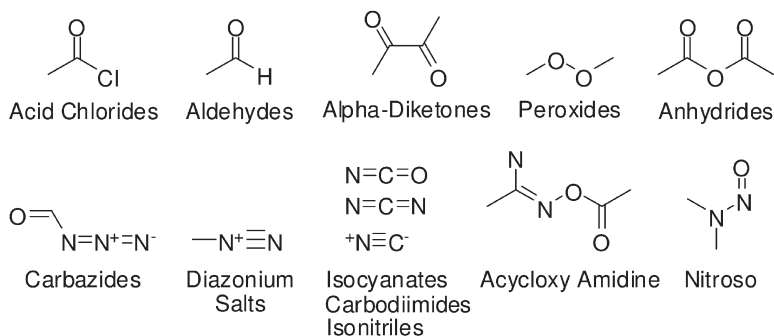


**Fig. 1.14** Examples of privileged substructure motifs for drug design. (a) GPCR-privileged motifs, (b) examples of GPCR ligands containing the spiropiperidine motif, (c) examples of protease and kinase inhibitor motifs, and two examples of generic privileged scaffolds

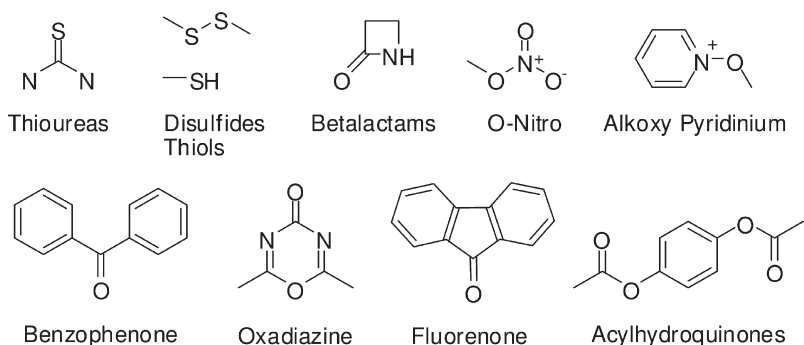
that have been shown to be present in many different biologically active molecules, including natural products. Keep in mind that the presence of a privileged motif in a molecule does not guarantee a desired binding behavior.

curs in GPCR inhibitors. Therefore, this motif might represent an attractive starting point for the design of ligands that bind to various GPCRs. Privileged motifs are also used as core structures (scaffolds) for combinatorial design and synthesis. Several companies offer specifically tailored molecular libraries around privileged motifs. Natural products provide useful molecular recognition motifs. For example, the benzopyran motif shown in Fig. 1.14c is present in more than 4000 substances including numerous natural products exhibiting diverse pharmacological activities. It can thus be considered a “generic druglike motif” that can be used as a scaffold for molecular design. It is important to keep in mind that a privileged motif alone does not necessarily induce a desired pharmacolog-

## Reactive groups



## Unsuitable groups

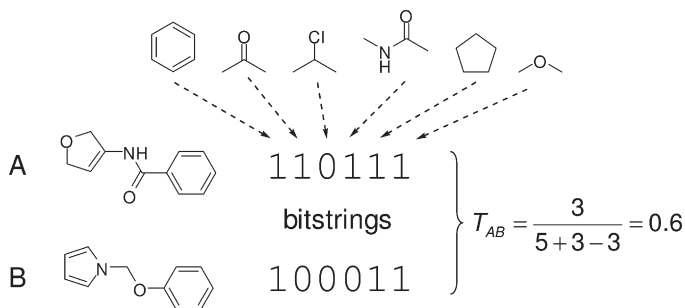


**Fig. 1.15** Examples of unwanted motifs. Typically, such substructure elements are avoided in drug design. Note: Thiols and betalactams are found in known drugs.

ical activity, and its presence does not unambiguously define the target receptor. It is the complete molecular structure, which defines the pharmacological behavior of a compound. A privileged motif is a *substructure* of a molecule, which itself represents the *superstructure*.

Substructure searching in chemical databases is often a first step in *virtual screening* with the aim being to find molecules that are structurally related to a query compound exhibiting a potentially similar activity to the query structure (cf. Chapter 4).

Figure 1.15 shows several substructure elements that are usually avoided in drug design, mostly reactive or unstable groups, or substructure elements that tend to lead to poor aqueous solubility of a compound (e.g., thiourea derivatives). Such lists of unwanted motifs are sometimes referred to as “flagging lists” or “black lists” by drug designers.



**Fig. 1.16** Hypothetical simplifying example of 2D molecular fingerprints and their Tanimoto similarity. Two molecules are encoded as bitstrings using a look-up table of predefined substructures. The Tanimoto similarity of the bitstrings is  $T_{AB}=0.6$  according to Eq. (1.6).

Substructure elements can be identified by automated similarity searching. Similarity searching can be achieved by comparing representations of molecules with respect to the substructure elements they contain. For this purpose molecules are represented as bitstrings (“2D fingerprints”, “molecular fingerprints”), where a bit indicates the presence (‘1’) or absence (‘0’) of a particular substructure (Fig. 1.16). Comparison of bitstrings can easily be performed by applying logic operators or similarity indices. A popular similarity index is the Tanimoto (Jaccard) coefficient (Eq. 1.6).

$$T_{AB} = \frac{c}{a + b - c} \quad (1.6)$$

where  $a$  is the number of bits set to 1 in molecule  $A$ ,  $b$  is the number of bits set to 1 in molecule  $B$ , and  $c$  is the number of set bits common to both  $A$  and  $B$ . The value of  $T$  is between zero and one, where a value of one indicates identical bitstrings, and a value of zero indicates maximal dissimilarity. Keep in mind that  $T_{AB}=1$  does not necessarily mean that the two molecules  $A$  and  $B$  are identical! The way the bitstrings are constructed can lead to different pair-wise similarity values, and the value of the Tanimoto coefficient (as well as of any other similarity measure) is only meaningful with regard to the type of molecular fingerprint representation used. The same is true for rules of thumb stating that a Tanimoto coefficient greater than a certain value (e.g., 0.85) indicates an identical chemotype in the two molecules compared. The meaning of a similarity value is strongly context-dependent.

There exist two principal ways to construct a 2D fingerprint: either by using a look-up table containing a set of predefined substructures (e.g., the so-called “MACCS keys”, or Ghose and Crippen fragments) or by exhaustive enumeration

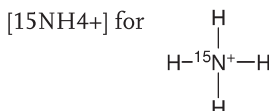
of all possible substructures of a molecule. The first approach has the advantage of comparably short fingerprints, where each position (each bit) corresponds to one substructure. The presence or absence of a substructure in a molecule can be immediately seen from the look-up table. The second concept (which often lacks the one-to-one matching of a bit position and a particular substructure if hashing is involved) leads to very long fingerprints with often more than 10 000 bits. Its advantage is its completeness: Since *all* substructures of a molecule are automatically encoded, the resulting similarity value obtained from the comparison of two such exhaustive fingerprints expresses the intuitive meaning of “structural similarity”.

## 1.10


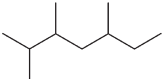
### Molecules as Strings

Many formats have been devised for representing molecular graphs as linear notations, for example, the InCHI representation developed by IUPAC, the Wiswesser Line Notation WLN – one of the earliest line notations –, or the Chemical Markup Language (CML) by Murray-Rust, which was designed for chemical information exchange over the Internet. String representations are ideally suited for storing, transmitting, and manipulating large numbers of molecules, which is a typical task in molecular design studies. The most frequently used linear molecule representation is probably the SMILES notation due to its comparative simplicity and readability. Weininger introduced the Simplified Molecular Input Line Entry System (SMILES) in 1988 as an intuitive method to annotate chemical structures in the form of compact text strings. Although SMILES are applicable to arbitrary chemical compounds, they primarily aim at small organic molecules. SMILES is a notation of the two-dimensional molecular graph.

By means of simple rules the graph representation can be expressed as a SMILES string: Atoms of the so-called “organic subset” (B, C, N, O, P, S, F, Cl, Br, I) are described by their element symbol. For all other atoms, square brackets (“[...]”) have to be used, e.g. [Na] for elemental sodium. Atoms, which are part of aromatic systems, are specified by lower case letters, e.g. ‘c’ for a carbon atom of the benzene ring. Hydrogen atoms are implicitly added to elements of the organic subset, therefore ‘C’ is a legal SMILES notation for methane (CH<sub>4</sub>), ‘O’ for water (H<sub>2</sub>O), ‘N’ for ammonia (NH<sub>3</sub>), and “CCO” for ethanol, just to give a few examples. Atom mass, chirality information, explicit hydrogens, and ionic charges can be optionally notated within the square brackets:

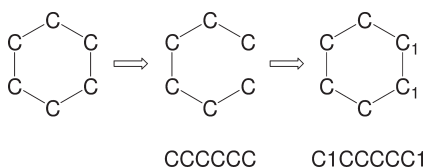


**Table 1.4** Molecular structures and their SMILES notation.

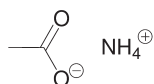
Structure	SMILES
	<chem>CCO</chem>
<chem>H2C=CH2</chem>	<chem>C=C</chem>
<chem>HC≡CH</chem>	<chem>C#C</chem>
	<chem>CC(C)C(C)CC(C)CC</chem>

Single bonds are represented by ‘-’, or may be omitted. Double, triple and aromatic bonds are denoted by ‘=’, ‘#’, and ‘:’, respectively. Table 1.4 shows some examples of bond symbol employment. For non-linear structures, it is necessary to have a possibility to indicate branches and rings. Round brackets (“(...)”) are used to indicate branching, recursive repetitions are allowed.

Cyclic structures are constructed in the following manner: An arbitrary bond in a ring has to be virtually broken, and a SMILES expression for the structure without rings has to be created. To indicate this virtual connection point, a number is used as a label for the two atoms which are connected by the broken bond:

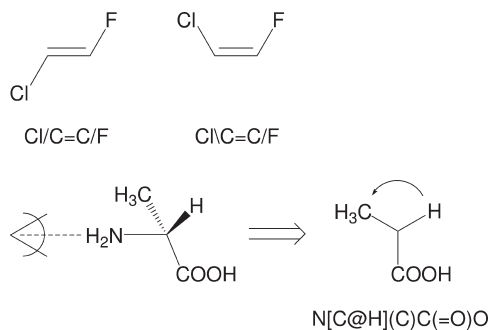


Disconnected structures, e.g. salts, are separated by a period:

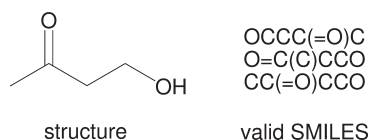


CC([O-])=O.[NH4+]

The SMILES language provides the possibility to annotate the configuration of double bonds and stereocenters. For double bonds, so-called “directional bonds” (symbols ‘/’ and ‘\’) are used. The absolute stereochemistry at chiral centers is defined by using the signs ‘@’ and “@@” to indicate anticlockwise or clockwise orientation of the substituents:

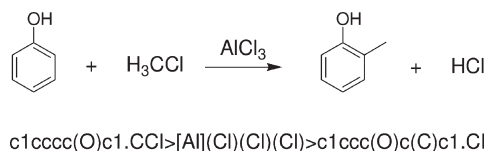


It is easy to see that SMILES are inconsistent, as there are different valid SMILES for the same molecular structure, for example:



To facilitate fast chemical database searching, Weininger (1988) proposed an algorithm which generates *unique* (or “canonical”) SMILES. For example, the canonical SMILES of Aspirin (Fig. 1.1c) is CC(=O)Oc1ccccc1C(=O)O. Further details about SMILES and unique SMILES can be found on the Daylight website ([www.daylight.com](http://www.daylight.com)).

SMILES also include a possibility to describe chemical reactions, the so-called *Reaction SMILES*. The reaction arrow is indicated by “ $\gg$ ”. If necessary, agents can be placed between the two ‘>’ symbols to indicate reaction conditions or leaving groups, for example:

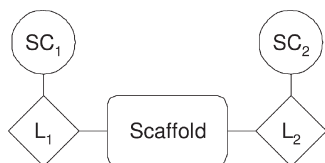


## 1.11

### Constructing Molecules from Strings

SMILES can be used for virtual molecule assembly from scaffolds, linkers, and side-chains by string concatenation. Side-chains are attached to the scaffold via linker groups (Fig. 1.17). This allows different chemical reactions to be considered implicitly by using different linker types and generic building block collections. It also simplifies virtual compound library enumeration since connecting functional groups may be completely left out in the set of building blocks. An

advantage of this conceptual idea of a combinatorial library is its simplicity and ease of implementation, yielding very short computing times. However, realistic chemical reactions cannot be modeled deliberately. For example, ring or scaffold formation during side-chain attachment cannot be modeled.

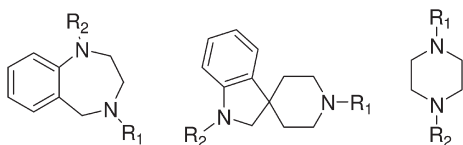


**Fig. 1.17** Schematic composition of a virtual reaction product.  
 $SC_i$  denotes a side-chain,  $L_j$  denotes a linker unit.

For such a virtual reaction all educts have to be formulated in an enhanced notation of SMILES: Special labels “[R1]”, “[R2]”, “[R3]”, etc., and “[A]” can be used to specify sites of variability (R) and attachment (A), respectively. The latter is necessary to enable directional concatenation of side-chains to linkers and linkers to scaffolds (Fig. 1.17). Examples typically look like this:

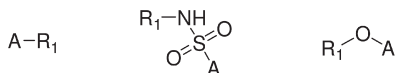
**Scaffolds:**

```
[R1]N2CCN([R2])C1=CC=CC=C1C2
[R2]N(C3=C2C=CC=C3)CC12CCN([R1])CC1
[R1]N1CCN([R2])CC1
```



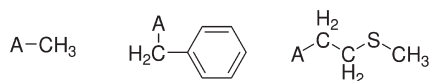
**Linkers:**

```
[A][R1]                (“pseudo linker” or “zero-bond” linker)
[A]S(=O)(N[R1])=O      (sulphonamide linker)
[A]O[R1]                (ether linker)
```

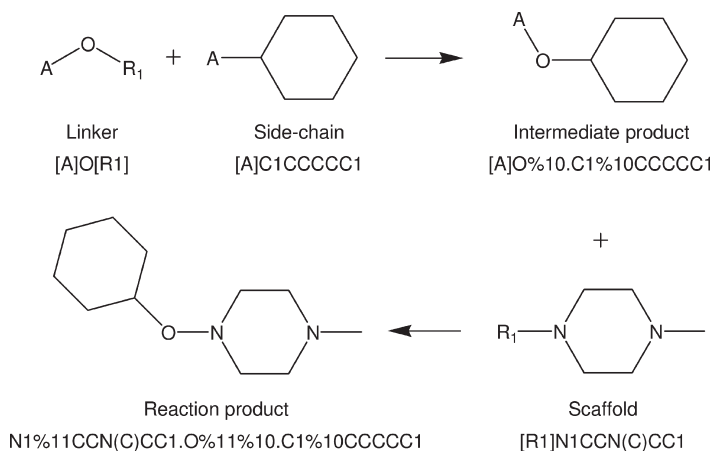


**Building blocks (side-chains):**

```
[A]C
[A]CC1=CC=CC=C1
S(CC[A])C
```



Reaction products are generated by concatenation of SMILES strings using “unsatisfied” ring closures. For example, according to the SMILES convention, either of the following notations for ethane is valid: “CC” or “C1.C1”. Following this scheme, one can use unsatisfied ring closures to form chemical bonds between the constituents of a reaction product, as shown in Fig. 1.18. The ‘%’ symbol denotes a specific building block. The resulting SMILES look unconventional – yet they are perfectly valid. This example shows how string representations of molecular building blocks provide a basis for the assembly of new chemical entities on the computer.



**Fig. 1.18** A virtual reaction with corresponding SMILES: The side-chain is connected to a linker having the side-chain’s A-group react with the linker’s R1-group forming intermediate product and a virtual A-R1 by-product (which is neglected for subsequent processing). The intermediate product’s A-group then undergoes reaction with the scaffold’s R1-group yielding the final reaction product and a second A-R1 by-product (also neglected).

## 1.12 From Elements to Atom Types

The usefulness of element symbols in molecular representations is limited. A major shortcoming in the context of molecular design is their lack of information about potential interactions that can be formed by the atoms of a molecule. The concept of “atom typing” provides a broader definition of an atom than just by

**Table 1.5** Examples of SMARTS expressions.

SMARTS	Meaning
*	any atom (so-called wildcard atom)
[+1]	all atoms with charge +1
[a]	all aromatic atoms
[r5]	all atoms in a five-membered ring
[#6]=:[#7]	carbon connected by a (double or aromatic) bond with a nitrogen
[\$(*O);\$(*CC)]	any atom that is connected to an aliphatic oxygen and to two sequential aliphatic carbons
[!c]~[#8,n]	an atom, that is not an aromatic carbon, which is connected by any bond to any oxygen or aromatic nitrogen

element. For many tasks in computer-assisted molecular design it is necessary to specify generic molecular substructures. For this purpose, SMARTS (SMILES Arbitrary Target Specification) was developed as an extension to the SMILES notation. SMARTS allows the intuitive specification of atom and bond properties that are required from a query (though there exist several such substructure query languages that could be used instead of SMARTS). Note that all SMILES are valid SMARTS expressions. Table 1.5 gives examples of SMARTS expressions. In SMARTS, atom and bond properties can be associated by logical operators: '!' for boolean NOT, ',' for OR, '&' for high precedence AND, ';' for low precedence AND. The so-called *recursive* SMARTS allow one to define the chemical environment of an atom.

There are many possibilities to define atom types according to different concepts. Most of them are grounded on the definition of *local atom environments*. A popular scheme is the SYBYL notation, which assigns a class type to each individual atom (Table 1.6). The identifiers can be used to represent atoms in SYBYL

**Table 1.6** Examples of atom types. "AND" denotes the logic (boolean) union.

Atom type	Description
C.3	sp <sup>3</sup> -hybridized carbon atom
C.2 (C.1)	sp <sup>2</sup> - AND sp-hybridized carbon atoms
C.ar	Aromatic carbon atom
C.cat	Carbon atom in positively charged groups (amidinium, guanidinium)
N.ar (N.2)	Aromatic nitrogen atom AND sp <sup>2</sup> -hybridized nitrogen atom
N.am	Nitrogen atom in amide bond
O.co2	Oxygen atom in carboxy groups
F	Fluorine atom
Met	Metal atom

Line Notation (SLN), a formal language for representation of molecules and reactions as strings, similar to SMILES and SMARTS. The idea of such class labels is to distinguish between an atom of the same element (C, N, O, P, S etc.) in different local environments, and to group atoms together. For example, carbon atoms in different hybridization states have different reactivity. Atom types represent an abstraction from pure molecular architecture and facilitate the calculation of molecular properties and SAR modeling.

### 1.13

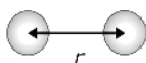
#### Entering the Third Dimension: Automatic Conformer Generation

Until now we have focused mainly on two-dimensional molecular representations. However, the molecular world is three-dimensional. Conformations are arrangements of the atoms of a molecule in space that can be converted by rotation(s) around a single bond, whereas configurations refer to those other arrangements whose interconversion requires bonds to be broken and then re-formed differently (e.g. *cis* and *trans* double bonds). There are two widely used principles of automated conformer generation from the molecular graph representation: *Force-field* methods relying on molecular mechanics, and *knowledge-based* heuristic approaches.

The basic idea of a molecular force-field is to treat a molecule as a mechanical object: Atoms are regular spheres that are connected by springs, and their pairwise interactions are expressed by terms of a potential energy function (Eq. 1.7). One further differentiates between *bonded* interactions and *non-bonded* intramolecular interactions, which leads to the general formulation of a molecular force-field equation:

$$\begin{aligned} \text{Energy} = & \text{Stretching Energy} + \text{Bending Energy} + \text{Torsion Energy} \\ & + \text{Non-bonded Interaction Energy.} \end{aligned} \quad (1.7)$$

Note that the non-bonded energy term represents the sum of the energies of all possible interacting non-bonded atoms. It often contains separate terms for estimating charge and vdW (van der Waals) interactions. Typically, the following terms or variations thereof are used for estimating the potential energy of a particular three-dimensional conformation:

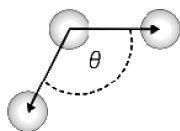


*Stretching Energy*

$$E_{\text{stretching}} = \sum_{\text{bonds}} k_b (r - r_0)^2 \quad (1.8)$$

The bond stretching energy equation (Eq. 1.8) estimates the bond vibration energy, where  $k_b$  controls the stiffness of the bond “spring”,  $r$  is the actual and

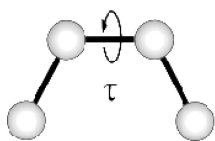
$r_0$  the equilibrium bond length. For different atom-pairs connected by a bond, unique parameter values are assigned.



*Bending Energy*

$$E_{\text{bending}} = \sum_{\text{angles}} k_{\theta} (\theta - \theta_0)^2 \quad (1.9)$$

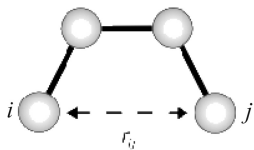
where  $k_{\theta}$  controls the stiffness of the angle “spring”,  $\theta$  is the actual and  $\theta_0$  the equilibrium angle. The bending energy equation (Eq. 1.9) is an estimate of the energy associated with vibration about the equilibrium bond angle. Again, different unique parameter values are used to describe different atom triplets and the associated bending energy.



*Torsion Energy*

$$E_{\text{torsions}} = \sum_{\text{torsions}} A [1 + \cos(n\tau - \Phi)] \quad (1.10)$$

where  $\tau$  is the dihedral angle spanned by the bonded atom quartet. Torsion energy (Eq. 1.10) is expressed as a periodic function. In this model,  $A$ ,  $n$ , and  $\Phi$  are empirically determined parameters that define the amplitude, periodicity, and shift of the cosine function.



*Non-bonded Energy*

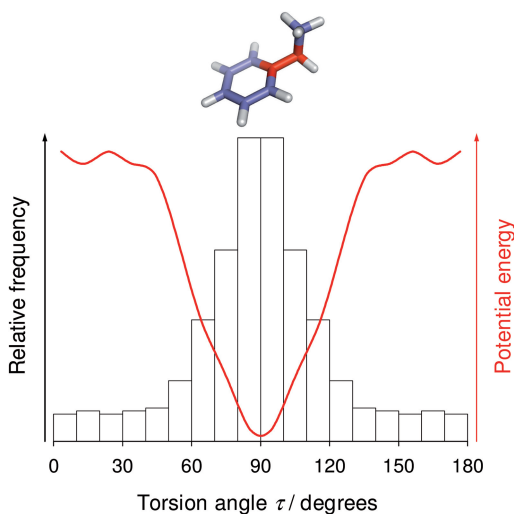
$$E_{\text{non-bonded}} = E_{\text{Coulomb}} + E_{\text{vdW}} = \sum_i \sum_j \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_i \sum_j \left( \frac{B_{ij}}{r_{ij}^{12}} - \frac{A_{ij}}{r_{ij}^6} \right) \quad (1.11)$$

where  $E_{\text{Coulomb}}$  and  $E_{\text{vdW}}$  estimate energies of point charge interactions (“Coulomb potential”) and van-der-Waals dispersive interactions, respectively. Here  $q_i$  and  $q_j$  are localized charges of atoms  $i$  and  $j$  at distance  $r_{ij}$  and  $\epsilon$  is the dielectricity constant. Parameters  $A$  and  $B$  of the vdW term define the depth and position of the potential energy wall for a given pair of non-bonded interacting atoms. In this model (Eq. 1.11), attractive interactions are expressed by the  $1/r^6$  term, and repulsion by the stronger  $1/r^{12}$  dependency (“Lennard-Jones” potential). Repulsion occurs when the interatomic distance falls below the sum of the contact radii

of the atoms, that is, in the case of collision (cf. Fig. 1.3). Partial charges can be calculated using *ab initio* methods like MOPAC or GAUSSIAN or estimated using empirical calculations like the Gasteiger–Marsili (1978) approach that is grounded on atom electronegativity.

Once a suitable force-field has been chosen, optimization techniques are employed to find minimum energy conformations of a molecule. Typically, conformations yielding energy values up to 25 kJ mol<sup>-1</sup> above the lowest energy conformation are considered “realistic”.

The second approach is complementary to molecular mechanics force-fields: knowledge-based potentials. In contrast to the general force-field equation (Eq. 1.7) the idea is to *implicitly* capture interaction energies. This is achieved by investigating known low-energy conformations of molecules. The Cambridge Structure Database provides such a constantly growing knowledge base with experimentally determined structures. Statistical analysis of conformations leads to histograms of molecular parameter values like interatomic distances, torsion angles, or ring conformations. Once sufficient data have been analyzed, these histograms are converted to obtain potential functions making a simple assumption: the maxima in the histograms correspond to preferred, low-energy conformations (Fig. 1.19). This is expressed as a log-odds score (Eq. 1.12): Preferred



**Fig. 1.19** Construction of a knowledge-based potential for a torsion angle. The histogram of observed torsion angle values was derived from experimentally determined conformations of ethylbenzene (rotation around the bond linking the ethyl group to the benzene ring). A potential energy function (curve) was derived using the *Inverse Boltzman Method*. The energy function has its minimum for the molecular conformation shown ( $\tau = 90^\circ$ ).

conformations have parameter values above the “background”. A crucial step is to define a suitable reference state  $p_{\text{ref}}$  which is used to define the expected “background” parameter distribution. Typically, an equal distribution is assumed, meaning that all values are expected to occur with the same frequency.

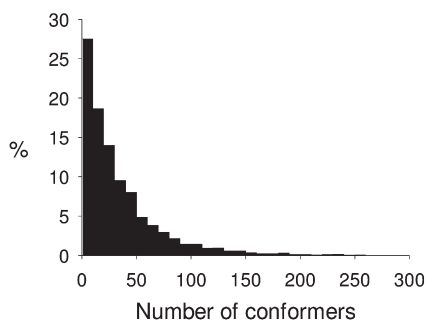
$$\Delta E_{ij}(r) \propto -\ln \frac{p_{ij}(r)}{p_{\text{ref}}} \quad (1.12)$$

where  $p_{ij}$  is the observed probability of finding atoms  $i$  and  $j$  at distance  $r$ . Such a relationship can be calculated for all parameter types (torsion angles, interatomic distances, etc.). Converting an experimentally observed distribution of parameter values to an empirical potential function is also referred to as the *Inverse Boltzmann Method*. For determination of a preferred conformation of a molecule, idealized values for its bond distances, torsional angles etc. are taken from look-up tables, and the resulting total energy is computed using the empirical potential functions.

#### 1.14

##### The “Bioactive” Conformation

Computational virtual screening – for example by similarity searching – is an often applied approach in pharmaceutical research that builds on knowledge to efficiently propose novel ligand candidates for a given receptor. Virtual screening for novel ligands corresponds to searching for molecules that comprise the necessary 3D arrangement of interacting groups which are essential for binding. A variety of computational methods is available for the estimation of ligand binding properties, among which automated docking methods and pharmacophore models are the most widely used applications (cf. Chapter 2). Such concepts are based on fitting a three-dimensional conformation to a receptor structure and thus fundamentally rely on the presence of the potential “bioactive” conformation (that is, a receptor-bound conformation) of the molecule under consideration. This conformation is not easy to find since it usually does not correspond to the global energy minimum conformation in an unbound state (Fig. 1.20). Even worse: For many ligands their bound conformation does not correspond to a minimum at all. Although it is clear that, in principle, each bioactive conformation could be reproduced computationally, there are still practical limits in the number of conformations that can be handled efficiently due to the exponential increase in the number of potential conformations with a growing number of rotatable bonds. In general, receptor-bound conformations are almost impossible to predict from the calculated ensemble of potential conformers. For example, semiempirical and *ab initio* calculations disfavor the planar conformer of acetylsalicylic acid, whereas force-field calculations imply that the planar conformer is more stable. This points to fundamental problems inherent to automatic conformer generators:



**Fig. 1.20** Number of conformers generated for molecules of a drug database (COBRA collection) using the conformer generation software Omega (OpenEye). On average, 36 conformations were generated.

- The *influence of solvent molecules* (water, ions) on the structure of a small molecule is usually neglected.
- *Flexible-fit effects* are usually not taken into consideration for energy calculation, that is, the influence of the receptor on the ligand upon binding.
- *Empirical and knowledge-based terms* of an energy function, which were derived from sets of experimentally determined structures, are only applicable within the structural diversity of these reference data – in other words, new molecular structures that were not covered by the reference compounds will be assigned wrong conformations.

One must keep these pitfalls in mind when handling computer-generated structures. Empirical studies actually revealed that it can be advantageous to use a single “preferred” conformer instead of a whole ensemble of multiple low-energy conformers for virtual screening (cf. Chapter 4).

## Literature

- D. K. Agrafiotis, A. C. Gibbs, F. Zhu, S. Izrailev, E. Martin, Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.* 2007, **47**, 1067–1086.
- L. Costantino, D. Barlocco, Privileged structures as leads in medicinal chemistry. *Curr. Med. Chem.* 2006, **13**, 65–85.
- P. Ertl, Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* 2003, **43**, 374–380.
- J. Gasteiger, M. Marsili, A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* 1978, **34**, 3181–3184.
- J. Gasteiger, T. Engel (Eds). *Cheminformatics – A Textbook*, Wiley-VCH, Weinheim 2003.
- M. Hann, B. Hudson, X. Lewell, R. Lively, L. Miller, N. Ramsden, Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 897–902.

- H. D. Höltje, W. Sippl, D. Rognan, G. Folkers, *Molecular Modeling – Basic Principles and Applications*, Wiley-VCH, Weinheim 2003.
- M. A. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, Wiley, New York 1990.
- F. E. Koehn, G. T. Carter, The evolving role of natural products in drug discovery, *Nat. Rev. Drug Discov.* 2005, **4**, 206–220.
- A. R. Leach, *Molecular Modeling – Principles and Applications*, Prentice Hall, Harlow 1996.
- A. R. Leach, V. J. Gillet, *An Introduction to Chemoinformatics*, Kluwer Academic Publishers, Dordrecht 2003.
- B. Lee, F. M. Richards, The interpretation of protein structures: Estimation of static accessibility, *J. Mol. Biol.* 1971, **55**, 379–400.
- C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.* 2001, **46**, 3–26.
- Y. C. Martin, A bioavailability score, *J. Med. Chem.* 2005, **48**, 3164–3170.
- Y. C. Martin, J. L. Kofron, L. M. Traphagen, Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 2002, **45**, 4350–4358.
- G. Moreau, P. Broto, The autocorrelation of a topological structure: A new molecular descriptor, *Nouv. J. Chim.* 1980, **6**, 359–360.
- D. J. Newman, G. M. Cragg, K. M. Snader, Natural products as sources of new drugs over the period 1981–2002, *J. Nat. Prod.* 2003, **66**, 1022–1037.
- J. Y. Ortholand, A. Ganesan, Natural products and combinatorial chemistry: back to the future, *Curr. Opin. Chem. Biol.* 2004, **8**, 271–280.
- M. Pastor, G. Cruciani, GRIND-INdependent Descriptors (GRIND): A novel class of alignment independent three-dimensional molecular descriptors, *J. Med. Chem.* 2000, **43**, 3233–3243.
- T. J. Richmond, Solvent accessible surface area and excluded volume in proteins, *J. Mol. Biol.* 1984, **178**, 63–89.
- P. Schneider, G. Schneider, Collection of bioactive reference compounds for focused library design, *QSAR Comb. Sci.* 2003, **22**, 713–718.
- D. M. Schnur, M. A. Hermsmeider, A. J. Tebben, Are target-family-privileged substructures truly privileged? *J. Med. Chem.* 2006, **49**, 2000–2009.
- D. Weininger, SMILES, A chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* 1988, **28**, 31–36.
- D. Weininger, A. Weininger, J. L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation., *J. Chem. Inf. Comput. Sci.* 1988, **29**, 97–101.
- Protein Database (PDB) [www.pdb.org](http://www.pdb.org)
- Cambridge Structural Database (CSD) [www.ccdc.cam.ac.uk/products/csd/](http://www.ccdc.cam.ac.uk/products/csd/)
- DrugBank <http://redpoll.pharmacy.ualberta.ca/drugbank/>
- Daylight Theory Manual [www.daylight.com/dayhtml/doc/theory/theory.toc.html](http://www.daylight.com/dayhtml/doc/theory/theory.toc.html)
- Relibase: A program for searching protein-ligand databases <http://relibase.ebi.ac.uk>