1
De Novo Design: From Models to Molecules

Gisbert Schneider and Karl-Heinz Baringhaus

> *Form ever follows function, and this is the law.*
> *Where function does not change, form does not change.*
>
> Louis Sullivan, American architect (1896) [1]

Innovative bioactive agents fuel sustained drug discovery and the development of new medicines. Future success in chemical biology and pharmaceutical research alike will fundamentally rely on the combination of advanced synthetic and analytical technologies that are embedded in a theoretical framework that provides a rationale for the interplay between chemical structure and biological effect. A driving role in this setting falls on leading edge concepts in computer-assisted molecular design, by providing access to a virtually infinite source of novel druglike compounds and guiding experimental screening campaigns. In this chapter, we present concepts and ideas for the representation of molecular structure, suggest predictive models of structure–activity relationships, and discuss approaches that have proved their usefulness and will contribute to future drug discovery by generating innovative bioactive agents. We also highlight some of the current prohibitive aspects of fully automated *de novo* design that will require attention for future methodological breakthroughs. This chapter provides an introduction to important pillars of *de novo* drug design, whereas the subsequent contributions presented in this book offer in-depth treatments of current trends, methods, and approaches together with numerous practical examples. We are confident that the reading will inspire.

1.1
Molecular Representation

Ever since the first atomic models of molecules have been conceived, scientists have used such models, and their associated concepts and language, to come up with innovative chemical agents that possess sought properties [2]. So far, we tend to think of a molecule in terms of sticks and balls when it comes

De novo Molecular Design, First Edition. Edited by Gisbert Schneider.

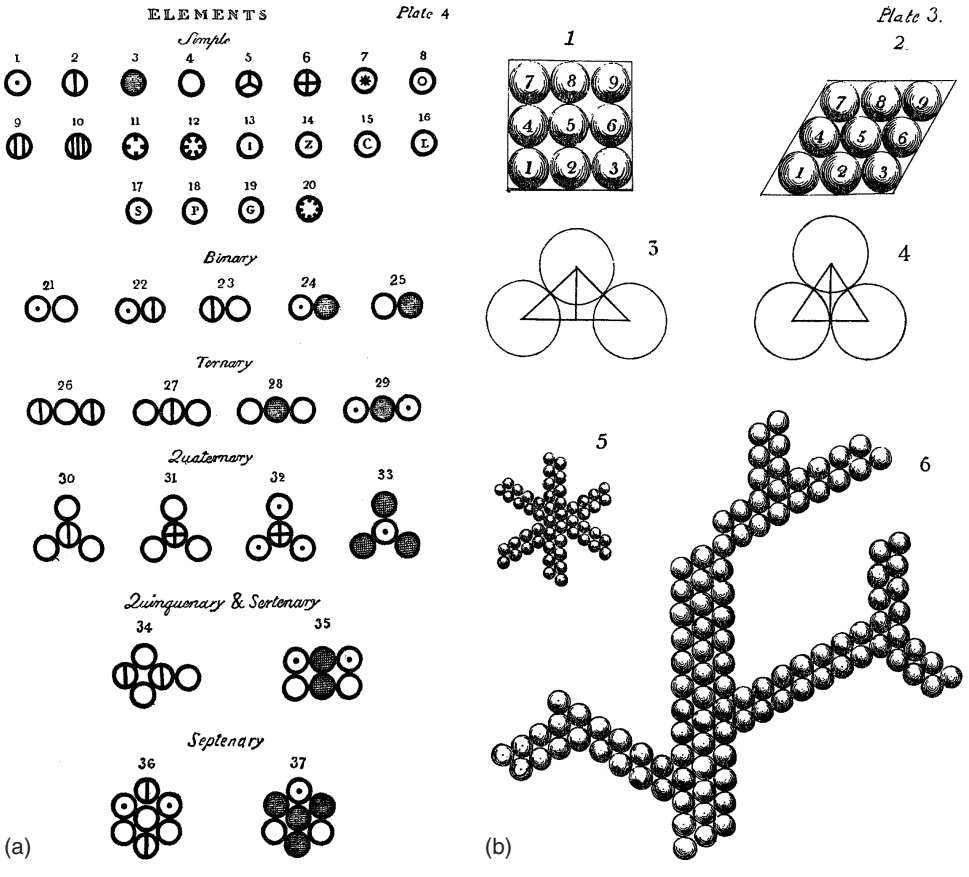


Figure 1.1 Atomic models of molecular structure as depicted in John Dalton's seminal book entitled *A New System of Chemical Philosophy* (1808). Panel (a) presents the "arbitrary signs chosen to represent the several chemical elements or ultimate particles." Panel (b) might be considered as an early molecular design study, as it depicts Dalton's view of various arrangements of water molecules. Note the similarity between these archaic philosophies and contemporary molecular models.

to visualize chemical structure. No doubt, simplistic representations have their justification for describing certain aspects of molecular constitution, configuration, and conformation and provide an intuitive access to "molecular architecture" (Figure 1.1). However, they fall far short of relating functional aspects to the objects we recognize as molecules. In the end, it is the desired *function* we wish to get from a molecular *structure*. "Form follows function" – this credo of modern architecture and industrial design is equally valid for molecular design, in particular in medicinal chemistry and chemical biology striving for new chemical entities (NCEs) as biologically active lead compounds and eventually future drugs.

Ideally, one would like to obtain a compound with a desired function directly from a design hypothesis, for example, a mathematical model that serves as a

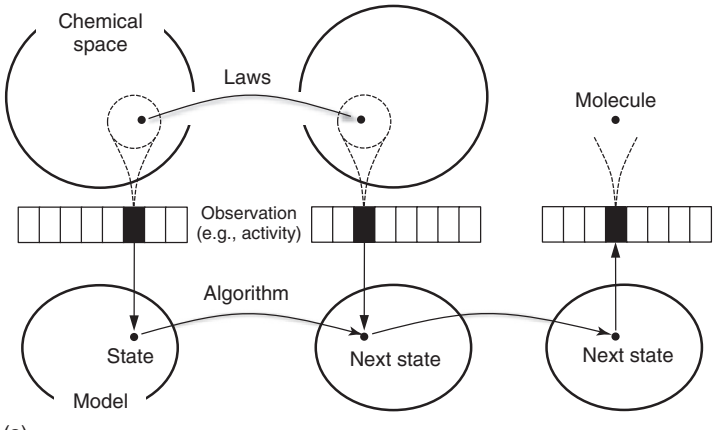
blueprint, without the need for exhaustive screening and meticulous optimization. In fact, *de novo* design means generating new molecules with desired properties ''from scratch.'' The concept of using transition functions that assign new states to objects, thereby observing *emergent system properties* [3, 4], has been well researched in fields such as complexity analysis, dynamical system, game theory, and systems biology [5]. In molecular design, we use models of the molecular world and expect a *trustworthy* model to correctly reflect aspects of the real world, so it can be used for predicting new molecules that possess the target property reflected in the model (Figure 1.2a). *De novo* design theory is tightly related to solving the *inverse quantitative structure–activity relationship* (SAR) *problem* or – to paraphrase from a philosophical point of view – finding the ''Urbild,''[1)] that is, the structural archetype associated with a molecular representation. In terms of mathematics, one tries to find an element x that is related to the value ξ: $\xi = f(x)$. In molecular design, x is a molecular structure from the set of all compounds (usually referred to as *chemical space*) and ξ is the representation (descriptor) of x computed by function f [8]. Typically, the representation of a compound is a real numbered value or set of values (vector representation), although other, for example, symbolic forms of representations have been suggested [9]. It is essential to realize that the representation of a chemical structure is always uniquely defined by the mapping function f, while there may exist – if defined – many possibly infinite numbers of molecules that have the exact same descriptor values (Figure 1.2b). As a basic illustration of this important point, consider the total charge descriptor of a molecule containing N atoms, which is computed as $\xi = f(x) = \sum_{i=1}^{N} q_i$, where q_i is the partial charge of atom i. Accordingly, it is easy to determine the total charge for a given molecular structure, but it there may be numerous chemically feasible compounds featuring the same total charge.

Generally speaking, molecular *de novo* design aims at generating new compounds that can be mapped to well-defined, preferred representations, that is, sets of descriptor values that characterize compounds with the desired biological or pharmacological activity. The challenge hereby is twofold, namely to

1) define a set of mathematical functions that characterize compounds with desired properties (i.e., they belong to the same equivalence class), and
2) for a given molecular representation, find corresponding Urbild compounds.

Consequently, as a prerequisite for successful design, we need an adequate representation of molecular structures and their physicochemical properties to allow the extraction of features that are responsible for a certain compound property or pharmacological activity (=function). Ideally, we need to understand the behavior of a molecule in different environments (e.g., in solution and in complex with a receptor) over time. Consequent physical treatment of molecular properties and dynamics can in principle be achieved based on solutions of the

1) The Urbild concept has multiple references and partly different meaning in mathematics and philosophy. See, for example, Refs. [6, 7].



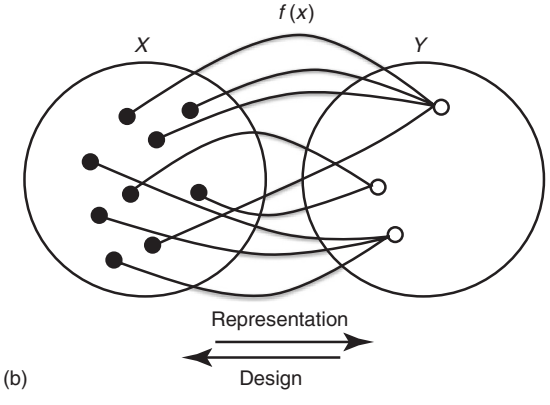


Figure 1.2 (a) Models of chemical space. (Adapted from Ref. [4].) Molecules in chemical space (real world) are lumped into an equivalence class (dotted circle) according to a structure–activity relationship model. In computer-based molecular design, appropriate algorithms act as transition functions so that changes of model states are faithfully reflected in the adaptation of molecular structure and function. (b) Molecular representation and design. A function $f : X \rightarrow Y$ transforms a molecular structure x to its corresponding molecular descriptor ξ. One may call x the "Urbild" of ξ. In molecular design applications, molecules are often mapped to numerical descriptor values by surjective functions, meaning that multiple elements of X might be turned into the same element of Y by applying f. This property of many molecular descriptor sets is exploited by *de novo* design, which aims at finding new molecules in X that can be mapped to pharmacologically meaningful representations.

Schrödinger equation (Eq. (1.1)).

$$\widehat{\mathrm{H}}\Psi = E\Psi \tag{1.1}$$

where $\widehat{\mathrm{H}}$ is the Hamilton operator defining the operations that need to be performed with the set of wave functions Ψ (psi) of the particles of a molecular system and

E is the system's potential energy. Of note, the square of the absolute value of the wave function, $|\Psi|^2$, may be interpreted as a probability density, thereby providing a *probabilistic* access to the rigid, finite "balls and sticks" of classical molecular models. The Schrödinger equation provides a rigorous theoretical foundation for *ab initio* quantum chemical (QC) and quantum mechanical (QM) calculations, which are grounded on a solid physical and mathematical framework without the necessity for empirical values or heuristics. Such calculations represent the formally most accurate way of calculating states and energies of molecular systems, allowing an assessment of conformational preferences, chemical reactivity, interaction potential, and so on. The problem is, however, that exact solutions of the Schrödinger equation cannot be obtained for molecules that are more complex than H_2^+, which currently renders druglike compounds with an average molecular mass of 300–500 Da out of reach. For such molecules of interest, approximations and generalizations are required that prohibit *exact* solutions to be found. For example, the Born–Oppenheimer approximation treats atom nuclei as fixed, and only the movement of electrons is considered. 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 wave functions (orbitals). Semiempirical approximations resulted in the Hückel theory of molecular orbitals (MOs), which can be used to derive a number of important molecular descriptors, for example, partial atomic charges and the electrostatic potential. Finally, combinations of methods that treat different parts of a system at different levels of precision permit QM calculations even for large molecular systems. While "rigorous" approaches seem perfectly suited for in-depth behavioral analysis of molecules and allow for fine-grained design and optimization, their application is currently limited because of high computational cost.

A drastic step in molecular modeling is in fact to neglect time-dependent behavior. Typically, molecules are treated as two-dimensional (2D) molecular graphs or as static three-dimensional (3D) space-filling rigid bodies with a defined surface (Figure 1.3). While such simplistic models may help us understand some basic aspects of conformational preference and molecular shape, it is important to keep in mind that they represent only crude approximations of the "true nature" of molecules. As we always work with models, it is of greatest importance that an appropriate molecular representation is applied to compound design. A molecular representation that allowed for successful drug design in one project is not necessarily generally applicable. Rather, it should always be considered as a *context-dependent* model with a local validity domain only. In general, abstraction-based object models use computation to implicitly solve complex underlying equation systems because closed-form mathematical models are unavailable or difficult to derive.

There actually are only very few molecular representations used in molecular design that are unambiguously related to their associated chemical Urbild. An example is given by the topological distance matrix $\mathbf{D}^{topo}$ that contains distance values as numbers of bonds connecting all pairs of atoms of a molecular graph along the shortest path (Table 1.1). As we will discuss later, contemporary *de*

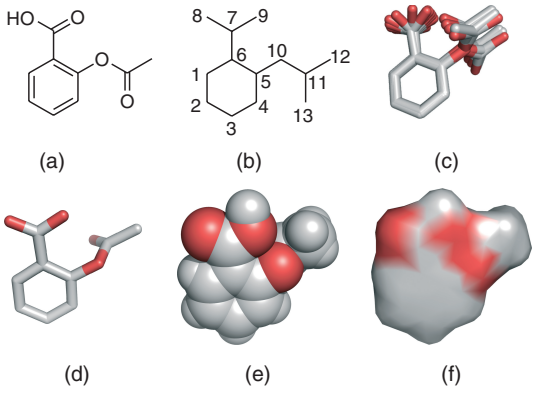


Figure 1.3 Commonly used models of molecular structure. (a) Acetylsalicylic acid is shown as a two-dimensional chemical graph, (b) the corresponding indexed molecular graph, (c) a computed ensemble of low-energy conformations, (d) and a crystal structure model with its (e) vdW and (f) SAS surfaces. For a definition of molecular surfaces, see Figure 1.4.

Table 1.1 Topological distance matrix of acetylsalicylic acid (cf. Figure 1.3b).

D^{topo}	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

novo design often relies on machine learning models that compute the mapping function $f(x)$ implicitly, for example, by kernel-function approaches, rather than employing precalculated descriptor values.

As most drug–receptor interactions are reversible and dominated by noncovalent interactions, the concepts of molecular surfaces and surface properties are relevant for drug design. Surfaces define an ''inside'' and an ''outside'' of a molecule and facilitate modeling of molecular objects as 3D bodies with a finite shape and volume. Thus, it is convenient to describe a molecule by a *continuous spatial function*. Grounded on the work of Lee and Richards, the solvent-accessible surface (SAS)

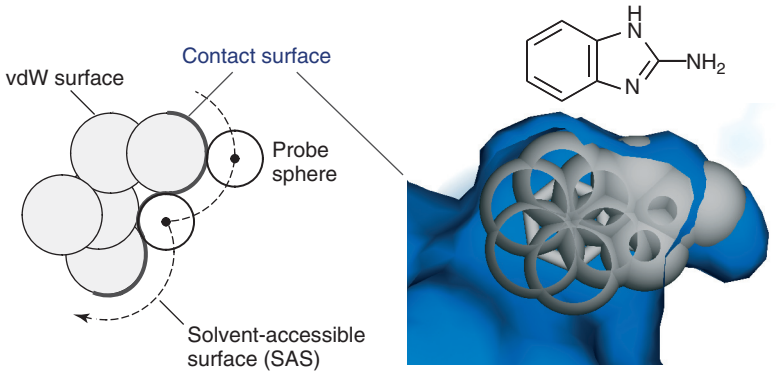


Figure 1.4 Definition of molecular surfaces. For calculation of the solvent-accessible molecular surface (SAS), several concepts exist. The SAS was originally defined by Lee and Richards [10] as the area traced out by the center of a probe sphere representing a solvent molecule as it is rolled over the van der Waals (vdW) surface (*left*). This SAS is slightly displaced from the vdW surface. The contact surface (Connolly surface), instead, consists of the part of the vdW surface that is directly accessible to the probe sphere plus the reentrant surface that covers the gaps between the atoms. The figure on the right presents a model of the complex formed between 2*H*-benzoimidazol-2ylamine and human tryptase (PDB ID: 2fpz) with the contact surface of the protein molecule in blue and the vdW surface of the ligand in gray. The latter is incompletely drawn so that the molecule's stick model becomes partly visible. Note that there are surface clashes because of close proximity ($d = 2.6$ Å) of the ligand's primary amine and the side chain of Asp^{189}, which forms a strong charge-assisted hydrogen bridge (not shown). It is recommendable to check X-ray complexes for potential modeling artifacts before using the structural models in ligand design studies.

in its various implementations probably is the most frequently used surface representation in drug design (Figure 1.4). For most current drug design applications, Connolly's definition is employed [11]. The "Connolly algorithm" uses a virtual solvent molecule represented as a probe sphere that is rolled over the molecule's van der Waals (vdW) surface. The radius of the sphere is often chosen to be 1.4 Å, which corresponds to half the vdW diameter of a water molecule. The resulting trace defines the SAS as the contact surface, which consists of parts of the vdW surface and the smoothing trace of the probe sphere. More recently, these rigid surface models have been increasingly abandoned in favor of probabilistic surface representations allowing to consider multiple conformations, time-dependent change, and uncertainty in structural modeling. For example, atoms may be represented by Gaussian functions with a width that corresponds to the atom's vdW radius or effective diameter. In this way, molecular surfaces are analytically computed as a mixture of Gaussians, which allows for quantitative flexible shape comparison between molecules [12, 13]. Overall, more than 2000 QM and empirical descriptors have been devised over recent decades [8], approximating characteristic features of molecular structure and molecular recognition, namely

- molecular shape,
- molecular distributions, and
- molecular interactions.

Although 2D molecular design methods have become a common standard, a molecule consists of a 3D shape, and ideally one would like to construct new compounds as spatial objects. Consequently, numerous molecular 3D descriptors and alignment methods have been proposed. Examples include CoMFA (*comparative molecular field analysis*) [14], Randic molecular profiles [15], 3D-MoRSE code (*3D-molecule representation of structures based on electron diffraction*) [16], invariant moments and radial scanning and integration [17], radial distribution function descriptors [18], WHIM (*weighted holistic invariant molecular descriptors*) [19], USR (*ultrafast shape recognition*, based on statistical moments) [20], ROCS (*rapid overlay of chemical structures*, based on Gaussian densities) [21], VolSurf (*volumes and surfaces* of 3D molecular fields) [22], GETAWAY (*geometry, topology, and atom weights assembly*) [23], and shrinkwrap surfaces [24], to name some prominent representatives. As an illustrative example of how contemporary shape representation looks like, we selected spherical harmonics descriptors of molecular surfaces. Spherical harmonics have been used in molecular modeling and design as a global feature-based parameterization method of molecular shape [25]. The spherical harmonics decomposition can be viewed as a generalization of the Fourier decomposition to three dimensions. They are solutions to Laplace's differential equation in spherical coordinates of the object to be represented. In other words, spherical harmonics can be used to model the shape of a molecular structure at different levels of sophistication, that is, different levels of model abstraction from the exact atomic structure of the molecular object. As illustrated in Figure 1.5, an ellipsoid shape roughly approximates a 3D conformation of the cyclooxygenase-2 (COX-2)

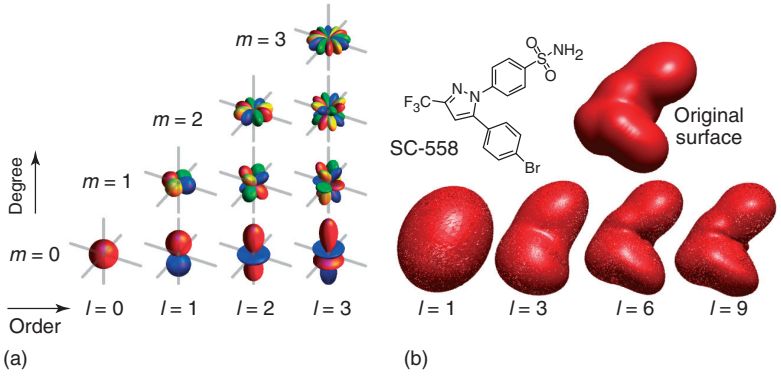


Figure 1.5 Molecular shape representation by spherical harmonics. (a) Spherical harmonics of different order and degree, with negative real (blue), positive real (red), negative imaginary (green), and positive imaginary (yellow) parts of the function. (b) The reconstruction of the molecular surface of selective cyclooxygenase-2 inhibitor SC-558 (PDB ID: 6cox) using spherical harmonics of order up to $l = 9$. Detail increases with higher order of the mathematical function.

inhibitor SC-558. Expanding this simplistic representation yields more and more fine-grained resolution. Spherical harmonics descriptors of molecular shape and volume have successfully been applied to, for example, molecular similarity searching and protein pocket analysis and comparison [26]. Such approaches have the appeal to extend our classical view of solid and static molecular surfaces and allow for the consideration of essential *dynamic* molecular properties for drug design, for example, conformational flexibility [27] and 3D pharmacophores of molecular fragments [28].

1.2 The Molecular Design Cycle

In the beginning of a drug design project, one may be faced with several scenarios depending on the already available knowledge about the drug target and its ligands and SARs. The aim is to boil down the number of relevant compounds for biochemical assaying as efficiently as possible, which means with a minimum expense of substance, time, and money. *In silico* automation of the whole design process is one way to go.

Understanding SARs is essential not only for "wet" medicinal chemistry but equally for successful computational optimization of a pharmacologically or otherwise biologically active substance. Once an SAR model is available, it is possible to perform *rational* drug design. Most importantly, any successful application of artificially optimizing systems requires a fundamental characteristic of the underlying fitness landscape (search space), namely the *principle of strong causality* [29]. For the field of drug design, this concept has been reformulated as the *chemical similarity principle* by Maggiora and Johnson [30]. Systematic compound optimization therefore requires a smooth response function or neighborhood behavior, so that small changes in molecular structure result in only small changes of biological function. A fitness function (scoring function and objective function) guiding the molecular design process must therefore be chosen and constructed wisely. Otherwise, any systematic optimization will likely fail. Fitness landscapes in molecular design possess characteristic features [31], which one encounters while optimizing a molecular structure, and are dominated by large unexplored areas (Figure 1.6). For example, there may be perceived "activity cliffs" or regions of "flatland." The

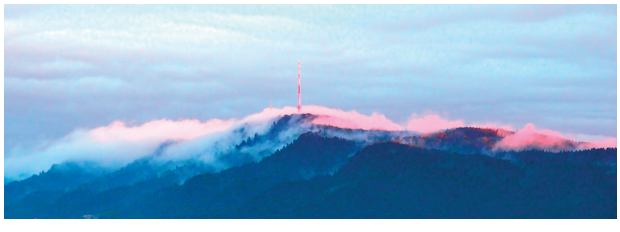


Figure 1.6 A "fitness landscape" featuring a plateau-like global optimum, several local optima, and many uncharted areas (fog) that elude straightforward optimization.

overall shape of a fitness landscape is determined by the molecular representation chosen and the underlying specific SAR of the receptor–ligand interaction under investigation [32]. Accordingly, the search for optima is often limited to finding a local optimum in proximity to the start position rather than converging on the global optimum.

The response of a linear system to small changes in its parameters, that is, alterations of molecular structure, is usually in direct proportion to the stimulation resulting in a ''smooth'' response. For nonlinear systems, however, a small change in the parameters can produce a large qualitative difference in the output. In other words, the designed new molecules do not behave as expected. There are many observations of such behavior in drug design, and we have to assume that SARs are generally nonlinear. As we do not know *a priori* which parts of a molecule are crucial determinants of bioactivity, we tend to believe that *any* small change of structure will only slightly affect molecular function. This way of thinking is often not appropriate, and the definition and quantitative description of chemical similarity is a critical issue for molecular design. Therefore, a robust model of the underlying SAR is a prerequisite for rational optimization. Of note, such a model does not (and most likely will not) necessarily have to be interpretable in terms of ''simple'' features. Rather, we should accept ''black box'' mathematical models for the purpose of automated molecule optimization and *de novo* design. The task for the molecular designer is to use such a complex model to come up with clear chemical rules for synthesis.

Computers are not required for *de novo* design. In fact, we can easily conceive of a Gedanken experiment, which demonstrates the process of structure generation by inductive learning as an instance of *adaptive* model building (formulation of a new or modified hypothesis) and assessing the value of the model (testing of the hypothesis). The Bongard problem[2)] shown in Figure 1.7 exemplifies such an adaptive optimization process [33]. With increasing knowledge about the structure of the search space including the number of compounds that have been tested experimentally up to a point and the distribution of actives and inactives, better models are formed resulting in higher hit rates during activity determination. While the human mind can do very well in this game, mathematical model building can synergistically assist in decision making. It is important to realize that not only the actual compound construction process is adaptive but also the model building process. However, in practical molecular design studies, often a static SAR model is used and the software searches for compounds that satisfy the model. As we will explore later (Section 1.3), there are algorithmic concepts of parallel model refinement and compound design. For the further interested reader, we recommend a general introduction to the theory of complex adaptive systems by Miller and Page [34].

De novo molecule design produces novel molecular structures with desired properties based on model of the fitness landscape. In this attempt, a medicinal

2) Solution to the Bongard-problem of Figure 1.7 Left: Both dots at same side of neck. Right: Dots at different sides of neck.

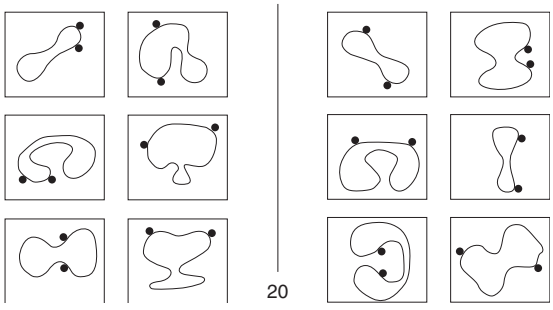
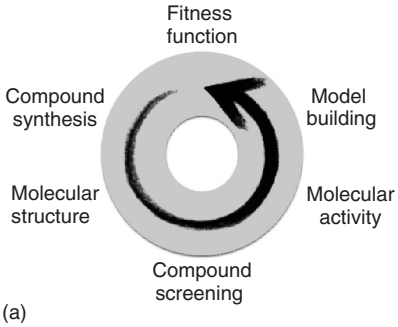


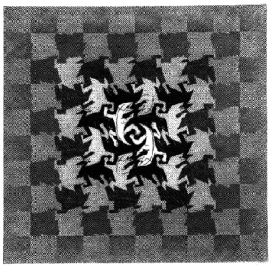
Figure 1.7 Bongard problem number 20 [33], as an example of inductive model building. Which feature separates the two classes of "molecules"? Find the classifier ("structure–activity relationship" model)! While doing the exercise, rationalize the steps of creating a vocabulary (how to represent the molecules) and adaptive hypothesis modification and testing (how to formulate and test the model).

chemist – and equally *de novo* molecule design software – is confronted with a virtually infinite search space, both in terms of possible models and possible chemical structures. Instead of the systematic construction and evaluation of each individual compound, navigation in the molecular design process relies on the *principle of local optimization*, that is, only a fraction of all potential screening candidates are actually constructed and evaluated and the design process converges on a local or "practical" optimum. Just like two medicinal chemists are likely to propose different molecules as "promising solutions," multiple runs with different *de novo* design software tools will likely produce different compounds because of the nature of the model and search algorithm employed. The trick is to incorporate as much chemical knowledge as possible about the structure of the fitness landscape into the design algorithm.

In computational *de novo* design, a virtual search agent mimics a medicinal chemist, and fitness functions perform virtual assays. In the ideal case, such an "*in silico* laboratory" suggests readily synthesizable, potent molecular structures. *Positive design* restricts this virtual optimization process to small regions of chemical space that have a higher probability to find molecules with the desired properties. *Negative design*, in contrast, defines criteria that help to prevent adverse properties and unwanted chemical structures. It is vital to understand that *de novo* design will rarely yield novel lead structures with nanomolar activity in the first place. Instead, generated structures will often represent molecules that require significant further chemical optimization. The molecular design cycle in Figure 1.8 pinpoints the basic steps of adaptive design, and Figure 1.9 illustrates an example of molecular design that started from weakly active compound **1** and by iterative structural refinement (intermediate **2**) ended with the potent and selective compound **3**. While this example was performed as interplay between computational molecular modeling and human decision making, all parts of the molecular design cycle can be performed *in silico*, thereby enabling fully automated *de novo* molecule design. Basically, the following three questions have to be addressed by a *de novo* design program:



(a)



(b)

Figure 1.8 (a,b) Adaptive molecular design cycle, and an artistic inspiration of adaptive optimization. (M.C. Escher's "Development I" ©2012 The M.C. Escher Company-Holland. All rights reserved. *www.mcescher.com*.) Initial models of the fitness function are simplistic and coarse. They are iteratively refined in consecutive rounds of virtual or real compound synthesis and testing – similar to the evolving shapes in the Escher artwork. Different molecular representations are required to properly capture the respective levels of abstraction from the atomistic chemical structure in each pass through the cycle, for example, connectivity, shape descriptors, molecular fragments, pharmacophore features, and charge models.

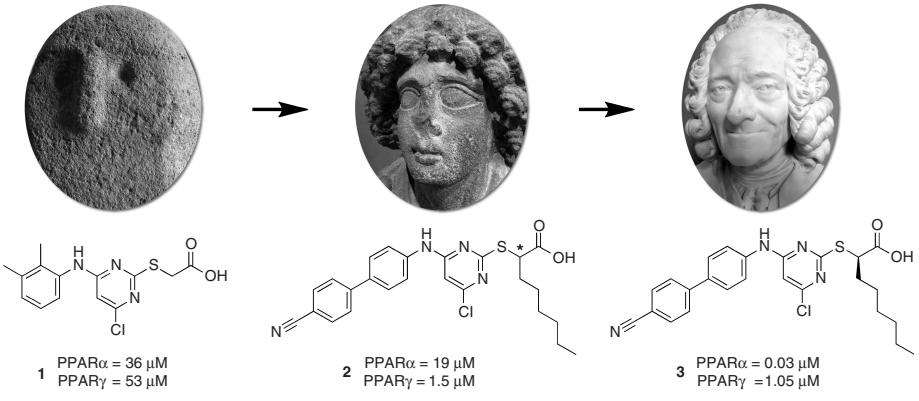


Figure 1.9 Example of adaptive molecular design. Pirinixic acid (*left*) was derivatized to yield a potent agonist of peroxisome proliferator-activated receptor subtype alpha (PPARα, *right*) in two design steps. The depictions above the chemical structures represent adaptive "design models" used for compound selection. Each step adds detail so that the initially coarse model is iteratively refined to obtain a specific structure–activity relationship.

1) How to assemble candidate compounds? (*problem of construction*)
2) How to represent molecules and assess their quality? (*problem of scoring*)
3) How to navigate in search space? (*problem of optimization*)

There are many implementations of *de novo* design algorithms using combinations of methods for performing these tasks. No matter how the various programs try to solve these challenges, almost all of them follow the fundamental

concept of mimicking the iterative adaptive process of drug discovery: molecules are generated, subsequently tested for activity, and the test results form the basis of the next round of (virtual) the synthesis. Search and assembly strategies correspond to the intellectual and technical work of a chemist, whereas scoring complies with testing the compounds for activity in a biological assay. According to Koza, artificial adaptive systems need to possess common essential elements [35], namely

- structures that undergo adaptation,
- initial structures (starting solutions),
- fitness measure that evaluates the structures,
- operations to modify the structures,
- state (memory) of the system at each stage,
- method for designating a result,
- method for terminating the process, and
- parameters that control the process.

The most important aspect of adaptive optimization probably is the system's ''memory.'' Keeping track of the past progress during an optimization process and learning from previous experience helps us make informed decisions for planning the next steps. Memory turns a blind or random search into rational design. There are many ways to implement memory in a molecular design program, and several instances are presented in the subsequent chapters of this book. Importantly, for molecular design, the memory should be adaptive, that is, it should enable large steps toward an optimum in flat regions of the fitness landscape but at the same time allow for fine-tuning a solution by small structural variations near the summit of an activity hill. A straightforward algorithm that implements an implicitly adaptive memory is the *evolution strategy*, which was conceived by Rechenberg in the late 1960s and paved the way for *genetic algorithms*, *genetic programming*, and many other adaptive optimization techniques. The simplest so-called (1, λ) evolution strategy can be formulated in just a few lines of *pseudo*-code:

```
1 Initialize parent (ξ^P, σ^P, F^P);
2 For each generation:
3  Generate λ variations (ξ^V, σ^V, F^V) of the parent (ξ^P, σ^P, F^P):
4   σ^V = abs(σ^P + G);
5   ξ^V = ξ^P + σ^V × G;
6   Calculate fitness F^V;
7  Select best variation according to F^V;
8  (ξ^P, σ^P, F^P) = (ξ^V, σ^V, F^V)^best;
9 End.
```

This stochastic algorithm is based on the interplay between variation (*lines 4–5*) and selection (*lines 7–8*) operators (Figure 1.10). The (μ, λ) notion implies the number of parents μ and offspring, λ, and the fact that the parent(s) do not participate in selection, that is, ''death'' after producing offspring. (Note that in a ($\mu + \lambda$) strategy the parents participate in the selection). In the above-mentioned

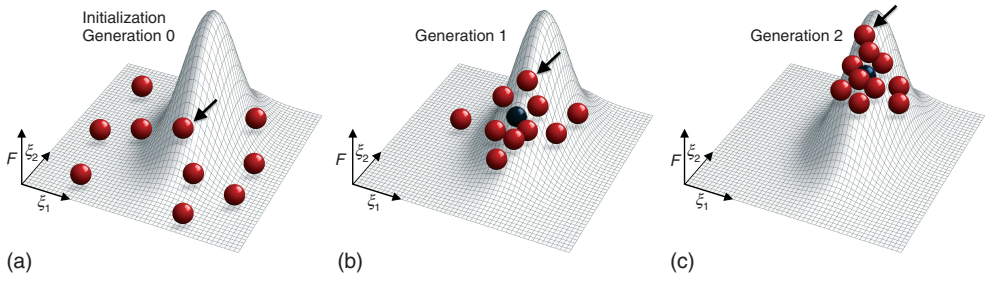


Figure 1.10 (a–c) Population-based optimization. In the example, compound library of 10 molecules (red balls) converges on the optimum of a fitness landscape model, guided by a (1, 10) *Evolution Strategy*. The arrows indicate the "winning compound" with the greatest fitness value F among all members of the population. This best solution survives and acts as the single parent (black ball) of a new generation. The width of the distribution of offspring (stepsize σ) in each generation is adaptive, that is, here it automatically assumes small values close to the optimum. Note that the parent compound does not participate in the selection of the winner.

pseudo-code, a molecular structure is represented by $\boldsymbol{\xi}$ (the molecular descriptor vector), its memory is the so-called stepsize parameter σ, and F its fitness value. G is a Gaussian-distributed *pseudo*-random number.[3)] There are three essential ingredients to this algorithm:

1) Each individual object undergoing optimization consists of three variables ($\boldsymbol{\xi}$, σ, and F) (*line 1*).
2) New solutions (individuals and molecules) are generated as a mutation of the parent object (*lines 4 and 5*). Note that the stepsize values differ for each object, as they are variations of the parental value. There are other methods for stepsize mutation than the one shown here.
3) The stepsize value of the best (fittest) member of the generation n is passed on to the parent of the next generation $n+1$, thereby the new parent inherits the memory of the most successful stepsize from the previous generation (*line 8*). This rule implements the ability of the process to adapt the local structure of the fitness landscape.

1.3 Receptor–Ligand Interaction

Ideally, one would directly compute the affinity of a newly designed molecule to its macromolecular target(s) [36] (cf. Chapter 16). The resulting fitness landscape guiding the *de novo* design process would then be expressed as a function relating the free energy of binding (Gibbs energy) ΔG to a given molecular representation

3) Approximately Gaussian-distributed, zero-centered *pseudo*-random numbers can be computed by the Box–Muller method: $G(i, j) = \sqrt{-2\ln(i)}\ \sin(2\pi j)$, where i and j are *pseudo*-random numbers in]0,1].

ξ (descriptor space). In fact, we will see later in several chapters of this book that such predictive functions may (i) either be developed from scratch (empirical or knowledge-based approach), provided that a sufficient number of accurate reference examples are available, or (ii) be derived from first principles (physically motivated approach). In general, we aim at designing noncovalent ligands, as the complex formation between drug molecules and their macromolecular targets is governed by noncovalent interactions. Covalent binders exist and have their applications – for example, in antitumor therapy – but typically reversible pharmacological effects are desirable, as covalent binding of drugs or their reactive metabolites can lead to various forms of drug toxicity. The reversible bimolecular interaction between a protein P and a ligand L forming the complex PL (Figure 1.11) can be formulated in a simplified manner (Eq. (1.2)):

$$\mathrm{R} + \mathrm{L} \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} \mathrm{RL} \tag{1.2}$$

Note that this scheme does not consider any other interactions that accompany the formation of a receptor–ligand complex *in vivo*, for example, migration of a

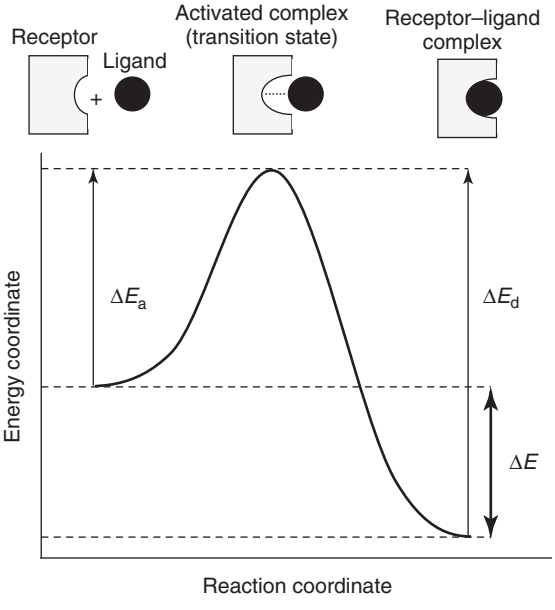


Figure 1.11 Energy diagram for a reversible bimolecular interaction between a receptor macromolecule and a small-molecular ligand. Changes in energy (energy coordinate) lead to changes in the "reaction coordinate." ΔE_a and ΔE_d denote the activation energies required for association (forward reaction = complex formation) and dissociation (backward reaction). ΔE is the overall change in energy for the interaction, which is here negative by definition, that is, the energy level of the receptor–ligand complex is below the energy level of the free interaction partners. The activated complex represents a transient state of loose association between the receptor and the ligand.

ligand to the active site, activation of second messenger transduction processes, or interaction with the solvent, membrane, and other macromolecules.

The free energy change ΔG that accompanies a receptor–ligand interaction has been defined by J. W. Gibbs in 1873 and is often referred to as *Gibbs energy* (Eq. (1.3)).

$$\Delta G = \Delta H - T\Delta S \tag{1.3}$$

The free energy change is governed by two contributions: the *enthalpic* and *entropic* terms. The change in enthalpy ΔH corresponds to the molecular forces involved in the receptor–ligand interaction, whereas the change in entropy $T\Delta S$ can be understood as the amount of energy in a reaction system that cannot be used to do thermodynamic work or – from a statistical point of view – as the change in the degrees of freedom (uncertainty) of a molecular system. Generally, an overall increase in entropy favors the formation of a ligand–receptor complex, whereas an overall increase in enthalpy disfavors the interaction. Any reversible ligand–receptor interaction is the result of enthalpic and entropic contributions, which partially compensate each other. Depending on the dominating forces, one distinguishes between the enthalpy-driven and entropy-driven formations of a receptor–ligand complex.

The experimentally accessible equilibrium constant of an interaction, K_{eq}, is directly related to the change of Gibbs energy for a given receptor–ligand complex. It is defined as the quotient of the rate constant of the forward binding k_{on} and the backward dissociation k_{off} (Eq. (1.4)). The square brackets indicate concentrations of the receptor–ligand complex [RL], the free receptor [R], and the ligand [L]. The reciprocal of the equilibrium constant is termed the *dissociation constant*, K_d (Eq. (1.5)). The terms *binding constant* or *inhibition constant*, denoted as K_i, are more often used as synonyms for the dissociation constant. Note that this definition is not consistently used in the literature, and confusion easily arises from the improper use of these terms. The meaning of K_d can be explained at the molecular level: If the free ligand concentration reaches the value of K_d, then 50% of the ligand binding sites of the receptor are occupied.

$$K_{eq} \equiv \frac{k_{forward}}{k_{backward}} = \frac{[RL]}{[R] \cdot [L]} \tag{1.4}$$

$$K_d = \frac{[R] \cdot [L]}{[RL]} \tag{1.5}$$

The Gibbs energy change of a bimolecular interaction is calculated from Eq. (1.4):

$$\Delta G = \Delta G^\circ + \overline{R}T \ln\left(\frac{[RL]}{[R] \cdot [L]}\right) \tag{1.6}$$

where ΔG° is the *standard free energy change of interaction*, that is, the change of Gibbs free energy that accompanies the formation of the complex from their component elements at equilibrium standard state (the ''standard state'' is by definition at 25 °C and a pressure of 100 kPa; for elements, $\Delta G^\circ \equiv 0$), $\overline{R}$ is the gas (or molar) constant ($\overline{R} = 1.99\,\text{cal} \times \text{mol}^{-1} \times \text{K}^{-1} = 8.31\,\text{J} \times \text{mol}^{-1} \times \text{K}^{-1}$), and

T is the absolute temperature in Kelvin ($T = {}^{\circ}\mathrm{C} + 273.15$). $\Delta G = 0$ at *steady-state* (equilibrium) conditions of the interaction, so that Eq. (1.5) relates the standard free energy change to the dissociation constant of a reaction. An important consequence of Eq. (1.5) is that it permits to calculate the energy of a receptor–ligand interaction from the experimentally obtained equilibrium constant.

$$\begin{aligned}\Delta G^{\circ} &= -\overline{R}T\ln\left(\frac{[\mathrm{RL}]}{[\mathrm{R}][\mathrm{L}]}\right) = -\overline{R}T\ln(K_{\mathrm{eq}}) \\ &= -\overline{R}T\ln\left(\frac{1}{K_{\mathrm{d}}}\right) = \overline{R}T\ln(K_{\mathrm{d}}) = 2.303\cdot\overline{R}T\log(K_{\mathrm{d}})\end{aligned} \tag{1.7}$$

To get an idea of the order of magnitude of ΔG° for a strong ligand–receptor interaction at body temperature ($37\,{}^{\circ}\mathrm{C} + 273.15 = 310.15\,\mathrm{K}$): for $K_{\mathrm{d}} = 10\,\mathrm{nM}$ ($=10^{-8}\,\mathrm{mol}\times\mathrm{l}^{-1}$), ΔG° is $-47.4\,\mathrm{kJ}\times\mathrm{mol}^{-1}$. As a rule of thumb, the experimentally determined binding constant assumes values between $10^{-3}\,\mathrm{mol}\times\mathrm{l}^{-1}$ (millimolar range) and $10^{-12}\,\mathrm{mol}\times\mathrm{l}^{-1}$ (picomolar range), corresponding to Gibbs energy values between approximately -17 and $-70\,\mathrm{kJ}\times\mathrm{mol}^{-1}$ in aqueous solution.

The enthalpic term can be attributed to noncovalent interaction energies resulting from the formation and disruption of

- hydrogen bridges (also termed *hydrogen-bonds*; with ionic interactions and covalent bonds as extreme forms of hydrogen bridges [37]),
- arene–arene (aromatic) and arene–charge interactions, and
- dispersive interactions between dipoles or induced dipoles (vdW interactions).

Hydrogen-bonding and aromatic interactions are usually considered as *directed* and dispersive (lipophilic) interactions as *undirected*. In a simplistic view, which still dominates *de novo* design, hydrogen-bonding patterns are often considered to critically influence the selectivity of a ligand–receptor interaction because of their directed nature. Reality, however, is more complex as hydrogen-bonding interactions can have vastly different energies, between 0.25 and 40 $\mathrm{kcal}\times\mathrm{mol}^{-1}$, and the more appropriate term would be *hydrogen bridge* to express this fact (Figure 1.12). According to Desiraju, *"a hydrogen-bond, X–H···A, is an interaction wherein a hydrogen atom is attracted to two atoms, X and A, rather than just one and so acts like a bridge between them"* [37]. Critical revision of our current view of receptor–ligand interaction is required for progress in *de novo* design. Often, we limit the design process to modeling enthalpic interactions by playing with sticks and colored balls. The reality is more subtle and delicate. In addition, it should always be kept in mind that an X-ray structure downloaded from the Protein Data Bank (PDB) or obtained through own experiments is nothing but a model that was obtained by fitting simplistic molecule representations into observed electron densities. An X-ray structural model is not reality – it can nevertheless be extremely useful for *de novo* design when appropriately used.

The overall entropic contribution to ligand binding results from the changes of the degrees of freedom of *all* interaction partners on the complex formation. It is important to keep in mind that ligand–receptor interactions do not take place in vacuum. Drug–receptor interactions typically occur in aqueous solution,

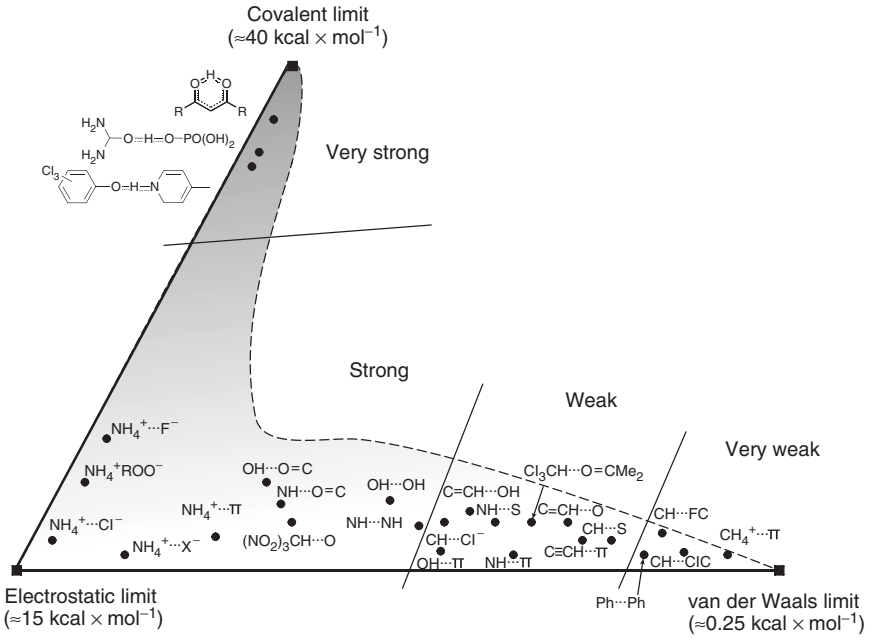
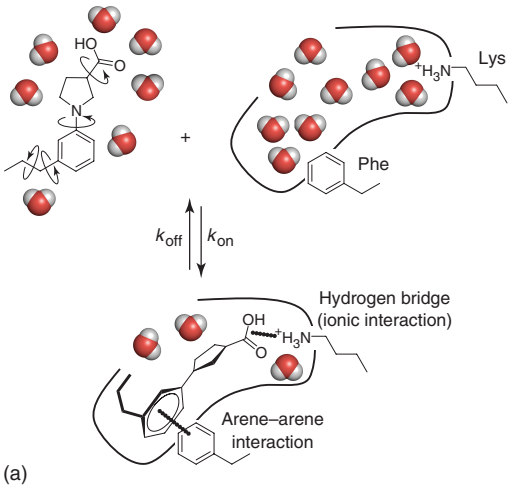
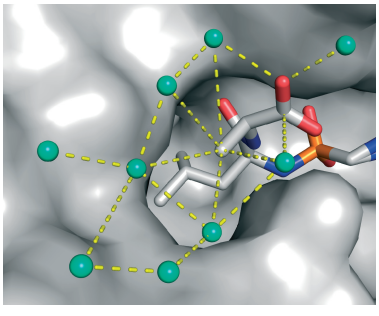


Figure 1.12 The strength of hydrogen bridge can be very different depending on the interacting atoms and their local environment. (This depiction was adapted from Ref. [37].)

and solvent molecules contribute to the entropic term of Eq. (1.3). The accurate determination of the role of water molecules in a ligand–receptor interaction still is one of the biggest challenges in molecular modeling and design. Figure 1.13a shows a sketch of a drug–protein interaction in water. Both the free ligand and the protein are fully solvated before complex formation. Water molecules undergo hydrogen-bonding interactions with other water molecules, the ligand, and the amino acid residues at the protein surface. On ligand binding, the ligand and the protein surface residues interact with each other forming favorable interactions, but at the same time, their conformational freedom is reduced (reduction of entropy). The bound ligand conformation is often referred to as the *bioactive conformation*, although this is not necessarily correct (also note that the ligand conformation observed in a complex needs not necessarily correspond to the lowest energy conformation in vacuum or solvent [38, 39], and conformational sampling techniques are required for proper evaluation of a meaningful conformer ensemble [40, 41]). The loss of degrees of freedom of the receptor and the ligand during complex formation is countered by an increase of entropy resulting from the release of receptor-bound water molecules into the bulk solvent. In particular, the release of water from hydrophobic surface patches inside the binding pocket into the solvent contributes favorably to the entropy term. This ''hydrophobic effect'' can be the driving force of ligand–receptor association. Water molecules adopt an entropically unfavorable ordered structure near hydrophobic surfaces because



(a)



(b)

Figure 1.13 (a) Schematic of the ligand–receptor binding process, during which all species coexist in a solvated state. The interaction between the ligand and the protein surface involves the release of water molecules from the binding cavity into the solvent as well as the loss of conformational mobility of both ligand and receptor. (b) Water molecules of the first solvation layer around a peptide-like thermolysin inhibitor (PDB ID: 3t74). Note the five water molecules around the terminal methyl group of the ligand. Structurally small modifications of the ligand can cause massive rearrangements of the surrounding water network.

they cannot form polar contacts with the protein at these sites. Once removed from these strained structures, their degrees of freedom are markedly increased (entropic contribution), and newly formed contacts with bulk water (enthalpic contribution) additionally contribute to an overall negative change of free energy. The contribution of the hydrophobic effect to complex formation is approximately proportional to the size of the lipophilic surface area shed by the ligand, which is often in the range of 80–200 J/(mol × Å^2).

As a general guideline for ligand design, hydrophobic surface patches of the ligand-binding pocket should be covered by hydrophobic parts of the ligand. The bound ligand conformation always tends to maximize the lipophilic interaction between lipophilic parts of the ligand and corresponding parts of the binding pocket.

In a series of elegant thermodynamic and crystallographic studies, Klebe and coworkers [42] deciphered hydrogen-bonding networks of water molecules in ligand binding sites and characterized the effects of their rearrangement on ligand affinity. It turned out that while structurally slightly different ligands virtually adopted the identical binding mode, large observed enthalpy/entropy changes are related to rearrangements of the first bound ligand solvation layer (Figure 1.13b). Consequently, it is insufficient to consider only the ligand and a structural model of the receptor (binding pocket). Contiguously connected water networks must be considered for receptor-based *de novo* design – especially, if one is interested in

obtaining accurate quantitative ΔG estimations. The influence of solvent molecules on the observed, measured activity of a ligand could also help to explain "activity cliffs" and "magic methyls," that is, cases when the *chemical similarity principle* seemingly does not hold [43].

Many proteins also contain so-called structural water: deeply buried water molecules below the surface. On average, approximately one water molecule per amino acid is found in the high-resolution X-ray structures deposited in the PDB [44]. It is generally assumed that "freezing in" a water molecule in a fixed position inside a protein generates a significant entropic cost. These water molecules are believed to stabilize the protein structure by forming strong hydrogen bonds with polar amino acid residues. Surprisingly, computational studies by Fischer and Verma [45] revealed the opposite: the protein actually becomes more flexible. They found that "*... this effect must be common in proteins, because the large entropic cost of immobilizing a single water molecule* ($-T\Delta S = 20.6\,kcal \times mol^{-1}$ *[...] for the lost translational and rotational degrees of freedom) can only be partly compensated by water–protein interactions, even when they are nearly perfect [...] leaving no room for a further decrease in entropy from protein tightening*." What makes this observation so important for the calculation of protein–ligand interaction energies and protein structure-based *de novo* drug design is the necessity to consider protein flexibility when generating novel ligands by matching them with binding sites.

One can expect immediate progress for receptor-based *de novo* design from a combination of flexible pocket models with advanced methods for shape and pharmacophore matching (cf. Chapter 4). Such a scoring scheme could include extended pharmacophoric features allowing, for example, for "strong," "medium," and "weak" hydrogen bridges, better consideration of arene–arene interactions and geometries, as well as explicit solvent molecules, and would allow for moderate pocket and ligand adaptation during the actual ligand construction, thereby possibly avoiding artifact ligand poses [46].

Consequently, studying, understanding, and predicting binding energies are of seminal importance for molecular design. Figure 1.14 presents experimentally

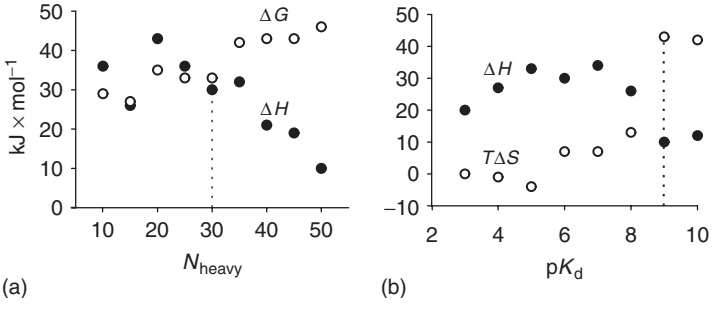


Figure 1.14 Thermodynamics of ligand binding in medicinal chemistry projects. (Adapted from Refs. [47, 48].) Highly potent ligands are often structurally complex and obtained by entropy-driven optimization. (a) N_{heavy}: number of non-hydrogen atoms and (b) pK_d: negative logarithm of the dissociation constant K_d.

measured binding thermodynamics of compounds that were optimized by medicinal chemistry. Apparently, for ligands exceeding approximately 30 non-hydrogen atoms, ΔG is increasingly less driven by enthalpic contributions, and entropic effects clearly govern complex formation of highly potent compounds with their macromolecular target exhibiting a K_d value in the single-digit nanomolar range ($pK_d > 8$) [48]. These thermodynamic data suggest that the overall potency that can be obtained through the formation of specific directed interactions is limited. In fact, numerous SAR studies reveal that favorable binding enthalpy is more difficult to achieve for highly potent ligands, which in turn affects the target selectivity of the compounds [49].

1.4 Modeling Fitness Landscapes

As we will see in more detail in throughout this book, there are numerous metrics and indices that can be used to compute a single value expressing an estimate of the drug- and lead-likeness of bioactive compounds using a pK_d estimate (cf. Chapters 2 and 11) [47]. These values may even be considered for preliminary compound prioritization in *de novo* design. Among the most prominent ones are the *ligand efficiency* [*LE*, Eq. (1.8)] [50], which relates a compound's potency to its size expressed as the number of non-hydrogen atoms (N_{heavy}), and one of its derivatives, the *ligand-efficiency-dependent lipophilicity* [*LELP*, Eq. (1.9)] that corrects *LE* by the influence of lipophilicity (log*P*) on potency [51].

$$\mathrm{LE} = \frac{-\overline{R}T \ \ln(K_d \text{ or } pK_d)}{N_{heavy}} \tag{1.8}$$

$$\mathrm{LELP} = \frac{\log P}{\mathrm{LE}} \tag{1.9}$$

Still, compound optimization from primary hits to pharmaceutical lead structures by organic synthesis is largely guided by the chemical feasibility and tractability of the candidate compounds, and the specific knowledge and intuition of the medicinal chemists involved. In this context, it is advisable to start a molecular design project by sampling compounds from chemical space to obtain a reasonably diverse pool for modeling activity landscapes. Maximum diversity methods aim at covering the variability of the complete compound pool within a carefully chosen small subset. Cell- and dissimilarity-based clustering and partitioning methods are often employed for this purpose. *Diverse* compound sets often represent reasonable starting points for screening campaigns, whereas *focused* libraries, in contrast, typically contain substances only from a certain region (activity island) of the chemical space. Generic filtering steps for drug- and lead-like compounds in conjunction with target-specific prediction and selection tools have been shown to be suited for designing activity-enriched focused libraries [47]. A selection of unwanted fragments and substructures are shown in Figure 1.15. For example,

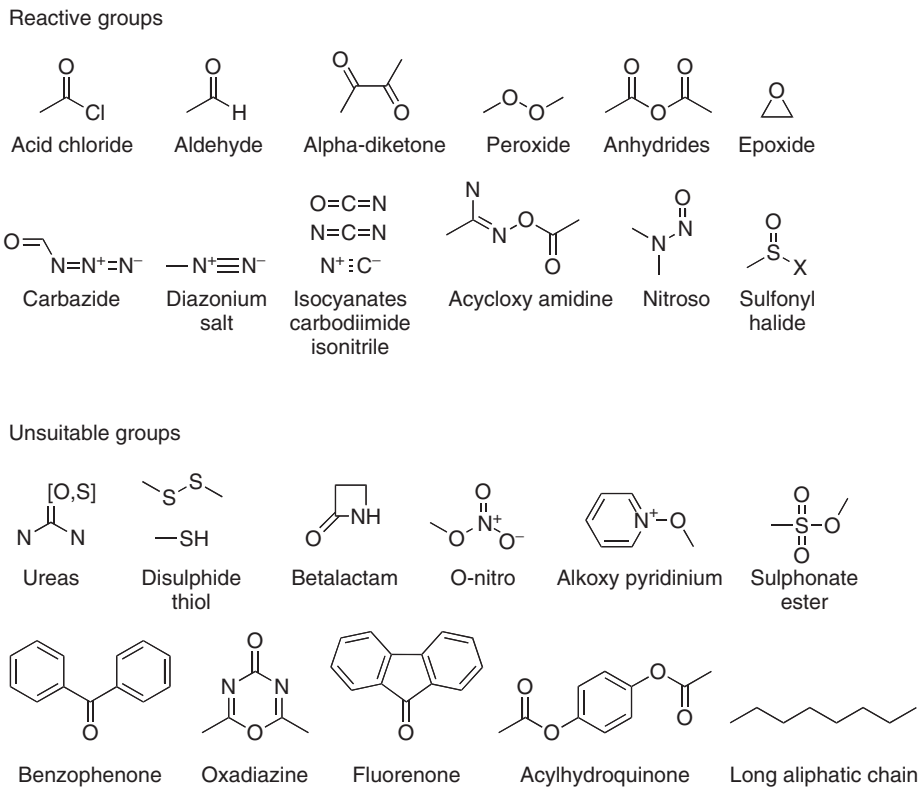


Figure 1.15 Examples of functional groups and substructures that are usually undesirable for drug design. Note that exceptions from these guidelines can sometimes be well motivated, for example, to obtain covalent inhibitors.

the REOS (*rapid elimination of swill*) approach is well suited for eliminating clearly undesirable compounds [52].

Once a set of reference compounds is available, one may start with actual modeling an activity landscape from these data. Fitness landscapes offer a modeling approach that assists synthetic chemists in decision-making and molecular design by visualizing and rationalizing structure–activity and structure–property relationships. A common theme and often a necessity are the transformation of raw data to a new coordinate system, where the axes of the new space represent ''factors'' or ''latent variables'' – features that might help explain the shape of the original data distribution. Fitness landscapes and their visualization have been a research topic in computational medicinal chemistry for approximately two decades [53]. For example, *principal component analysis* (PCA) [54] and *projection to latent structures* (PLS) [55] yield linear, statistically interpretable SAR models and data projections from typically high-dimensional descriptor spaces. The underlying mathematical models and the solutions provided by nonlinear projection are often more accurate, but at the same time evade immediate chemical interpretation.

Despite this apparent drawback, nonlinear projection techniques such as the *self-organizing map* (SOM, Kohonen network) [56], *multidimensional scaling* (MDS) [57], and *stochastic proximity embedding* (SPE) [58] – to name just some of the most prominent approaches – have demonstrated their particular usefulness for fitness landscape modeling. Their appeal lies in the ability to appropriately reflect the typically nonlinear dependencies between a structural (constitution-, topology-, or conformation-based) molecular representation and some measured bioactivity or property. Visualization of fitness landscapes at various project stages provides a selection criterion that is based not only on the actives found so far, but equally accounts for the inactive compounds.

In a proof-of-concept study, researchers at ETH and Roche pursued an advanced approach to fitness landscape visualization that results in easily interpretable biological response surfaces in chemical space (LiSARD, *ligand-induced structure–activity relationship display*) [59]. The LiSARD algorithm generates interactive graphics that can be used as intuitive roadmaps for molecular design and optimization. As a first practical application, they analyzed human somatostatin receptor subtype 5 receptor (hSST5R) antagonists. This class-A G-protein-coupled receptor is involved in several physiological processes, for example, *N*-methyl-D-aspartate (NMDA) receptor activation and control of hormonal secretion [60]. In a chemogenomics study aimed at finding nonpeptide hSST5R antagonists, approximately 3000 compounds of which the majority belonged to four structural classes were synthesized and tested [61]. Figure 1.16 demonstrates the adaptive nature of the corresponding landscape models. Depending on the number of compounds synthesized and tested, increasingly more fine-grained landscape models are obtained. In the example, using two-thirds of the data, the final shape of the landscape is clearly visible. Keep in mind that even the first approximate landscape model computed from only 100 compounds correctly structures chemical space into desired (blue) and ''tabu'' regions (red). Having access to such knowledge at an early project stage provides valuable information for hit prioritization and helps focus on relevant areas in chemical space so that optimized lead structures can be identified faster. Monitoring the SAR landscape over project duration certainly is a desirable feature for medicinal chemists to explore innovative structural variations of a chemotype, avoid walking in circles, and escape areas with potential *off*-target liabilities. In fact, multiple activities and properties can be displayed simultaneously in fitness landscapes, thereby enabling multidimensional optimization with the aim to avoid compounds that have an undesired pharmacological activity and property profile. Figure 1.17 shows such landscapes that were obtained from combining the individual landscapes for *on*- and *off*-targets. Such guidelines for ''polypharmacological'' design consider multiple targets (or properties) simultaneously. The *de novo* design process will aim at generating molecules that occupy the regions of predicted high activity without the need for separate fitness functions.

Avoiding undesired properties or regions in a fitness landscape is referred to as *negative design*, whereas ''positive design'' describes the attempt to engineer molecules that exhibit a desired property or function. These terms were originally coined by Richardson for the field of protein engineering and design but have now

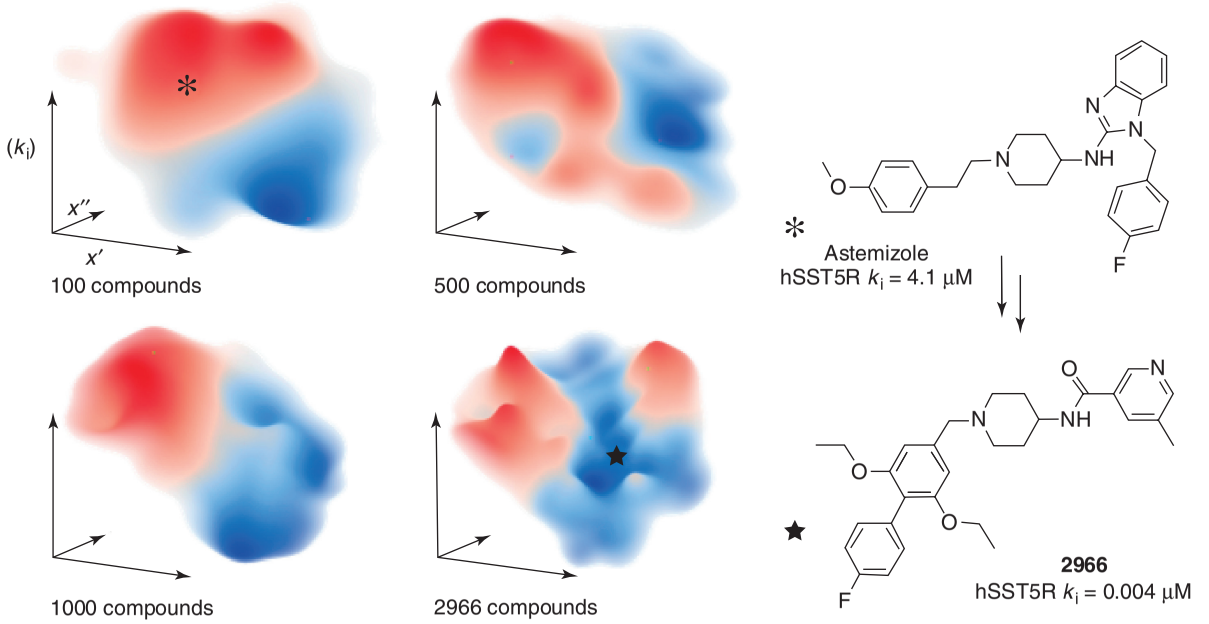


Figure 1.16 Adaptive evolution of the structure–activity landscape for hSST5R agonists over project time. The snapshots contain increasing levels of detail that can be captured depending on the available number of compounds synthesized and tested. Note that active and inactive compounds contribute equally to the model (*blue* = low K_i, *red* = high K_i). The project started with the reference astemizole as a template and ended with compound 2966. The asterisk and star symbols show the location of the two compounds in the fitness landscape. Landscapes were computed with the software LiSARD. Compounds were represented by the CATS pharmacophore descriptor and projected using stochastic neighbor embedding (SNE).

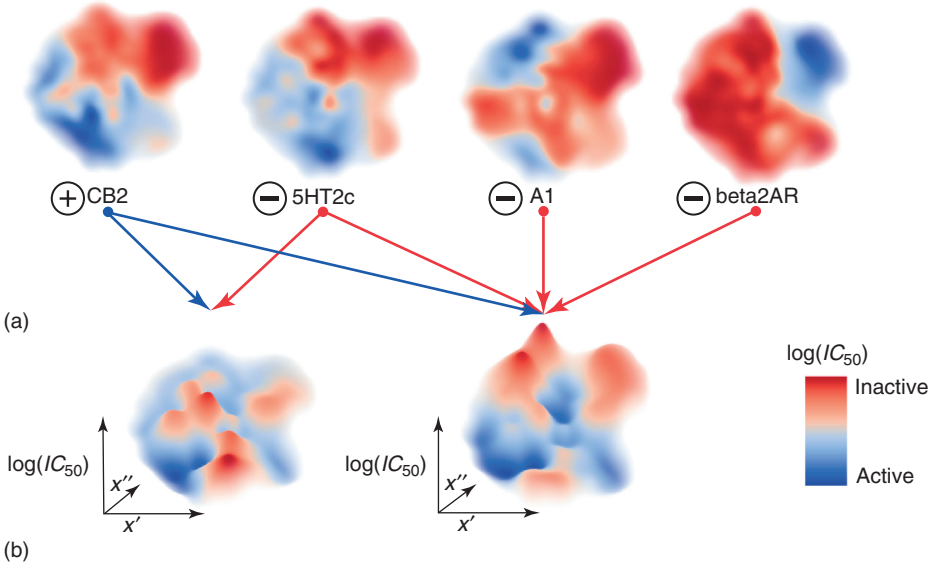


Figure 1.17 Examples of "polypharmacological" fitness landscapes. Two such landscapes (b) resulted from combining different target-specific landscapes (a). The individual landscapes were modeled using activity data for selected GPCRs (CB2, cannabinoid receptor 2; 5HT2c, serotonin receptor 2c; A1, adenosine receptor 1; and beta2AR, beta-2 adrenergic receptor). Plus signs designate desired target activity (*on*-targets), and minus signs indicate *off*-targets. Arbitrary mixtures of activity landscapes are possible, so that the design tasks can be combined in a single fitness landscape.

entered the world of small-molecule design [62]. After first-pass filtering of candidate compounds to eliminate the bulk of unwanted molecules (negative design), we can apply target-specific focused design. This can be done on the fly during the actual compound construction process, or post hoc by evaluating the *in silico* generated compounds. Automated compound classification and scoring enables rapid computational compound processing. The process requires appropriate predictive functions that perform pattern recognition and feature extraction.

A straightforward filtering routine that is specific for the molecular design task at hand, uses binary classifiers solving two-class problems for elimination of potentially unwanted molecules from a compound library or enrich a library with molecules predicted to reveal some kind of desired activity. The basic idea is to define two classes of compounds, one sharing a desired property (the *positive* set) and another lacking this property (the *negative* set). Consequently, a binary classifier is obtained for rapid first-pass compound scoring. Currently, four classifier systems are most often used in these applications: the naïve Bayes approach, feedforward artificial neural networks, support vector machines (SVMs), and Gaussian process (GP) models. These methods originate from the field of machine learning and virtual screening [63–65], which has massive impact on molecular *de novo* design methodology and enabled adaptive fitness landscape

modeling [66]. It is convenient to formulate these classifiers in terms of adaptive learning machines that improve with additionally available data. During first stage, the learning machine is presented with labeled samples, which are basically *n*-dimensional vectors with a class membership label attached (e.g., ''active'' = 1 and ''inactive'' = −1). The learning machine generates a classifier function for prediction of the class label assigned to the input coordinates (pattern). During the second stage, the generalization ability of the model is tested. Numerous performance indices have been suggested to obtain a realistic estimate of the prospective model accuracy [67]. It is common to use Matthews' correlations coefficient [68], the receiver-operator characteristic (ROC), area under the curve (AUC) [69], and the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) metric for this purpose [70]. Considering the pitfalls of such performance indices [71], it might not always be wise to employ the ''best-performing'' model owing to well-known issue arising from overfitting and erroneous estimation of a model's applicability domain, that is, its portion of chemical space for which reliable predictions can be made [72, 73]. Irrespective of the scoring strategy chosen, we must not forget that our current understanding of the physical forces governing ligand–receptor interaction is incomplete, and gaining a decimal point in computational precision is meaningless if the underlying model does not translate into compounds with improved activity.

1.4.1 Naïve Bayes Classifier

The naïve Bayes classifier is a fast and simple yet surprisingly effective classification algorithm with numerous applications in virtual screening and molecular design [74]. It is based on the assumption of conditional independence of features. The basic idea behind the naive Bayes classifier is *Bayes theorem*: Let C and X denote two events, then $P(C|X) = P(X|C)P(C)/P(X)$. If we define C as the hypothesis (target class) and X with the data (compound to be classified), the previous equation relates the *posterior* probability $P(C|X)$ of the hypothesis given the data to the *prior* probability $P(C)$, the probability of the data given the hypothesis $P(X|C)$, and the probability of the data $P(X)$. In other words, $P(C|X)$ is our ''belief'' in the hypothesis after we have seen the data, given the data, and our prior ''belief.'' This can be paraphrased as ''posterior = likelihood × prior/evidence'' [75]. In the *Bayesian* interpretation, ''probability'' indicates the degree of personal belief in a proposition, in contrast to the *frequentist* interpretation of probability as the relative frequency of occurrence of an event. A characteristic of the Bayesian approach is the concept of *prior* and *posterior* probability distributions, which measure what is known about a variable before and after the data have been considered. The naïve Bayes classifier is cheap to train and evaluate. Importantly, it also allows the addition of training samples later on by adjusting the relative frequencies used to estimate the probabilities. Another advantage is that it computes *probabilities* instead of plain ''predictions.'' Thereby, one also obtains a measure of *confidence* into the prediction. Although the assumption of conditional feature independence

is normally not valid in practice, naïve Bayes classifiers often work well anyway, even in high dimensions. This is owed to the fact that the order of magnitude of the probabilities is more important for classification than their exact values – as long as the dependencies between the molecular descriptors within a class are not too strong, the naïve Bayes classifier will perform reasonably well. Another effect of the conditional independence assumption is that redundant descriptors will have greater influence on the prediction, reducing performance. This caveat needs to be considered and can be alleviated by feature selection. Hopkins and coworkers [76] demonstrated a recent application of the approach for the polypharmacological *de novo* design of G-protein-coupled receptor (GPCR) ligands (cf. Chapter 12).

1.4.2 Artificial Neural Network

Feedforward neural networks (also called *multilayer Perceptron*s, MLPs) are a type of artificial neural networks that have found widespread application in virtual screening and *de novo* molecular design [77] and were pioneered in the field of chemistry by Gasteiger and coworkers [78]. They are *universal function approximators* modeled loosely after biological nervous systems [79]. The design of MLPs follows concepts of natural nervous systems such as neurons, axons, dendrites, and parallel information processing. An MLP is made up of basic units called *neurons*, which are organized into *layers* (Figure 1.18). An *input layer*, where each input neuron represents a descriptor, one or more *hidden layers* and an *output layer*, which produces the prediction result. The neurons of different layers are fully connected to each other. Depending on the allowed connections, one differentiates between *feedforward networks*, which correspond to acyclic directed graphs, and *recurrent networks*, where cycles are allowed. Numerous variations of this principle have been conceived over the past two decades. Among them, *associative* networks, that is, combinations of an ensemble of feedforward MLPs and the *k*-nearest neighbor technique, might be particularly useful for drug discovery and design [80]. Another idea is to combine several MLPs (also in combination with other prediction models) by a jury network (Figure 1.18). Such a cascaded machine-learning model has the advantage to often be more robust than the individual first-stage models.

1.4.3 Support Vector Machine

The SVM belongs to a class of machine-learning algorithms for classification and regression that are based on the ''kernel trick'' [81]. The latter is a general method, which allows algorithms that can be formulated in terms of *inner products* (also dot product, scalar product) to be systematically extended to nonlinear cases. For example, kernel PCA is such a nonlinear descendant of standard linear PCA [82]. In kernel-based classifier learning, a hyperplane, which optimally separates the training samples in a (nonlinearly) transformed space defined by the kernel

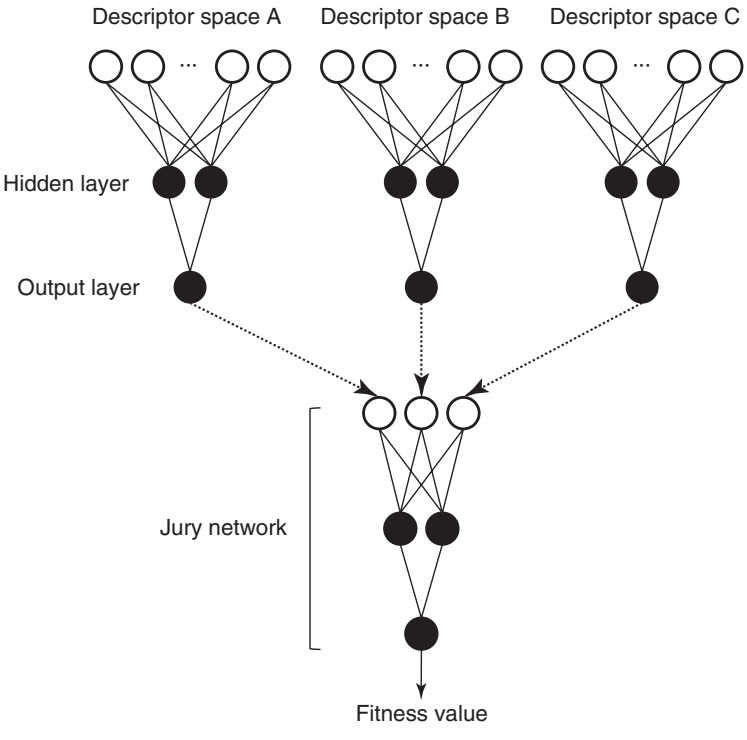


Figure 1.18 Cascaded neural network model that combines three first-stage models, whose output values are fed into a jury network. In this example, three different molecular representations (descriptor spaces A, B, and C) are fed into separate feedforward neural networks, so that the jury prediction ("fitness value") represents a weighted model of these descriptor worlds.

function, is found by solving a convex quadratic optimization problem. Of note, the solution depends only on a subset of the training samples, the *support vectors*, which define the optimal hyperplane, that is, the one with maximal margin (Figure 1.19). SVMs have received considerable attention in drug discovery and design [83], mainly because of solid theoretical foundations as well as good and robust performance in practice. They offer a method of choice for classification, as compound datasets are often not linearly separable in the chosen descriptor space, because either the molecular representation does not provide appropriate information or the problem is ill-posed [84].

1.4.4 Gaussian Process

GP models originate from Bayesian statistics and have only recently been added to the molecular designer's toolbox. Their first applications in molecular modeling

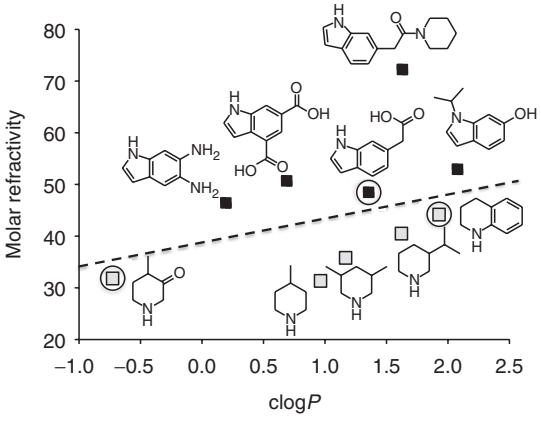


Figure 1.19 Linear separation of two classes of molecules (indoles, filled squares; piperidines, open squares) in a two-dimensional descriptor space (clog *P*, molar refractivity). The dashed line is the optimal hyperplane, with corresponding support vectors highlighted (circled). SVMs implicitly work in *very* high-dimensional spaces that allow for a linear separation of data that are nonlinearly related in the original descriptor space (kernel trick). Note that in the simplifying example shown the computation was done on standardized descriptors; otherwise, the difference in scale between the values of the two descriptors would have given more weight to molar refractivity, leading to a different hyperplane and support vectors. This hyperplane would have also separated the two classes, but with a worse generalization performance on new samples.

were quantitative structure–activity relationship (QSAR) regression models aimed at predicting aqueous solubility [85], blood–brain barrier penetration [86], and hERG (human ether-á-go-go related gene) inhibition [87]. Meanwhile, they represent a method of choice if quantitative property and activity predictions are required [88], with continuously increasing numbers of applications in drug design. For example, compound **4** (Figure 1.20) was identified as a subtype-selective agonist of transcription factor peroxisome-proliferator-activated receptor-gamma (PPARγ) by a predictive GP model that was trained with known PPAR ligands [89]. A particular advantage of GPs is that they provide error estimates with their predictions [90]. In GP modeling of molecular properties, one defines a kernel function to model molecular similarity. Compound information enters GP models only via this function. This is done by computing molecular descriptors (physicochemical property vectors) or by graph kernels that are defined directly on the molecular graph. From a family of functions that are potentially able to model the underlying SAR (prior), only functions that agree with the data are retained. The weighted average of the retained functions (posterior) acts as predictor, and its variance as an estimate of the confidence in the prediction (Figure 1.20). Variance is small near reference data, that is, for molecules similar to known ligands, and increases with growing distance. Importantly, the predictions and confidence estimates can be calculated analytically.

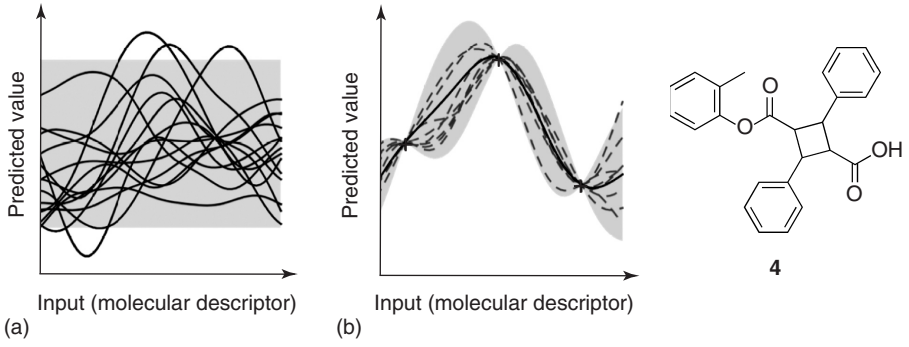


Figure 1.20 (a) Nonlinear Bayesian regression with Gaussian processes starts with a family of functions that map input data to activity (predicted value). (b) This prior is then combined with measured data (crosses). Only functions close to the observed data are retained. Averaging over the remaining functions yields the final predictor (solid line) and its variance (shaded area) as confidence estimate (domain of applicability). Compound **4** was identified as a subtype-selective PPARγ agonist using a Gaussian process model.

1.5 Strategies for Compound Construction

While virtual screening of large compound collection may be used for finding active compounds and, to a limited extent, performs scaffold hopping from known drugs or other bioactive reference compounds, a structure generator is required to conceive of innovative molecules that have not been synthesized or suggested before [72]. Ideally, one would perform an exhaustive enumeration of all possible molecular structures with a certain number of non-hydrogen atoms. The most common elements in drugs are C, N, O, P, and S (Table 1.2), and their coordination is well known. In fact, Reymond and coworkers compiled a large collection of virtual molecule structures containing up to 17 non-hydrogen atoms following this idea [92, 93]. The resulting "chemical universe databases" (GDB) contain up to 166 billion organic molecules, of which more than 99% have never been synthesized. This observation clearly demonstrates that there is ample opportunity for drug design. Further advances in computer hardware technology and distributed computation will undoubtedly facilitate complete analyses of such huge numbers of molecules. The GDB approach follows an atom-based construction strategy. *De novo* design methods additionally rely on fragment-based compound assembly and construct molecules on the fly, rather than precompiling a database for virtual screening. Both approaches have obvious advantages and drawbacks, which are discussed in depth in Chapters 5, 6, 10, 13–15, and 17. Probably the most compelling arguments for using fragments for molecular design are their inherent biomorphic qualities (fragments bind macromolecular targets with a high LE), and the possibility to directly employ fragments as synthons for organic synthesis (reaction-driven *de novo* design; cf. Chapter 10). Thereby, a design algorithm suggests not only new bioactive compounds but also a motivated synthesis pathway, and has access to a

Table 1.2 Relative occurrence of element types in 12 647 drugs and pharmaceutical lead compounds (COBRA (Collection of Bioactive Reference Analogs) database v12.6) [91].

H	C	O	N	F	S	Cl	P	Other	Σ
46.2%	40.0%	6.2%	5.6%	0.8%	0.6%	0.4%	0.1%	0.1%	100%

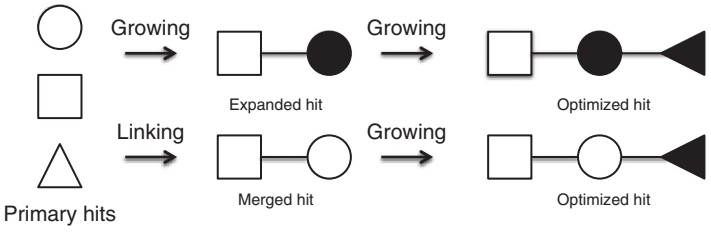


Figure 1.21 Strategies for fragment growing and linking. (Adapted from Ref. [95].) Open symbols indicate experimentally determined or computationally suggested start fragments (primary hits), filled symbols are newly added fragments, and lines represent linker moieties.

significantly larger chemical space than enumerated compound databases. It is generally assumed that fragment-based approaches, in contrast to strictly atom-based construction methods, offer a shortcut to generate new ligands in a chemically more meaningful way and also dramatically reduce the *total* size of the search space. If fragments commonly occurring in drugs are used for molecular assembly, the designed compounds will have a chance of being druglike, chemically stable, and synthetically feasible [94]. The most frequently applied fragment assembly strategies are growing and linking (Figure 1.21).

A general assumption of fragment-based compound assembly is that fragment contributions to the ligand binding energy are additive [96]. This principle approximately holds true if the binding mode and orientation of the individual fragments are only marginally perturbed in the *de novo* generated product. Still, there are numerous reports of unexpected ligand binding modes and nonadditive, that is, nonlinear, fragment contributions to ΔG [97, 98]. Nonadditivity can result in strong discrepancies between the computed sum of fragment affinities, A + B, and the affinity of the ligated compound AB [99, 100]. Such an example is presented in Figure 1.22 [101]. Here, the experimentally determined free energy of binding ΔG of the product AB, a potent factor Xa inhibitor ($K_i = 2$ nM), exceeds the sum of the individual fragment contributions by $-14\,\mathrm{kJ} \times \mathrm{mol}^{-1}$, although only a single bond was used as a fragment linker. For further information on rigorous free-energy-based molecular design, see Chapter 16.

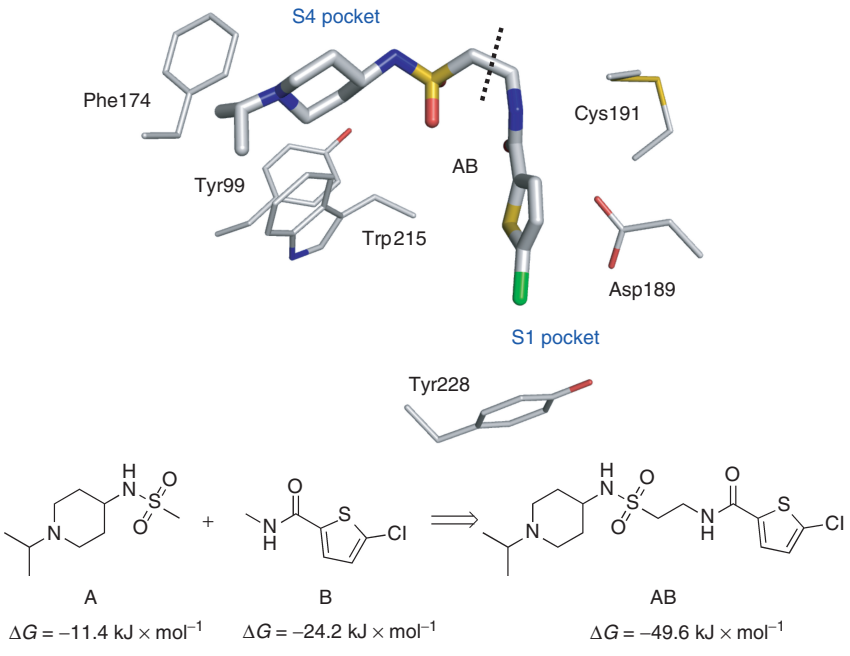


Figure 1.22 Fragment superadditivity: the experimentally determined free energy of binding ΔG of the product AB, a potent factor Xa inhibitor ($K_i = 2$ nM), exceeds the sum of the individual fragment contributions, A + B, by $-14\ \mathrm{kJ} \times \mathrm{mol}^{-1}$. A single bond was added as linker (dashed line). The cartoon shows the enzyme–ligand complex (PDB ID: 4a7i). Interacting residue side chains are highlighted.

As a shortcut to obtain reliable estimations of binding energies during fragment linking, one might explicitly account for the effect of linker elements during the construction process. Possibly though, the evaluation of the full ligand product rather than fragments only is obligatory. Despite the great appeal of fast algorithmic solutions for fragment-based exhaustive or *global* combinatorial product evaluation that implement the additivity principle, the actual practical applicability of these techniques can hardly be assessed *a priori*. As a workaround, both stochastic and deterministic *local* optimization strategies that score the full product have become a frequently pursued molecular design strategy. It is fair to say that the compound assembly task in *de novo* design can be regarded as solved. Even if these techniques do not guarantee finding the globally (computed) best solution in chemical space, they still identify new bioactive ligands with good success rates. In fact, an adaptive trade-off between conservative and exploratory designs can be helpful in hit identification. For example, the ligand-based fragment-growing tool DOGS (*design of genuine structures*) can be tuned to a desired ratio of scaffold exploitation/exploration during candidate compound assembly (cf. Chapter 10) [102]. Similarly, evolutionary

design algorithms are easily adaptable to produce desired scaffold diversity [103, 104].

Only a small fraction of all molecules amenable to virtual construction can in fact be synthesized in a reasonable time frame and with acceptable effort. *De novo* design programs tackle this issue by employing rules to guide the assembly process. Such rules attempt to reflect chemical knowledge and thereby avoid the formation of implausible or unstable structures. For example, some assembly approaches prevent connections between certain atom types, and finally the formation of unwanted substructures [105, 106]. Other strategies employ chemistry-driven retrosynthetic rules capturing general principles of reaction classes. A prominent example of this kind of rule set is RECAP (*retrosynthetic combinatorial analysis procedure*) [107], which is also employed by some *de novo* design tools. An early example is DREAM++ conceived by Kuntz and coworkers [108]. The software SYNOPSIS (*synthesize and optimize system in silico*) [109] follows a conceptually even more elaborate approach by connecting available molecular building blocks using a set of known chemical reactions. This enables the software to suggest reasonable synthesis pathways along with each final compound. Instead of accounting for synthetic accessibility by explicit reaction-based compound construction, one can also rely on *post hoc* synthesis planning with software such as CAESA [110], SYLVIA [111], or RouteDesigner [112] to come up with synthesis plans for *de novo* generated compounds.

1.6 Strategies for Compound Scoring

The first *de novo* design programs were exclusively based on *receptor-based* (also referred to as *structure-based*) scoring approaches, by which the quality of a designed compound is assessed by evaluating their potential to interact with a binding site on the receptor surface (cf. Chapter 4). Consequently, this idea is limited to target proteins for which there is a 3D structural model. Receptor-based tools were soon augmented by the development of *ligand-based* scoring schemes to circumvent this shortcoming (Table 1.3). Another motivation was the realization that – except for highly constrained systems – 3D *de novo* design *"in situ"* was too intractable at that time, because of the high computational costs of conformer generation and the attempt to explicitly consider synthetic tractability. The latter was soon partially achieved by the use of straightforward rules for fragment-based building block assembly (cf. Chapters 5, 6, and 10). While receptor-based scoring relies on the concept of ligand-pocket complementarity, ligand-based scoring schemes assess the similarity (or distance) to known reference ligands (templates) that exhibit the desired biological activity. Compounds designed under the objective to show high similarity to the reference are expected to have an increased probability to exhibit similar pharmacological properties.

Table 1.3 Chronological overview of selected *de novo* drug design software (a software name is given, otherwise the name of the first author) and the implemented compound scoring strategy.

***De novo* design method**	**Year of publication**	**Scoring strategy**	
		Receptor-based	**Ligand-based**
HSITE/2D skeletons [113]	1989	X	—
3D skeletons [114]	1990	X	—
Builder v1 [115]	1992	X	—
LUDI [116, 117]	1992	X	—
NEWLEAD [118]	1993	X	—
SPLICE [119]	1993	X	—
GroupBuild [120]	1993	X	—
CONCEPTS [121]	1993	X	—
SPROUT [122]	1993	X	—
MCSS and HOOK [123]	1994	X	—
GrowMol [124]	1994	X	—
Chemical Genesis [125]	1995	X	X
PRO_LIGAND [126]	1995	X	X
SMoG [127]	1996	X	—
CONCERTS [128]	1996	X	—
RASSE [129]	1996	X	—
PRO_SELECT [130]	1997	X	—
Skelgen [131]	1997	X	X
Nachbar [132]	1998	—	X
Globus [133]	1999	—	X
DycoBlock [134]	1999	X	—
LEA [135]	2000	—	X
LigBuilder [136]	2000	X	—
TOPAS [90]	2000	—	X
F-DycoBlock [137]	2001	X	—
ADAPT [138]	2001	X	—
Pellegrini and Field [139]	2003	X	X
SYNOPSIS [101]	2003	X	—
CoG [140]	2004	—	X
BREED [141]	2004	—	X
Nikitin [142]	2005	X	—
LEA3D [143]	2005	X	—
Flux [144]	2006	—	X
FlexNovo [145]	2006	X	—
Feher [146]	2008	—	X
GANDI [147]	2008	X	X
COLIBREE [105]	2008	—	X
SQUIRRELnovo [148]	2009	—	X
Hecht and Fogel [149]	2009	X	X
FOG [106]	2009	—	X
MED-hybridize [150]	2009	X	—
MEGA [151]	2009	X	X

Table 1.3 (*Continued*)

De novo design method	Year of publication	Scoring strategy	
		Receptor-based	Ligand-based
Fragment-shuffling [152]	2009	X	X
AutoGrow [153]	2009	X	—
NovoFLAP [154]	2010	—	X
PhDD [155]	2010	—	X
GARLig [156]	2010	X	—
DOGS [94]	2010	—	X
White and Wilson [157]	2010	—	X
Qsearch [158]	2011	—	X
EvoMD [159]	2011	—	X
Contour [102]	2012	X	—
MOEA [160]	2013	—	X
Ulrich [161]	2013	X	—

(Adapted from Ref. [162].)

1.6.1 Receptor-Based Scoring

Receptor-based approaches are closely related to computational strategies for automated receptor–ligand docking [163–165]. Both techniques share the objective to maximize the complementarity of the ligand with the binding site regarding shape and properties. Although rigorous free energy calculations have become feasible for drug design (cf. Chapter 16) [166], the most common approaches to estimate the quality of binding during the design process are the same as for molecular docking, where three main strategies have emerged [167]:

1) physically motivated force fields,
2) empirical scoring functions, and
3) knowledge-based scoring functions.

Simplistic physical force fields treat molecules as ensembles of spheres (atoms) connected by springs (bonds). Each spring has optimal values for length, torsion angles, and angles to other springs. Deviation from these optimal values results in strain, and accordingly, low strain energies correspond to favorable ligand conformations. Ligand interaction with the receptor is estimated by terms for nonbonded interactions; most commonly by Coulomb and vdW potentials, sometimes augmented by an explicit term for contributions of hydrogen bridges. A generalized force-field term for nonbonding interaction computes the interaction energy E between a ligand and a receptor for a given ligand–receptor complex (binding pose) (Eq. (1.10)).

$$E = \sum_{i=1}^{\text{ligand}} \sum_{j=1}^{\text{receptor}} \left[\frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right] \tag{1.10}$$

where A_{ij} and B_{ij} are parameters expressing repulsion and attraction of vdW interactions of atoms i and j at distance r_{ij}, q is the atomic point charge, and ε is the dielectric constant of the solvent. Despite its great relevance for the computation of accurate energies, the latter is difficult to assess for ligand binding pockets and represents a major source of error for force-field-based scoring in *de novo* design.

Empirical scoring functions are expressed as weighted sums of several contributing terms, where weights are determined by regression analysis. Weights are optimized in order to reproduce experimentally measured activity values (e.g., pIC_{50}, pK_d) of known ligand–receptor complexes. The individual terms represent different ligand–receptor interactions, which can be determined from a given binding pose. The free energy of binding is calculated as presented in Eq. (1.11):

$$\Delta G = \Delta G^0 \sum_{i=1}^{\#\text{interaction type}} \left[\Delta G_i \cdot \text{count}_i \cdot \text{penalty}_i \right] \tag{1.11}$$

where ΔG_i represents the contribution (adjusted weight) of interaction type i, count_i is the number of times this interaction type is observed in the given receptor–ligand complex, and penalty_i is a penalty function accounting for deviations from the ideal interaction geometries for some interaction types such as hydrogen bridges or aromatic interactions. The penalty must be determined for each observed interaction type. ΔG^0 is a ground term that is also adjusted during the fitting process.

Knowledge-based scoring functions rely on discrepancies between observable and expected distributions of atom pair occurrences. On the basis of the frequencies of atoms (or functional groups), one can calculate a background probability of the chance that two atoms (one from the receptor and the other from the ligand) are placed in a certain distance in a ''random'' ligand–receptor complex, given that they do not interact. This is compared to the counts of atom pairs observed in experimentally explored ligand–receptor complexes (training set) and finally transformed into interaction scores by an inverse formulation of the Boltzmann law [168, 169]. Atom pairs that occur more often than expected by chance result in negative interaction energies (attraction), whereas less frequently observed pairs score positive (repulsion). Ligand affinity in a given complex with a receptor is estimated by summing up individual scores of observed atom pairs derived from the training set. Equation (1.12) calculates the contribution of atom pairs between atom types i and j at distance r as the interaction energy of the ligand–receptor complex:

$$E(i,j) = -k_B T \ln \frac{p_{ij}^{\text{observed}}(r)}{p_{ij}^{\text{expected}}(r)} \tag{1.12}$$

where k_B is the Boltzmann constant, T is the absolute temperature, and p_{ij} are observed and background frequencies of atom pairs of type i and j at distance r. The total energy of binding is calculated as a sum of these terms for all pairs of atom types and a range of different distances.

1.6.2 Ligand-Based Scoring

In contrast to computing the complementarity of ligands with the binding site, ligand-based scoring schemes compare the *de novo* generated compounds to a template compound and compute a similarity index for the two molecules in some descriptor space. For compound comparison, one needs to select a model for compound representation (molecular descriptors) and a similarity metric. Some ligand-based *de novo* design programs rely on pharmacophore models for quality assessment (cf. Chapter 7). These methods compare molecules according to their spatial or topological arrangement of potential receptor–ligand interaction centers. Some tools employ pseudoreceptor techniques (cf. Chapter 9) and related techniques such as molecular field analysis (cf. Chapter 8) for scoring. These approaches calculate pharmacophoric and steric constraints of a hypothetical receptor pocket based on a 3D conformation of an active ligand or a ligand ensemble, and assess the score of a new compound by evaluating its complementarity to this virtual cavity model, thus forming a bridge between receptor- and ligand-based methods. Ligand-based scoring strategies can be based on either a single reference or an ensemble of known ligands. For example, a consensus pharmacophore model can be constructed from a multiple alignment of reference ligands.

1.7 Flashback Forward: A Brief History of *De Novo* Drug Design

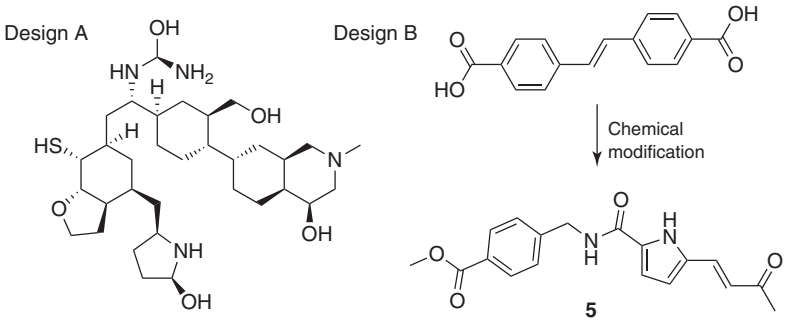
De novo design methods have been extensively reviewed during recent few years [170–180]. Here, we highlight some of the milestone developments. It goes without saying that our selection is subjectively biased, and we refer the reader to the literature for a broader overview of the field. For summaries of the state of the art of the computer-assisted design of proteins and nucleic acids, as potential drugs of the future, see Chapters 18–21 of this volume.

With the first structure-based *de novo* drug design study published in 1976 [181, 182], the whole game became professional approximately 25 years ago when the first computer applications were conceived for the purpose of fully automated molecular design [183–186]. At the time, the most prominent pioneering tools were ALADDIN [187], CAVEAT [188, 189], GENOA [190], and DYLOMMS [191]. Innovative scoring techniques, such as GRID [192], MCSS [193], DOCK [194], and CoMFA [14], for ligand–receptor affinity fostered this development. In the

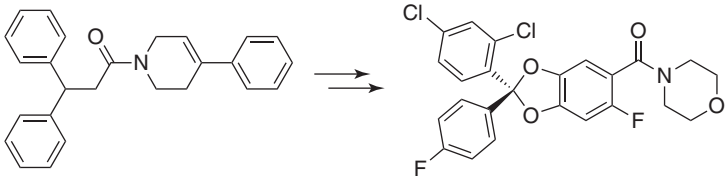
1990s, *de novo* design prospered for the first time resulting in groundbreaking applications [195–199] and algorithms – expressly GROW [200], GrowMol [124], LEGEND [116, 117], and LUDI [201, 202] representing some of the key players. In a seminal pioneering study from 1991 [24], Moon and Howe argued that "*Given detailed structural knowledge of the target receptor, it should be possible to construct a model of a potential ligand, by algorithmic connection of small molecular fragments, that will exhibit the desired structural and electrostatic complementarity with the receptor*." At the time, searching the space of candidate compounds was considered the most critical issue of the whole design process – compared to today the available computer hardware was rather limited. Molecular fragments as building blocks were primarily used to obtain a manageable search space. Peptides and peptide mimetics were a preferred molecule class for exploration by design. Currently, we are witnessing a strong renewed interest in peptide and protein *de novo* design (cf. Chapters 18 and 19), driven on the one hand by the realization that peptides actually represent ideal drug candidates and superb chemical probes, and on the other hand by modern chemical tricks that improve their pharmacokinetic and pharmacodynamic properties [203–205].

The early design studies typically relied on static X-ray structures providing the essential structural and pharmacophoric feature constraints for *in situ* ligand assembly. Evidently, rigid models of ligand–accommodating receptor cavities cannot account for induced- or flexible-fit phenomena that may be observed on fragment binding, which certainly has contributed to a somewhat limited enthusiasm and acceptance of *de novo* design by the medicinal chemistry community at the time. Some of the current molecular design tools explicitly allow for molecular flexibility, albeit sometimes at the price of strongly increased needs for computation time.

With the advent of reaction-driven compound fragmentation and assembly techniques as well as fast substructure-based prediction of "complexity," the issue of synthetic feasibility has been partially addressed (cf. Chapter 2). Despite several convincing applications, the accurate computer-based assessment of context-dependent building block reactivity still remains profoundly challenging – in particular when rapid estimations for high-throughput applications are mandatory like in *de novo* compound construction. The great importance of using a suitable set of fragments for virtual compound generation shall be highlighted exemplarily by two selected case studies. The first example describes the design of novel inhibitors of hepatitis C virus (HCV) helicase. Brancale and coworkers [206] equipped the receptor-based *de novo* design software LigBuilder [136] with two different sets of molecular building blocks, which resulted in the initial designs A and B, respectively. It is evident that the highly complex, unstable compound A is an attempt to fill the complete binding site, which most likely is a consequence of poor scoring as larger compounds often yield better scores. Design B – despite its nondrug-like structure – might be considered as a prototype ligand of HCV helicase, which was successfully converted into the chemically feasible inhibitor **5** ($IC_{50} = 260\,\text{nM}$).



Compounds **6** and **7** provide a second example of compound optimization from a *de novo* designed prototype to a potent lead structure. The software TOPAS produced a small series of structural suggestions that were further optimized as potent inverse agonists of cannabinoid receptor 1 (CB1) [207]. A single known reference compound served as a template for fragment-based virtual ligand assembly, guided by a topological pharmacophore model (CATS, *chemically advanced template search*) [208]. The initial design **6** had moderate activity ($K_i = 1500$ nM) but was chosen for subsequent optimization through iterative modeling, synthesis, and testing, which eventually led to the benzodioxole **7** ($K_i = 4$ nM) exhibiting desired *in vivo* efficacy [209].

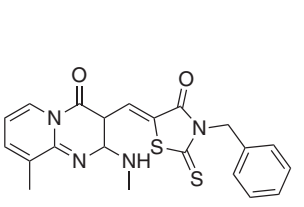


Original design **6** Optimized lead structure **7**

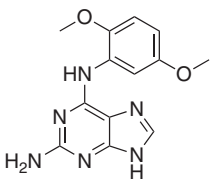
These selected examples confirm that profound chemical understanding is essential for successful application of computer-based *de novo* design tools. One cannot expect that these software tools deliver potent leads from scratch. Future drug design tools should incorporate as much medicinal chemistry knowledge as possible to facilitate candidate selection and increase their acceptance and utilization for drug discovery.

New algorithms, mainly stemming from the field of machine learning, as well as technological advances in computer sciences, for example, super computing, cloud computing, and GPU computing, have promoted a new wave of *de novo* design techniques [210]. Old software tools have not gone out of fashion, though. For example, the receptor-pocket-based LigBuilder software has recently been applied to come up with new inhibitors such as **8** ($IC_{50} = 6\,\mu$M) of eyes absent homolog 2 (EYA2) protein [211] and **9** ($IC_{50} = 0.4\,\mu$M) for VRAF murine sarcoma viral oncogene homolog B1 (BRAF) kinase [212]. A recent study employing the

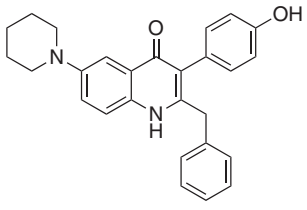
receptor-based design software LUDI yielded kinesin spindle protein (KSP) inhibitor **10** ($IC_{50} = 0.01\,\mu M$) as a potent and novel antimitotic lead [160].



8

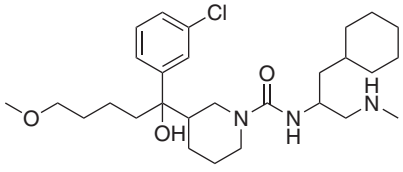


9



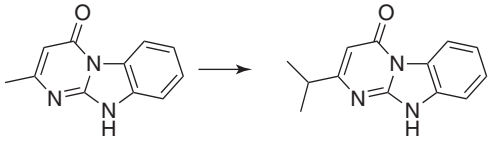
10

Despite the sustained success of some of the classic design tools, algorithm development has not ceased – quite the opposite! As an important advancement, the *de novo* design software Contour [213] includes solvation in the scoring process. Following the kernel trick of SVMs, its scoring function is able to model nonlinear functions using linear operations in a kernel-induced feature space. Compound construction by Contour employs an *in situ* combinatorial fragment-growing algorithm and was successfully applied to generate the new renin inhibitor **11** ($IC_{50} = 0.5\,nM$).



11

In a multiobjective design study, Ijzerman and coworkers [214] used their software tool MOEA (*multiobjective evolutionary algorithm*) to generate subtype-selective A_1 adenosine receptor antagonist **12** ($IC_{50} = 6\,\mu M$), which was further refined to obtain compound **13** ($IC_{50} = 0.3\,\mu M$). Again, kernel-based machine learning was used for the development of tailored scoring functions. SVMs based on molecular fingerprints for other adenosine receptor subtypes (hA_{2A}, hA_{2B}, and hA_3) served as negative objective functions, and in a combination with pharmacophore models, the molecular construction algorithm was steered toward the desired activity.



12 **13**

The previous example shows that known drug scaffolds can reemerge during *de novo* design. This observation provides an excellent starting point for drug

repurposing [215]. Recently, Schneider and coworkers employed their ligand-based software DOGS to come up with potent new type-II inhibitors of polo-like kinase1 (Plk1). Compound **14** ($IC_{50} = 0.2$ nM), which had been found previously by receptor-based virtual screening [216], served as design template, and by reaction-based fragment assembly, the software suggested compounds **15** and **16**, among other designs (Figure 1.23; cf. Chapter 10) [217]. Both molecules were readily amenable to chemical synthesis, following the one-step synthetic route (**15**: reductive amination; **16**: amide bond formation) suggested by the software without necessity for further optimization. While compound **15** represents a conservative design with a rather similar generic scaffold as the template, compound **16** features a remarkable scaffold hop (blue-colored graph structures in

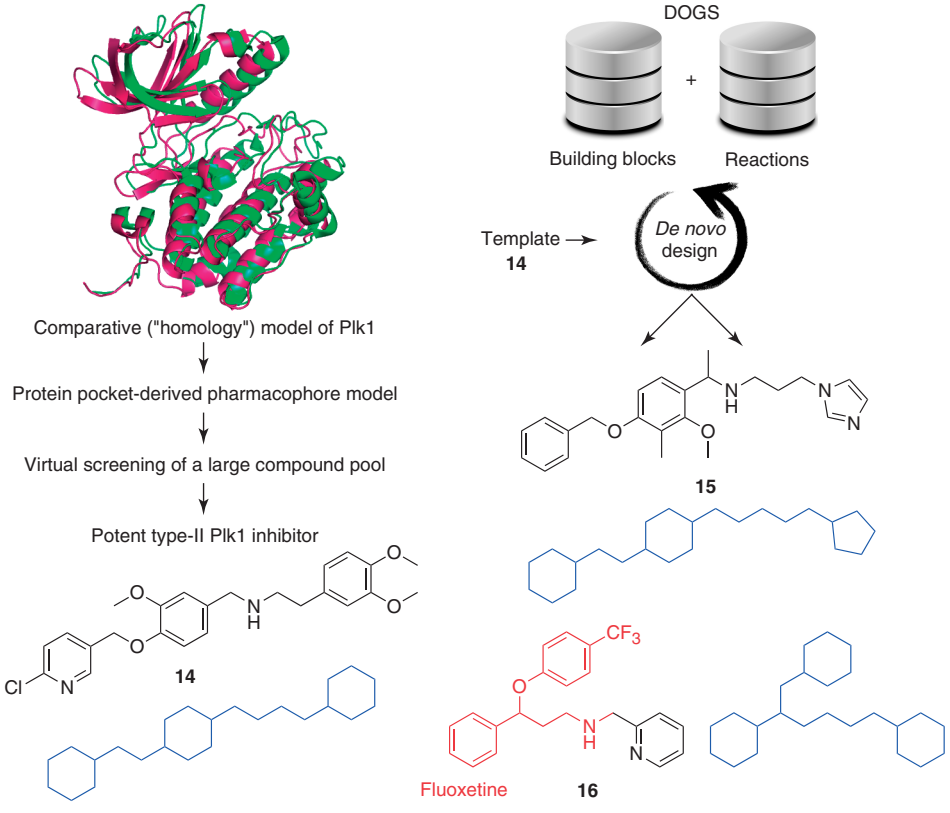
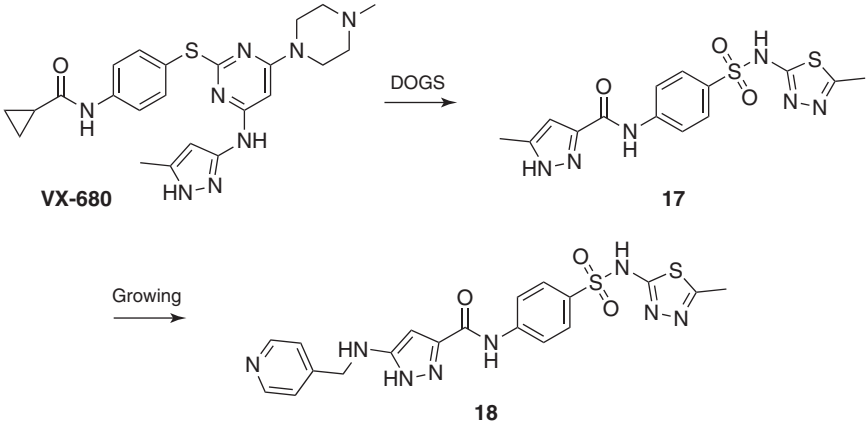


Figure 1.23 Receptor-based pharmacophore matching led to the discovery of compound **14**, a nanomolar type II inhibitor of human polo-like kinase 1 (Plk1). Taking this compound as template for ligand-based *de novo* design the software DOGS suggested the potent compounds **15** and **16**, which were synthesized as detailed by the software. The designed molecule **16** may be considered as a structural derivative of the antidepressant fluoxetine (red-colored substructure in compound **16**). Generic graph scaffolds of the compounds are shown in blue.

Figure 1.23). Both compounds exhibited low-nanomolar inhibitory activity against Plk1. Of note, the potential anticancer compound **16** induced significantly delayed cancer cell proliferation without affecting the vitality of nontransformed cells ($EC_{50} = 4\,\mu M$ against HeLa cells) and exhibited no inhibitory effects against a large panel of activated kinases [218]. Its LE is 0.66 (Eq. (1.8)). The computationally designed compound is a derivative of the antidepressant fluoxetine (Prozac®), for which the authors observed a similar but weaker cellular response profile.

In a similar study by the same authors, DOGS served for ligand-based *de novo* design to swiftly discover a new class of compounds efficiently blocking aurora A kinase (AurA) [219]. VX-680 is a *pan*-aurora inhibitor with high potency against AurA ($K_i = 0.6\,nM$), but its further development was abandoned because of observed QT interval prolongation issues in clinical trials. Thus, taking VX-680 as a template, the software DOGS suggested new chemotypes mimicking structural and pharmacophoric features of the template. The suggested designs were acquired by explicit scaffold hopping from the template. Compound **17** was synthesized and obtained in good yields following the synthesis pathway suggested by the software. Biochemical activity testing demonstrated moderate AurA inhibition by **17** ($IC_{50} \sim 10\,\mu M$). Molecule growing optimized this primary hit. Adding a molecular fragment resulted in compound **18**, which directly blocks AurA ($IC_{50} = 3\,\mu M$) and is potently active in cellular assays.



These representative studies provide proof of concept for *de novo* design as a premier tool for generating pioneering chemotypes in the absence of a structural model of the target protein and with minimal experimental effort needed. They also confirm the concept of reaction-driven, template-based *de novo* design as excellently suited for the rapid identification of novel bioactive molecules exhibiting a desired biological activity spectrum.

1.8 Conclusions

Current drug discovery is fueled by advanced high-throughput screening technology, fragment-based and parallel medicinal chemistry. Computer-based *de novo* design has only just begun to play a role in this game [170, 220, 221]. Recent substantial developments that enable *de novo* design in drug discovery are owed to reaction-driven compound assembly, multiobjective scoring, and fragment-based approaches. In the very near future, design software will be directly coupled to automated compound synthesizers, liquid handling robots, and microfluidic lab-on-a-chip systems. Once such a fully automated pipeline is realized, rapid feedback loops will become possible so that truly adaptive machine-learning and computer-based optimizations are performed. On the basis of the concept of "active learning," just another idea borrowed from the machine-learning field, autonomous robotic molecular design machines will support project teams in their attempt to find new medicines. Irrespective of such futuristic thoughts, tight cooperation between molecular designers, synthetic chemists, and biologists will remain indispensable for success. Already we are witnessing an increase in the number of *de novo* design applications that go all the way from the initial design via chemical synthesis to activity determination in both academic groups and pharmaceutical industry. The story does not end here but has reached out to larger, more complex molecules such as proteins and nucleic acids. So-called biologicals, biomimetics, and traditional small organic molecules alike will continue to provide ample opportunity for medicinal chemistry and chemical biology. Computer-assisted *de novo* drug design has the appealing advantage to be theoretically unlimited in compound diversity and intrinsically innovative. As soon as a surprising but reasonable suggestion of a candidate compound is made, *de novo* design has already fulfilled its purpose: to generate useful ideas and inspire, thereby providing an opportunity to overcome stalled drug discovery.

It will be most interesting to see how *de novo* design technology will develop during the next decade [222–224]. Structural novelty combined with synthetic feasibility might be more important for a *de novo* design than actual bioactivity, which can often be increased by means of medicinal chemistry [225]. In 1987, Sheridan *et al.*[226] wrote: "*Only a few novel bond 'frameworks' in which important pharmacophore atoms are held in the proper arrangement need to be found to suggest new areas for drug design and synthesis.*" This statement is true today as it was in the early days of computer-based drug design. The primary aim of *de novo* design tools is to fuel the creativity of chemists by making surprising and innovative suggestions.

Acknowledgments

The authors are most grateful to Petra Schneider, Tiago Rodrigues, Jan A. Hiss, Daniel Reker, Michael Reutlinger, and Nickolay Todoroff for inspiring discussion.

This work was supported by the ETH Zürich, the Swiss National Science Foundation (SNF grant no. 205321–134783), and the OPO-Foundation Zürich.

References

1. Sullivan, H.L. (1896) The Tall Office Building Artistically Considered. Lippincott's Monthly Magazine, issue 57, pp. 403–409, Citation taken from p. 408.
2. Dalton, J. (1842) *A New System of Chemical Philosophy*, John Weale, London, pp. 237–239.
3. Kaufmann, S. (1993) *Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, Oxford.
4. Holland, J.H. (1998) *Emergence — From Chaos to Order*, Oxford University Press, Oxford.
5. Freeman, W.J., Kozma, R., and Werbos, P.J. (2001) Biocomplexity: adaptive behavior in complex stochastic dynamical systems. *Biosystems*, **59**, 109–123.
6. Mashaal, M. (2006) *Bourbaki: A Secret Society of Mathematicians*, American Mathematical Society, Providence, RI.
7. Tillich, P. (1951) *Systematic Theology I*, University of Chicago Press, Chicago.
8. Todeschini, R. and Consonni, V. (2000) *Handbook of Molecular Descriptors*, Wiley-VCH Verlag GmbH, Weinheim.
9. King, R.D., Muggleton, S., Lewis, R.A., and Sternberg, M.J. (1992) Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 11322–11326.
10. Lee, B. and Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.
11. Connolly, M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.
12. Ballester, P.J. (2011) Ultrafast shape recognition: method and applications. *Future Med. Chem.*, **3**, 65–78.
13. Ebalunode, J.O. and Zheng, W. (2010) Molecular shape technologies in drug discovery: methods and applications. *Curr. Top. Med. Chem.*, **10**, 669–679.
14. Cramer, R. III, Patterson, D., and Bunce, J. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, **110**, 5959–5967.
15. Randic, M., Kleiner, A., and Alba, L.D. (1994) Distance/distance matrices. *J. Chem. Inf. Comput. Sci.*, **34**, 277–286.
16. Schuur, J. and Gasteiger, J. (1997) Infrared spectra simulation of substituted benzene derivatives on the basis of a 3D structure representation. *Anal. Chem.*, **69**, 2398–2405.
17. Robinson, D., Barlow, T., and Richards, G. (1997) The utilization of reduced dimensional representations of molecular structure for rapid molecular similarity calculations. *J. Chem. Inf. Comput. Sci.*, **37**, 943–950.
18. Hemmer, M., Steinhauer, V., and Gasteiger, J. (1999) Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.*, **19**, 151–164.
19. Gramatica, P., Corradi, M., and Consonni, V. (2000) Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. *Chemosphere*, **41**, 763–777.
20. Ballester, P.J. and Richards, W.G. (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.*, **28**, 1711–1723.
21. Grant, A., Gallardo, A., and Pickup, B. (1996) A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.*, **17**, 1653–1666.
22. Cruciani, G., Crivori, P., Carrupt, P.A., and Testa, B. (2000) Molecular fields in quantitative structurepermeation

relationships: the VolSurf approach. *J. Mol. Struct.*, **503**, 17–30.
23. Consonni, V., Todeschini, R., Pavan, M., and Gramatica, P. (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comput. Sci.*, **42**, 693–705.
24. Van Drie, J.H. (1997) ''Shrink-wrap'' surfaces: a new method for incorporating shape into pharmacophoric 3D database searching. *J. Chem. Inf. Comput. Sci.*, **37**, 38–42.
25. Ritchie, D. and Kemp, G. (1999) Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.*, **20**, 383–395.
26. Morris, R.J., Najmanovich, R.J., Kahraman, A., and Thornton, J.M. (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, **21**, 2347–2355.
27. Pérez-Nueno, V.I., Venkatraman, V., Mavridis, L., Clark, T., and Ritchie, D.W. (2011) Using spherical harmonic surface property representations for ligand-based virtual screening. *Mol. Inf.*, **30**, 151–159.
28. Jakobi, A.J., Mauser, H., and Clark, T. (2008) Parafrag–an approach for surface-based similarity comparison of molecular fragments. *J. Mol. Model.*, **14**, 547–558.
29. Rechenberg, I. (1973) *Evolutionsstrategie–Optimierung Technischer Systeme Nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, Stuttgart.
30. Johnson, M.A. and Maggiora, G.M. (eds) (1990) *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, Inc., New York.
31. Guha, R. (2011) The ups and downs of structure-activity landscapes. *Methods Mol. Biol.*, **672**, 101–117.
32. Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M.S., and Van Drie, J.H. (2009) Navigating structure-activity landscapes. *Drug Discov. Today*, **14**, 698–705.
33. Bongard, M.M. (1970) *Pattern Recognition*, Hayden Book Co., Spartan Books, Rochelle Park, NJ(Original publication: Проблема Узнавания, Nauka Press, Moscow, 1967).
34. Miller, J.H. and Page, S.E. (2007) *Complex Adaptive Systems - An Introduction to Computational Models of Social Life*, Princeton University Press, Pronceton, NJ, Oxford.
35. Koza, J.R. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, The MIT Press, Cambridge.
36. Böhm, H.J. and Schneider, G. (eds) (2003) *Protein-Ligand Interactions: From Molecular Recognition to Drug Design. Methods and Principles in Medicinal Chemistry*, Vol. 19 (eds R. Mannhold, H. Kubinyi, and G. Folkers), Wiley-VCH Verlag GmbH, Weinheim.
37. Desiraju, G.R. (2002) Hydrogen bridges in crystal engineering: interactions without borders. *Acc. Chem. Res.*, **35**, 565–573.
38. Butler, K.T., Luque, F.J., and Barril, X. (2009) Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem.*, **30**, 601–610.
39. Poehlsgaard, J., Harpsøe, K., Jørgensen, F.S., and Olsen, L. (2012) A robust force field based method for calculating conformational energies of charged drug-like molecules. *J. Chem. Inf. Model.*, **52**, 409–419.
40. Agrafiotis, D.K., Gibbs, A.C., Zhu, F., Izrailev, S., and Martin, E. (2007) Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.*, **47**, 1067–1086.
41. Foloppe, N. and Chen, I.J. (2009) Conformational sampling and energetics of drug-like molecules. *Curr. Med. Chem.*, **16**, 3381–3413.
42. Biela, A., Betz, M., Heine, A., and Klebe, G. (2012) Water makes the difference: rearrangement of water solvation layer triggers non-additivity of functional group contributions in protein-ligand binding. *ChemMedChem*, **7**, 1423–1434.

43. Dimova, D., Heikamp, K., Stumpfe, D., and Bajorath, J. (2013) Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets. *J. Med. Chem.*, **56**, 3339–3345.
44. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
45. Fischer, S. and Verma, C.S. (1999) Binding of buried structural water increases the flexibility of proteins. *Proc. Natl. Acad. Sci. U.S.A*, **96**, 9613–9615.
46. Bissantz, C., Kuhn, B., and Stahl, M. (2010) A medicinal chemist's guide to molecular interactions. *J. Med. Chem.*, **53**, 5061–5084.
47. Hann, M.M. and Keserü, G.M. (2012) Finding the sweet spot: the role of nature and nurture in medicinal chemistry. *Nat. Rev. Drug Discov.*, **11**, 355–365.
48. Ferenczy, G.G. and Keserü, G.M. (2010) Enthalpic efficiency of ligand binding. *J. Chem. Inf. Model.*, **50**, 1536–1541.
49. Freire, E. (2008) Do enthalpy and entropy distinguish first in class from best in class? *Drug Discov. Today*, **13**, 869–874.
50. Hopkins, A.L., Groom, C.R., and Alex, A. (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today*, **9**, 430–431.
51. Keserü, G.M. and Makara, G.M. (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.*, **8**, 203–212.
52. Walters, W.P. and Namchuk, M. (2003) Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.*, **2**, 259–266.
53. Reutlinger, M. and Schneider, G. (2012) Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph. Model.*, **34**, 108–117.
54. Pearson, K. (1901) *Philos. Mag.*, **2**, 559–572.
55. Ståhle, L. and Wold, S. (1986) On the use of some multivariate statistical methods in pharmacological research. *J. Pharmacol. Methods*, **16**, 91–110.
56. Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, **43**, 59–69.
57. Torgerson, W.S. (1952) Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**, 401–419.
58. Agrafiotis, D.K. and Xu, H. (2003) A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.*, **43**, 475–484.
59. Reutlinger, M., Guba, W., Martin, R.E., Alanine, A.I., Hoffmann, T., Klenner, A., Hiss, J.A., Schneider, P., and Schneider, G. (2011) Neighborhood-preserving visualization of adaptive structure-activity landscapes: application to drug discovery. *Angew. Chem. Int. Ed.*, **50**, 11633–11636.
60. Reisine, T. and Bell, G.I. (1995) Molecular biology of somatostatin receptors. *Endocrinol. Rev.*, **16**, 427–442.
61. Martin, R.E., Green, L.G., Guba, W., Kratochwil, N., and Christm, A. (2007) Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: a chemogenomics approach. *J. Med. Chem.*, **50**, 6291–6294.
62. Richardson, J.S., Richardson, D.C., Tweedy, N.B., Gernert, K.M., Quinn, T.P., Hecht, M.H., Erickson, B.W., Yan, Y., McClain, R.D., and Donlan, M.E. (1992) Looking at proteins: representations, folding, packing, and design. *Biophys. J.*, **63**, 1185–1209.
63. Gertrudes, J.C., Maltarollo, V.G., Silva, R.A., Oliveira, P.R., Honório, K.M., and da Silva, A.B. (2012) Machine learning techniques and drug design. *Curr. Med. Chem.*, **19**, 4289–4297.
64. Chen, B., Harrison, R.F., Papadatos, G., Willett, P., Wood, D.J., Lewell, X.Q., Greenidge, P., and Stiefl, N. (2007) Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput. Aided Mol. Des.*, **21**, 53–62.

65. Melville, J.L., Burke, E.K., and Hirst, J.D. (2009) Machine learning in virtual screening. *Comb. Chem. High Throughput Screening*, **12**, 332–343.
66. Schneider, G. and So, S.S. (2001) *Adaptive Systems in Drug Design*, Landes Bioscience, Austin, TX.
67. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., and Nielsen, H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
68. Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
69. Clark, R.D. and Webster-Clark, D.J. (2008) Managing bias in ROC curves. *J. Comput. Aided Mol. Des.*, **22**, 141–146.
70. Truchon, J.F. and Bayly, C.I. (2007) Evaluating virtual screening methods: good and bad metrics for the ''early recognition'' problem. *J. Chem. Inf. Model.*, **47**, 488–508.
71. Nicholls, A. (2011) What do we know? Simple statistical techniques that help. *Methods Mol. Biol.*, **672**, 531–581.
72. Funatsu, K., Miyao, T., and Arakawa, M. (2011) Systematic generation of chemical structures for rational drug design based on QSAR models. *Curr. Comput. Aided Drug Des.*, **7**, 1–9.
73. Roy, K. (2007) On some aspects of validation of predictive quantitative structure-activity relationship models. *Expert Opin. Drug Discov.*, **2**, 1567–1577.
74. Bender, A. (2011) Bayesian methods in virtual screening and chemical biology. *Methods Mol. Biol.*, **672**, 175–196.
75. Duda, R., Hart, P., and Stork, D. (2001) *Pattern Classification*, John Wiley & Sons, Inc., New York.
76. Besnard, J., Ruda, G.F., Setola, V., Abecassis, K., Rodriguiz, R.M., Huang, X.P., Norval, S., Sassano, M.F., Shin, A.I., Webster, L.A., Simeons, F.R., Stojanovski, L., Prat, A., Seidah, N.G., Constam, D.B., Bickerton, G.R., Read, K.D., Wetsel, W.C., Gilbert, I.H., Roth, B.L., and Hopkins, A.L. (2012) Automated design of ligands to polypharmacological profiles. *Nature*, **492**, 215–220.
77. Zou, J., Han, Y., and So, S.S. (2008) Overview of artificial neural networks. *Methods Mol. Biol.*, **458**, 15–23.
78. Zupan, J. and Gasteiger, J. (1999) *Neural Networks in Chemistry and Drug Design–an Introduction*, Wiley-VCH Verlag GmbH, Weinheim.
79. Bishop, C. (1996) *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.
80. Tetko, I.V. (2008) Associative neural network. *Methods Mol. Biol.*, **458**, 185–202.
81. Schölkopf, B. and Smola, A. (2002) *Learning with Kernels*, MIT Press, Cambridge.
82. Schölkopf, B., Smola, A., and Müller, K.R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.
83. Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.*, **26**, 5–14.
84. Wassermann, A.M., Geppert, H., and Bajorath, J. (2011) Application of support vector machine-based ranking strategies to search for target-selective compounds. *Methods Mol. Biol.*, **672**, 517–530.
85. Schwaighofer, A., Schroeter, T., Mika, S., Laub, J., ter Laak, A., Sülzle, D., Ganzer, U., Heinrich, N., and Müller, K.R. (2007) Accurate solubility prediction with error bars for electrolytes: a machine learning approach. *J. Chem. Inf. Model.*, **47**, 407–424.
86. Obrezanova, O., Csanyi, G., Gola, J.M., and Segall, M.D. (2007) Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.*, **47**, 1847–1857.
87. Hansen, K., Rathke, F., Schroeter, T., Rast, G., Fox, T., Kriegl, J.M., and Mika, S. (2009) Bias-correction of regression models: a case study on hERG inhibition. *J. Chem. Inf. Model.*, **49**, 1486–1496.
88. Schwaighofer, A., Schroeter, T., Mika, S., and Blanchard, G. (2009)

How wrong can we get? A review of machine learning approaches and error bars. *Comb. Chem. High Throughput Screening*, **12**, 453–468.

89. Rupp, M., Schroeter, T., Steri, R., Zettl, H., Proschak, E., Hansen, K., Rau, O., Schwarz, O., Müller-Kuhrt, L., Schubert-Zsilavecz, M., Müller, K.R., and Schneider, G. (2010) From machine learning to natural product derivatives that selectively activate transcription factor PPARgamma. *ChemMedChem*, **5**, 191–194.
90. Rasmussen, C. and Williams, C. (2006) *Gaussian Processes for Machine Learning*, MIT Press, Cambridge.
91. Schneider, P. and Schneider, G. (2003) Collection of bioactive reference compounds for focused library design. *QSAR Comb. Sci.*, **22**, 713–718.
92. Reymond, J.L. and Awale, M. (2012) Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem. Neurosci.*, **3**, 649–657.
93. Ruddigkeit, L., Blum, L.C., and Reymond, J.L. (2013) Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.*, **53**, 56–65.
94. (a) Schneider, G., Lee, M.L., Stahl, M., and Schneider, P. (2000) *De novo* design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comp. Aided Mol. Des.*, **14**, 487–494. (b) Schneider, G., Clément-Chomienne, O., Hilfiger, L., Schneider, P., Kirsch, S., Böhm, H.J., and Neidhart, W. (2000) Virtual screening for bioactive molecules by evolutionary *de novo* design. *Angew. Chem. Int. Ed.*, **39**, 4130–4133.
95. Keserü, G.M. and Makara, G.M. (2006) Hit discovery and hit-to-lead approaches. *Drug Discov. Today*, **11**, 741–748.
96. Hajduk, P.J. (2006) Fragment-based drug design: how big is too big? *J. Med. Chem.*, **49**, 6972–6976.
97. Ichihara, O., Barker, J., Law, R.J., and Whittaker, M. (2011) Compound design by fragment-linking. *Mol. Inf.*, **30**, 298–306.
98. Babaoglu, K. and Shoichet, B.K. (2006) Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.*, **2**, 720–723.
99. Murray, C.W. and Verdonk, M.L. (2002) The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aided Mol. Des.*, **16**, 741–753.
100. Hubbard, R.E., Chen, I., and Davis, B. (2007) Informatics and modeling challenges in fragment-based drug discovery. *Curr. Opin. Drug Discov. Dev.*, **10**, 289–297.
101. Nazaré, M., Matter, H., Will, D.W., Wagner, M., Urmann, M., Czech, J., Schreuder, H., Bauer, A., Ritter, K., and Wehner, V. (2012) Fragment deconstruction of small, potent factor Xa inhibitors: exploring the superadditivity energetics of fragment linking in protein-ligand complexes. *Angew. Chem. Int. Ed.*, **51**, 905–911.
102. Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., and Schneider, G. (2012) DOGS: reaction-driven *de novo* design of bioactive compounds. *PLoS Comput. Biol.*, **8**, e1002380.
103. Schneider, G., Hartenfeller, M., Reutlinger, M., Tanrikulu, Y., Proschak, E., and Schneider, P. (2009) Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol.*, **27**, 18–26.
104. Hiss, J.A., Hartenfeller, M., and Schneider, G. (2010) Concepts and applications of "natural computing" techniques in *de novo* drug and peptide design. *Curr. Pharm. Des.*, **16**, 1656–1665.
105. Hartenfeller, M., Proschak, E., Schüller, A., and Schneider, G. (2008) Concept of combinatorial de novo design of druglike molecules by particle swarm optimization. *Chem. Biol. Drug Des.*, **72**, 16–26.
106. Kutchukian, P.S., Lou, D., and Shakhnovich, E.I. (2009) FOG: Fragment optimized growth algorithm for the de novo generation of molecules

occupying druglike chemical space. *J. Chem. Inf. Model.*, **49**, 1630–1642.

107. Lewell, X.Q., Judd, D., Watson, S., and Hann, M. (1998) RECAP–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.*, **38**, 511–522.

108. Makino, S., Ewing, T.J.A., and Kuntz, I.D. (1999) DREAM++: flexible docking program for virtual combinatorial libraries. *J. Comput. Aided Mol. Des.*, **13**, 513–532.

109. Vinkers, H.M., de Jonge, M.R., Daeyaert, F.F., Heeres, J., Koymans, L.M., van Lenthe, J.H., Lewi, P.J., Timmerman, H., Van Aken, K., and Janssen, P.A. (2003) Synopsis: synthesize and optimize system *in silico*. *J. Med. Chem.*, **46**, 2765–2773.

110. Gillett, V.J. (1995) SPROUT, HIPPO and CAESA: tools for *de novo* structure generation and estimation of synthetic accessibility. *Perspect. Drug Discov. Des.*, **3**, 34–50.

111. Zaliani, A., Boda, K., Seidel, T., Herwig, A., Schwab, C.H., Gasteiger, J., Claussen, H., Lemmen, C., Degen, J., Pärn, J., and Rarey, M. (2009) Second-generation *de novo* design: a view from a medicinal chemist perspective. *J. Comput. Aided Mol. Des.*, **23**, 593–602.

112. Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S.Y., Johnson, A.P., Major, S., Wade, R.A., and Ando, H.Y. (2009) Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.*, **49**, 593–602.

113. (a) Danziger, D.J. and Dean, P.M. (1989) Automated site-directed drug design: a general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proc. R. Soc. London, Ser. B*, **236**, 101–113. (b) Lewis, R.A. and Dean, P.M. (1989) Automated site-directed drug design: the concept of spacer skeletons for primary structure generation. *Proc. R. Soc. London, Ser. B*, **236**, 125–140.

114. Gillett, V.A., Johnson, A.P., Mata, P., and Sike, S. (1990) Automated structure design in 3D. *Tetrahedron Comp. Methodol.*, **3**, 681–696.

115. Lewis, R.A., Roe, D.C., Huang, C., Ferrin, T.E., Langridge, R., and Kuntz, I.D. (1992) Automated site-directed drug design using molecular lattices. *J. Mol. Graph.*, **10**, 66–78.

116. Nishibata, Y. and Itai, A. (1991) Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron*, **47**, 8985–8990.

117. Nishibata, Y. and Itai, A. (1993) Confirmation of usefulness of a structure construction program based on three-dimensional receptor structure for rational lead generation. *J. Med. Chem.*, **36**, 2921–2928.

118. Tschinke, V. and Cohen, N.C. (1993) The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypothesis. *J. Med. Chem.*, **36**, 3863–3870.

119. Ho, C.M.W. and Marshall, G.R. (1993) SPLICE: a program to assemble partial query solutions from three-dimensional database searches into novel ligands. *J. Comput. Aided Mol. Des.*, **7**, 623–647.

120. Rotstein, S.H. and Murcko, M.A. (1993) GroupBuild: a fragment-based method for *de novo* drug design. *J. Med. Chem.*, **36**, 1700–1710.

121. Pearlman, D.A. and Murcko, M.A. (1993) CONCEPTS: new dynamic algorithm for *de novo* design suggestion. *J. Comput. Chem.*, **14**, 1184–1193.

122. (a) Gillet, V.J., Johnson, A.P., Mata, P., Sike, S., and Williams, P. (1993) SPROUT: a program for structure generation. *J. Comput. Aided Mol. Des.*, **7**, 127–153. (b) Gillet, V.J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z., and Johnson, A.P. (1994) SPROUT: recent developments in the *de novo* design of molecules. *J. Comput. Aided Mol. Des.*, **34**, 207–217. (c) Mata, P., Gillet, V.J., Johnson, A.P., Lampreia, J., Myatt, G.J., Sike, S., and Stebbings, A.L. (1995) SPROUT: 3D structure generation using templates. *J. Chem. Inf. Comput. Sci.*, **35**, 479–493.

123. (a) Eisen, M.B., Wiley, D.C., Karplus, M., and Hubbard, R.E. (1994) HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins*, **19**, 199–221. (b) Caflish, A., Miranker, A., and Karplus, M. (1993) Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *J. Med. Chem.*, **36**, 2142–2167.

124. Bohacek, R.S. and McMartin, C. (1994) Multiple highly diverse structures complementary to enzyme binding sites: results of extensive application of a *de novo* design method incorporating combinatorial growth. *J. Am. Chem. Soc.*, **116**, 5560–5571.

125. Glen, R.C. and Payne, A.W. (1995) A genetic algorithm for the automated generation of molecules within constraints. *J. Comput. Aided Mol. Des.*, **9**, 181–202.

126. (a) Clark, D.E., Frenkel, D., Levy, S.A., Li, J., Murray, C.W., Robson, B., Waszkowycz, B., and Westhead, D.R. (1995) PRO LIGAND: an approach to de novo molecular design. 1. Application to the design of organic molecules. *J. Comput. Aided Mol. Des.*, **9**, 13–32. (b) Waszkowycz, B., Clark, D.E., Frenkel, D., Li, J., Murray, C.W., Robson, B., and Westhead, D.R. (1994) PRO LIGAND: an approach to de novo molecular design. 2. Design of novel molecules from molecular field analysis (MFA) models and pharmacophores. *J. Med. Chem.*, **37**, 3994–4002. (c) Westhead, D.R., Clark, D.E., Frenkel, D., Li, J., Murray, C.W., Robson, B., and Waszkowycz, B. (1995) PRO LIGAND: an approach to *de novo* molecular design. 3. A genetic algorithm for structure refinement. *J. Comput. Aided Mol. Des.*, **9**, 139–148. (d) Frenkel, D., Clark, D.E., Li, J., Murray, C.W., Robson, B., Waszkowycz, B., and Westhead, D.R. (1995) PRO LIGAND: an approach *de novo* molecular design. 4. Application to the design of peptides. *J. Comput. Aided Mol. Des.*, **9**, 213–225. (e) Clark, D.E. and Murray, C.W. (1995) PRO LIGAND: an approach to *de novo* molecular design. 5. Tools for the analysis of generated structures. *J. Chem. Inf. Comput. Sci.*, **35**, 914–923. (f) Murray, C.W., Clark, D.E., and Byrne, D.G. (1995) PRO LIGAND: an approach to *de novo* molecular design. 6. Flexible fitting in the design of peptides. *J. Comput. Aided Mol. Des.*, **9**, 381–395.

127. (a) DeWitte, R.S. and Shakhnovich, E.I. (1996) SMoG *de novo* design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.*, **118**, 11733–11744. (b) Ishchenko, A.V. and Shakhnovich, E.I. (2002) SMall Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein–ligand interactions. *J. Med. Chem.*, **45**, 2770–2780. (c) Grzybowski, B.A., Ishchenko, A.V., Kim, C.Y., Topalov, G., Chapman, R., Christianson, D.W., Whitesides, G.M., and Shakhnovich, E.I. (2002) Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 1270–1273.

128. Pearlman, D.A. and Murcko, M.A. (1996) CONCERTS: dynamic connection of fragments as an approach to *de novo* ligand design. *J. Med. Chem.*, **39**, 1651–1663.

129. Luo, Z., Wang, R., and Lai, L. (1996) RASSE: a new method for structure-based drug design. *J. Chem. Inf. Comput. Sci.*, **36**, 1187–1194.

130. (a) Murray, C.W., Clark, D.E., Auton, T.R., Firth, M.A., Li, J., Sykes, R.A., Waszkowycz, B., Westhead, D.R., and Young, S.C. (1997) PRO_SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *J. Comput. Aided Mol. Des.*, **11**, 193–207. (b) Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V., and Mee, R.P. (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor

complexes. *J. Comput. Aided Mol. Des.*, **11**, 425–445.

131. (a) Todorov, N.P. and Dean, P.M. (1997) Evaluation of a method for controlling molecular scaffold diversity in *de novo* ligand design. *J. Comput. Aided Mol. Des.*, **11**, 175–192. (b) Todorov, N.P. and Dean, P.M. (1998) A branch-and-bound method for optimal atom-type assignment in *de novo* ligand design. *J. Comput. Aided Mol. Des.*, **12**, 335–350. (c) Stahl, M., Todorov, N.P., James, T., Mauser, H., Böhm, H.J., and Dean, P.M. (2002) A validation study on the practical use of automated *de novo* design. *J. Comput. Aided Mol. Des.*, **16**, 459–478.

132. (a) Nachbar, R.B. (1998) Molecular evolution: a hierarchical representation for chemical topology and its automated manipulation. Proceedings of the 3rd Annual Genetic Programming Conference, pp. 246–253; (b) Nachbar, R.B. (2000) Molecular evolution: automated manipulation of hierarchical chemical topology and its application to average molecular structures. *Genet. Program. Evolv. Mach.*, **1**, 57–94.

133. Globus, A., Lawton, J., and Wipke, W.T. (1999) Automatic molecular design using evolutionary algorithms. *Nanotechnology*, **10**, 290–299.

134. (a) Liu, H., Duan, Z., Luo, Q., and Shi, Y. (1999) Structure-based ligand design by dynamically assembling molecular building blocks at binding site. *Proteins*, **36**, 462–470. (b) Zhu, J., Yu, H., Fan, H., Liu, H., and Shi, Y. (2001) Design of selective inhibitors of cyclooxygenase-2 dynamic assembly of molecular building blocks. *J. Comput. Aided Mol. Des.*, **15**, 447–463.

135. Douguet, D., Thoreau, E., and Grassy, G. (2000) A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm. *J. Comput. Aided Mol. Des.*, **14**, 449–466.

136. Wang, R., Gao, Y., and Lai, L. (2000) LigBuilder: a multi-purpose program for structure-based drug design. *J. Mol. Model.*, **6**, 498–516.

137. Zhu, J., Fan, H., Liu, H., and Shi, Y. (2001) Structure-based ligand design for flexible proteins: application of new F-DycoBlock. *J. Comput. Aided Mol. Des.*, **15**, 979–996.

138. Pegg, S.C.H., Haresco, J.J., and Kuntz, I.D. (2001) A genetic algorithm for structure-based *de novo* design. *J. Comput. Aided Mol. Des.*, **15**, 911–933.

139. Pellegrini, E. and Field, M.J. (2003) Development and testing of a *de novo* drug-design algorithm. *J. Comput. Aided Mol. Des.*, **17**, 621–641.

140. Brown, N., McKay, B., Gilardoni, F., and Gasteiger, J. (2004) A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.*, **44**, 1079–1087.

141. Pierce, A.C., Rao, G., and Bemis, G.W. (2004) BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease. *J. Med. Chem.*, **47**, 2768–2775.

142. Nikitin, S., Zaitseva, N., Demina, O., Solovieva, V., Mazin, E., Mikhalev, S., Smolov, M., Rubinov, A., Vlasov, P., Lepikhin, D., Khachko, D., Fokin, V., Queen, C., and Zosimov, V. (2005) A very large diversity space of synthetically accessible compounds for use with drug design programs. *J. Comput. Aided Mol. Des.*, **19**, 47–63.

143. Douguet, D., Munier-Lehmann, H., Labesse, G., and Pochet, S. (2005) LEA3D: a computer-aided ligand design for structure-based drug design. *J. Med. Chem.*, **48**, 2457–2468.

144. (a) Fechner, U. and Schneider, G. (2006) Flux (1): a virtual synthesis scheme for fragment-based *de novo* design. *J. Chem. Inf. Model.*, **46**, 699–707. (b) Fechner, U. and Schneider, G. (2007) Flux (2): comparison of molecular mutation and crossover operators for ligand-based *de novo* design. *J. Chem. Inf. Model.*, **47**, 656–667.

145. Degen, J. and Rarey, M. (2006) FlexNovo: structure-based searching in large fragment spaces. *ChemMedChem*, **1**, 854–868.

146. Feher, M., Gao, Y., Baber, C., Shirley, W.A., and Saunders, J. (2008) The use of ligand-based *de novo* design for

scaffold hopping and sidechain optimization: two case studies. *Bioorg. Med. Chem.*, **16**, 422–427.

147. Dey, F. and Caflisch, A. (2008) Fragment-based *de novo* ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.*, **48**, 679–690.

148. (a) Proschak, E., Sander, K., Zettl, H., Tanrikulu, Y., Rau, O., Schneider, P., Schubert-Zsilavecz, M., Stark, H., and Schneider, G. (2009) From molecular shape to potent bioactive agents II: fragment-based *de novo* design. *ChemMedChem*, **4**, 45–48. (b) Proschak, E., Zettl, H., Tanrikulu, Y., Weisel, M., Kriegl, J.M., Rau, O., Schubert-Zsilavecz, M., and Schneider, G. (2009) From molecular shape to potent bioactive agents I: bioisosteric replacement of molecular fragments. *ChemMedChem*, **4**, 41–44.

149. Hecht, D. and Fogel, G.B. (2009) A novel *in silico* approach to drug discovery via computational intelligence. *J. Chem. Inf. Model.*, **49**, 1105–1121.

150. Moriaud, F., Doppelt-Azeroual, O., Martin, L., Oguievetskaia, K., Koch, K., Vorotyntsev, A., Adcock, S.A., and Delfaud, F. (2009) Computational fragment-based approach at PDB scale by protein local similarity. *J. Chem. Inf. Model.*, **49**, 280–294.

151. Nicolaou, C.A., Apostolakis, J., and Pattichis, C.S. (2009) *De novo* drug design using multiobjective evolutionary graphs. *J. Chem. Inf. Model.*, **49**, 295–307.

152. Nisius, B. and Rester, U. (2009) Fragment shuffling: an automated workflow for three-dimensional fragment-based ligand design. *J. Chem. Inf. Model.*, **49**, 1211–1222.

153. Durrant, J.D., Amaro, R.E., and McCammon, J.A. (2009) AutoGrow: a novel algorithm for protein inhibitor design. *Chem. Biol. Drug Des.*, **73**, 168–178.

154. Damewood, J.R., Lerman, C.L., and Masek, B.B. (2010) NovoFLAP: a ligand-based *de novo* design approach for the generation of medicinally relevant ideas. *J. Chem. Inf. Model.*, **50**, 1296–1303.

155. Huang, Q., Li, L.L., and Yang, S.J. (2010) PhDD: a new pharmacophore-based *de novo* design method of drug-like molecules combined with assessment of synthetic accessibility. *J. Mol. Graph. Model.*, **28**, 775–787.

156. Pfeffer, P., Fober, T., Hüllermeier, E., and Klebe, G. (2010) GARLig: a fully automated tool for subset selection of large fragment spaces via a self-adaptive genetic algorithm. *J. Chem. Inf. Model.*, **50**, 1644–1659.

157. White, D. and Wilson, R.C. (2010) Generative models for chemical structures. *J. Chem. Inf. Model.*, **50**, 1257–1274.

158. Lippert, T., Schulz-Gasch, T., Roche, O., Guba, W., and Rarey, M. (2011) *De novo* design by pharmacophore-based searches in fragment spaces. *J. Comput. Aided Mol. Des.*, **25**, 931–945.

159. Wong, S.S., Luo, W., and Chan, K.C. (2011) EvoMD: an algorithm for evolutionary molecular design. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 987–1003.

160. Jiang, C., Yang, L., Wu, W.T., Guo, Q.L., and You, Q.D. (2011) *De novo* design, synthesis and biological evaluation of 1,4-dihydroquinolin-4-ones and 1,2,3,4-tetrahydroquinazolin-4-ones as potent kinesin spindle protein (KSP) inhibitors. *Bioorg. Med. Chem.*, **19**, 5612–5627.

161. Urich, R., Wishart, G., Kiczun, M., Richters, A., Tidten-Luksch, N., Rauh, D., Sherborne, B., Wyatt, P.G., and Brenk, R. (2013) *De novo* design of protein kinase inhibitors by in silico identification of hinge region-binding fragments. *ACS Chem. Biol.*, **8**, 1044–1052.

162. Hartenfeller, M. and Schneider, G. (2011) Enabling future drug discovery by *de novo* design. *WIREs Comput. Mol. Sci.*, **1**, 742–759.

163. Kitchen, D.B., Decornez, H., Furr, J.R., and Bajorath, J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, **3**, 935–949.

164. Wang, B., Westerhoff, L.M., and Merz, K.M. Jr., (2007) A critical assessment of

the performance of protein-ligand scoring functions based on NMR chemical shift perturbations. *J. Med. Chem.*, **50**, 5128–5134.

165. Coupez, B. and Lewis, R.A. (2006) Docking and scoring–theoretically easy, practically impossible? *Curr. Med. Chem.*, **13**, 2995–3003.

166. Michel, J., Foloppe, N., and Essex, J.W. (2010) Rigorous free energy calculations in structure-based drug design. *Mol. Inf.*, **29**, 570–578.

167. Schneider, G. and Böhm, H.J. (2002) Virtual screening and fast automated docking methods. *Drug Discov. Today*, **7**, 64–70.

168. Gohlke, H. and Klebe, G. (2001) Statistical potentials and scoring functions applied to protein-ligand binding. *Curr. Opin. Struct. Biol.*, **11**, 231–235.

169. Grzybowski, B.A., Ishchenko, A.V., Shimada, J., and Shakhnovich, E.I. (2002) From knowledge-based potentials to combinatorial lead design in silico. *Acc. Chem. Res.*, **35**, 261–269.

170. Schneider, G. and Fechner, U. (2005) Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discov.*, **4**, 649–663.

171. Dean, P.M. (2007) Chemical genomics: a challenge for *de novo* drug design. *Mol. Biotechnol.*, **37**, 237–245.

172. Mauser, H. and Guba, W. (2008) Recent developments in *de novo* design and scaffold hopping. *Curr. Opin. Drug Discov. Dev.*, **11**, 365–374.

173. Jorgensen, W.L. (2009) Efficient drug lead discovery and optimization. *Acc. Chem. Res.*, **42**, 724–733.

174. Pirard, B. (2011) The quest for novel chemical matter and the contribution of computer-aided *de novo* design. *Expert Opin. Drug Discov.*, **6**, 225–231.

175. Warr, W.A. (2011) Some trends in chem(o)informatics. *Methods Mol. Biol.*, **672**, 1–37.

176. Kutchukian, P.S. and Shakhnovich, E.I. (2010) *De novo* design: balancing novelty and confined chemical space. *Expert Opin. Drug Discov.*, **5**, 789–812.

177. Konteatis, Z.D. (2010) In silico fragment-based drug design. *Expert Opin. Drug Discov.*, **5**, 1047–1065.

178. Hartenfeller, M. and Schneider, G. (2011) *De novo* drug design. *Methods Mol. Biol.*, **672**, 299–323.

179. Sheng, C. and Zhang, W. (2013) Fragment informatics and computational fragment-based drug design: an overview and update. *Med. Res. Rev.*, **33**, 554–598.

180. Nicolaou, C.A., Kannas, C., and Loizidou, E. (2012) Multi-objective optimization methods in *de novo* drug design. *Mini Rev. Med. Chem.*, **12**, 979–987.

181. Beddell, C.R., Goodford, P.J., Norrington, F.E., Wilkinson, S., and Wootton, R. (1976) Compounds designed to fit a site of known structure in human haemoglobin. *Br. J. Pharmacol.*, **57**, 201–209.

182. Beddell, C.R., Goodford, P.J., Stammers, D.K., and Wootton, R. (1979) Species differences in the binding of compounds designed to fit a site of known structure in adult human haemoglobin. *Br. J. Pharmacol.*, **65**, 535–543.

183. Gund, P., Wipke, W.T., and Langridge, R. (1974) Computer searching of a molecular structure file for pharmacophoric patterns. *Comput. Chem. Res. Educ. Technol.*, **3**, 5–21.

184. Martin, Y.C., Bures, M.G., and Willett, P. (1990) in *Reviews in Computational Chemistry*, Vol. 1 (eds K. Lipkowitz and D. Boyd), Wiley-VCH Verlag GmbH, Weinheim, pp. 213–263.

185. Sheridan, R.P. and Venkataraghavan, R. (1987) Designing novel nicotinic agonists by searching a database of molecular shapes. *J. Comput. Aided Mol. Des.*, **1**, 243–256.

186. Lewis, R.A. and Dean, P.M. (1989) Automated site-directed drug design: the formation of molecular templates in primary structure generation. *Proc. R. Soc. London, Ser. B*, **236**, 141–162.

187. Van Drie, J.H., Weininger, D., and Martin, Y.C. (1989) ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric,

and substructure searching of three-dimensional molecular structures. *J. Comput. Aided Mol. Des.*, **3**, 225–240.

188. Bartlett, P.A., Shea, G.T., Telfer, S.J., and Waterman, S. (1989) in *Molecular Recognition in Chemical and Biological Problems*, Vol. 78 (ed. S.M. Roberts), Royal Society of Chemistry, London, pp. 182–196.

189. Lauri, G. and Bartlett, P.A. (1994) CAVEAT: a program to facilitate the design of organic molecules. *J. Comput. Aided Mol. Des.*, **8**, 51–66.

190. Carhart, R.E., Smith, D.H., Gray, N.A.B., Nourse, J.G., and Djerassi, C. (1981) GENOA: a computer program for structure elucidation utilizing overlapping and alternative substructures. *J. Org. Chem.*, **46**, 1708–1718.

191. Wise, M., Cramer, R.D., Smith, D., and Exman, I. (1983) in *Quantitative Approaches to Drug Design* (ed. J.C. Dearden), Elsevier, Amsterdam, pp. 145–146.

192. Goodford, P.J. (1995) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849–857.

193. Miranker, A. and Karplus, M. (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins*, **11**, 29–34.

194. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. (1982) A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, **161**, 269–288.

195. Jackson, R.C. (1995) Update on computer-aided drug design. *Curr. Opin. Biotechnol.*, **6**, 646–651.

196. Böhm, H.J. (1996) Current computational tools for *de novo* ligand design. *Curr. Opin. Biotechnol.*, **7**, 433–436.

197. Bohacek, R.S. and McMartin, C. (1997) Modern computational chemistry and drug discovery: structure generating programs. *Curr. Opin. Chem. Biol.*, **1**, 157–161.

198. Marrone, T.J., Briggs, J.M., and McCammon, J.A. (1997) Structure-based drug design: computational advances. *Annu. Rev. Pharmacol. Toxicol.*, **37**, 71–90.

199. Kubinyi, H. (1998) Combinatorial and computational approaches in structure-based drug design. *Curr. Opin. Drug Discov. Dev.*, **1**, 16–27.

200. Moon, J.B. and Howe, W.J. (1991) Computer design of bioactive molecules: a method for receptor-based *de novo* ligand design. *Proteins*, **11**, 314–328.

201. Böhm, H.J. (1992) The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.*, **6**, 61–78.

202. Böhm, H.J. (1992) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided Mol. Des.*, **6**, 593–606.

203. Nestor, J.J. Jr., (2009) The medicinal chemistry of peptides. *Curr. Med. Chem.*, **16**, 4399–4418.

204. Vanhee, P., van der Sloot, A.M., Verschueren, E., Serrano, L., Rousseau, F., and Schymkowitz, J. (2011) Computational design of peptide ligands. *Trends Biotechnol.*, **29**, 231–239.

205. Renukuntla, J., Vadlapudi, A.D., Patel, A., Boddu, S.H., and Mitra, A.K. (2013) Approaches for enhancing oral bioavailability of peptides and proteins. *Int. J. Pharm.*, **447**, 75–93.

206. Kandil, S., Biondaro, S., Vlachakis, D., Cummins, A.C., Coluccia, A., Berry, C., Leyssen, P., Neyts, J., and Brancale, A. (2009) Discovery of a novel HCV helicase inhibitor by a *de novo* drug design approach. *Bioorg. Med. Chem. Lett.*, **19**, 2935–2937.

207. Rogers-Evans, M., Alanine, A., Bleicher, K., Kube, D., and Schneider, G. (2004) Identification of novel cannabinoid receptor ligands via evolutionary *de novo* design and rapid parallel synthesis. *QSAR Comb. Sci.*, **26**, 426–430.

208. Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) 'Scaffold-hopping' by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed.*, **38**, 2894–2896.

209. Alig, L., Alsenz, J., Andjelkovic, M., Bendels, S., Bénardeau, A., Bleicher, K., Bourson, A.,

David-Pierson, P., Guba, W., Hildbrand, S., Kube, D., Lübbers, T., Mayweg, A.V., Narquizian, R., Neidhart, W., Nettekoven, M., Plancher, J.M., Rocha, C., Rogers-Evans, M., Röver, S., Schneider, G., Taylor, S., and Waldmeier, P. (2008) Benzodioxoles: novel cannabinoid-1 receptor inverse agonists for the treatment of obesity. *J. Med. Chem.*, **51**, 2115–2127.

210. Mitchell, W. and Matsumoto, S. (2011) Large-scale integrated super-computing platform for next generation virtual drug discovery. *Curr. Opin. Chem. Biol.*, **15**, 553–559.

211. Park, H., Ryu, S.E., and Kim, S.J. (2012) Structure-based *de novo* design of Eya2 phosphatase inhibitors. *J. Mol. Graph. Model.*, **38**, 382–388.

212. Park, H., Jeong, Y., and Hong, S. (2012) Structure-based *de novo* design and biochemical evaluation of novel BRAF kinase inhibitors. *Bioorg. Med. Chem. Lett.*, **22**, 1027–1030.

213. Ishchenko, A., Liu, Z., Lindblom, P., Wu, G., Jim, K.C., Gregg, R.D., Claremon, D.A., and Singh, S.B. (2012) Structure-based design technology contour and its application to the design of renin inhibitors. *J. Chem. Inf. Model.*, **52**, 2089–2097.

214. van der Horst, E., Marqués-Gallego, P., Mulder-Krieger, T., van Veldhoven, J., Kruisselbrink, J., Aleman, A., Emmerich, M.T., Brussee, J., Bender, A., and Ijzerman, A.P. (2012) Multi-objective evolutionary design of adenosine receptor ligands. *J. Chem. Inf. Model.*, **52**, 1713–1721.

215. Ekins, S. and Williams, A.J. (2011) Finding promiscuous old drugs for new uses. *Pharm. Res.*, **28**, 1785–1791.

216. Keppner, S., Proschak, E., Schneider, G., and Spänkuch, B. (2009) Identification and validation of a potent type II inhibitor of inactive polo-like kinase 1. *ChemMedChem*, **4**, 1806–1809.

217. Schneider, G., Geppert, T., Hartenfeller, M., Reisen, F., Klenner, A., Reutlinger, M., Hähnke, V., Hiss, J.A., Zettl, H., Keppner, S., Spänkuch, S., and Schneider, P. (2011) Reaction-driven *de novo* design, synthesis and testing of potential type II kinase inhibitors. *Future Med. Chem.*, **3**, 415–424.

218. Spänkuch, B., Keppner, S., Lange, L., Rodrigues, T., Zettl, H., Koch, C.P., Reutlinger, M., Hartenfeller, M., Schneider, P., and Schneider, G. (2013) Drugs by numbers: reaction-driven *de novo* design of potent and selective anticancer leads. *Angew. Chem. Int. Ed.*, **52**, 4676–4681.

219. Rodrigues, T., Roudnicky, F., Koch, C.P., Kudoh, T., Reker, D., Detmar, M., and Schneider, G. (2013) *De novo* design and optimization of Aurora A kinase inhibitors. *Chem. Sci.*, **4**, 1229–1233.

220. Loving, K., Alberts, I., and Sherman, W. (2010) Computational approaches for fragment-based and de novo design. *Curr. Top. Med. Chem.*, **10**, 14–32.

221. Langdon, S.R., Ertl, P., and Brown, N. (2010) Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Mol. Inf.*, **29**, 366–385.

222. Bailey, D. and Brown, D. (2001) High-throughput chemistry and structure-based design: survival of the smartest. *Drug Discov. Today*, **6**, 57–59.

223. Bleicher, K.H., Böhm, H.J., Müller, K., and Alanine, A.I. (2003) Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.*, **2**, 369–378.

224. Schneider, G. (2010) Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.*, **9**, 273–276.

225. Krüger, B.A., Dietrich, A., Baringhaus, K.H., and Schneider, G. (2009) Scaffold-hopping potential of fragment-based *de novo* design: the chances and limits of variation. *Comb. Chem. High Throughput Screening*, **12**, 383–396.

226. Sheridan, R.P., Rusinko, A. III, Nilakantan, R., and Venkataraghavan, R. (1989) Searching for pharmacophores in large coordinate data bases and its use in drug design. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 8165–8169.