# Part One
# Principles

*Virtual Screening*. Edited by C. Sotriffer
Copyright © 2011 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-32636-5

# 1
# Virtual Screening of Chemical Space: From Generic Compound Collections to Tailored Screening Libraries

*Markus Boehm*

## 1.1
## Introduction

Today's challenge of making the drug discovery process more efficient remains unchanged. The need for developing safe and innovative drugs, under the increasing pressure of speed and cost reduction, has shifted the focus toward improving the early discovery phase of lead identification and optimization. "Fail early, fail fast, and fail cheap" has often been quoted as the key principle contributing to the overall efficiency gain in drug discovery. While high-throughput screening (HTS) of large compound libraries is still the major source for discovering novel hits in the pharmaceutical industry, virtual screening has made an increasing impact in many areas of the lead identification process and has evolved into an established computational technology in modern drug discovery over the past 10 years.

Traditionally, virtual screening is conducted simply by searching the company proprietary database of its compound collections, and this approach continues to be a mainstream application. However, the continuous development of novel and more sophisticated virtual screening methods has opened up the possibility to search also for compounds that do not necessarily exist in physical form in a screening collection. Such compounds can be obtained either from a multitude of external sources, such as compound libraries from commercial vendors, or from public or commercial databases. Even more, virtual screening can deal with molecules that purely exist as virtual entities derived from *de novo* design ideas or enumeration of combinatorial libraries. Taken to its extreme, any molecule conceivable by the human mind can in theory be evaluated by virtual screening. This has led to the concept of chemical space comprising the entire collection of all possible molecules – real and imaginary – that could be created. Since such a chemical space is huge, it is crucial for the success of drug discovery to identify those regions in chemical space that contain molecules of oral druglike quality that are likely to be biologically active. Virtual screening has the unique capability of not only searching the small fraction of chemical space occupied by compounds in existing screening collections but also exploring new and so far

*Virtual Screening*. Edited by C. Sotriffer
Copyright © 2011 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-32636-5

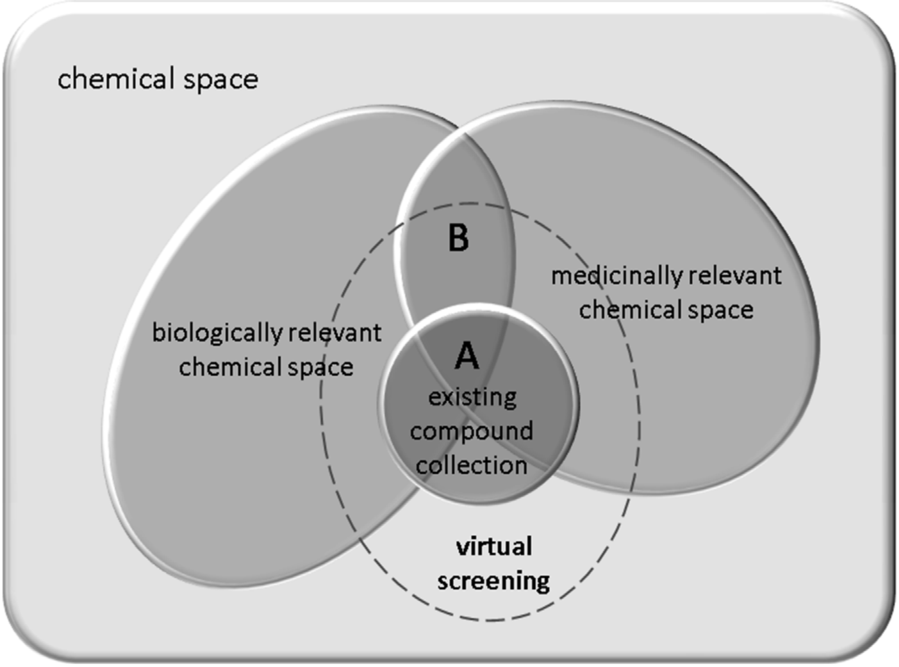
chemical space
B
medicinally relevant chemical space
biologically relevant chemical space
A
existing compound collection
virtual screening

**Figure 1.1** Regions of biologically and medicinally relevant chemical space within the continuum of chemical space. Only a small portion of chemical space has been sampled by existing compound collections, which led to the discovery of drugs (A). Virtual screening has the unique opportunity to expand into unexplored chemical space to find new pockets of space where drugs are likely to be discovered (B).

undiscovered regions (Figure 1.1). The challenge for the future is to better define and systematically explore those promising areas in chemical space.

## 1.2 Concepts of Chemical Space

Despite the fact that the term *chemical space* has received widespread attention in drug discovery, only few concrete definitions have been proposed. Lipinski suggested that chemical space “can be viewed as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars” [1]. More concrete, chemical space can be defined as the entire collection of all meaningful chemical compounds, typically restricted to small organic molecules [2]. To navigate through the vastness of chemical space, compounds can be mapped onto the coordinates of a multidimensional descriptor space. Each dimension represents various properties describing the molecules, such as physicochemical or topological properties, molecular fingerprints, or similarity to a given reference compound [3]. Depending on the particular descriptor and property set used for defining a chemical space, the representation of compounds in this chemical space varies. Thus, the relative distribution of molecules within the chemical space and the relationship between them strongly depend on the chosen descriptor set. The consequence of this is that changes in chemical representation of molecules are likely to result in changes in their neighborhood relationship. This aspect is important to keep in mind when it comes to measuring diversity or similarity within a set of molecules.

How vast is chemical space? Various estimates of the size of chemical space have been proposed. The number of all possible, small organic compounds ranges anywhere from $10^{18}$ to $10^{180}$ molecules [4]. The first attempt to systematically enumerate all molecules of up to 13 heavy atoms applying basic chemical feasibility rules resulted in less than $10^9$ structures [5]. However, with every additional heavy atom the number of possible structures grows exponentially due to the combinatorial explosion of enumeration. Thus, it is estimated that with less than 30 heavy atoms more than $10^{63}$ molecules with a molecular weight of less than 500 can be generated, predicted to be stable at room temperature and stable toward oxygen and water [6]. Compared to the estimated number of atoms in the entire observable universe ($10^{80}$), it seems that for all practical purposes chemical space is infinite and any attempt to fully capture it even with computational methods appears to be futile. Even more, in contrast to the number of compounds in a typical screening collection of large pharmaceutical companies ($10^6$) it becomes clearly obvious that only a tiny fraction of chemical space is examined.

One might ask why hit identification in drug discovery is successful, despite the fact that only a very limited set of compounds within the entire chemical space is being probed. It has been hypothesized that existing screening collections are not just randomly selected from chemical space, but are already enriched with molecules that are likely to be recognized by biological targets [7]. Many synthesized compounds have been derived from natural products, metabolites, protein substrates, natural ligands, and other biogenic molecules. Hence, a certain “biogenic bias” is inherently built into existing screening libraries resulting in an increased chance of finding active hits. This observation indicates that, given the vast and infinite size of chemical space, the goal should not be to exhaustively sample the entire space but to identify those regions that contain compounds likely to be active against biological targets (biologically relevant chemical space).

Another limiting factor is that not all biologically active molecules have the desired physicochemical properties required for oral drugs. There are many aspects important for a biologically active compound to become a safe and orally administered drug, such as absorption, permeability, metabolic stability, or toxicity. The concept of *druglikeness* has been introduced to determine the characteristics necessary for a drug likely to be successful. Over time, this has been further extended toward leadlike criteria with more stringent rules and guidelines recommended for compounds in a screening collection (Section 1.3). It is generally assumed that molecules have an increased chance to be successfully developed into a medicine when they satisfy lead- and druglike criteria (medicinally relevant chemical space).

Unfortunately, not much is known about the size and regions of biologically and medicinally relevant chemical space. Current definitions of such relevant spaces often rely on the knowledge of existing, mostly orally administered drugs, and are limited by the chemical diversity of historical screening collections and by the biological diversity of known druggable targets. On one side, the data accumulated so far suggest that compounds active against certain target families (e.g., GPCRs, kinases, etc.) tend to cluster together in specific regions of chemical space [8]. For individual targets from those families, the relevant chemical space seems to be well defined, and the likelihood of finding drugs in these defined regions is high (Section 1.5). On the other side, there are many target classes that have been deemed as difficult or undruggable, such as certain proteases or phosphatases. Also, a fairly unresolved area in drug discovery is the identification of small molecule modulators of protein–protein interactions in biological signaling cascades [9]. It is assumed that the chemical space represented by traditional screening collections is inadequate to successfully tackle these “tough targets,” and new regions of chemical space need to be explored. Possible sources of chemical matter potentially occupying such unexplored regions of space can be derived from natural products or through emerging technologies such as diversity-oriented synthesis for generating natural product-like combinatorial libraries (Section 1.4).

## 1.3
## Concepts of Druglikeness and Leadlikeness

It has been demonstrated that the lead development stage contributes 40% to the overall attrition rate throughout the whole drug development process, beginning from the first assay development to final registration [10]. Therefore, it is assumed that significant improvements can be realized in the early phase of lead identification and development. In-depth analysis of marketed oral drugs led to the introduction of druglikeness that defines the physicochemical properties that determine key issues of drug development, such as absorption and permeability. Lipinski’s influential analysis of compounds failing to become orally administered drugs resulted in the well-known “rule of five” [11]. In short, the rule predicts that poor absorption or permeation of a drug is more likely to occur when there are more than 5 H-bond donors, 10 H-bond acceptors, the molecular weight is greater than 500, and the calculated log $P$ is greater than 5 (Table 1.1). The concept of *druglikeness* has been widely accepted and embraced by scientists in drug discovery nowadays, with many variations and extensions of the original rules, and it has served its purpose well to help optimize pharmacokinetic properties of drug candidate molecules [12, 13].

The rules defining druglikeness, however, should not necessarily be applied to lead molecules. One of the reasons is the observation that, on average, compounds in comparison to their initial leads become larger and more complex during the lead optimization phase, and the associated physicochemical properties (e.g., molecular weight, calculated log $P$, etc.) increase accordingly [14, 15]. To ensure

**Table 1.1** Comparison of properties typically used for leadlikeness and druglikeness criteria.

| Properties | Leadlikeness | Druglikeness |
|---|---|---|
| Molecular weight (MW) | $\leq 350$ | $\leq 500$ |
| Lipophilicity (clog *P*) | $\leq 3.0$ | $\leq 5.0$ |
| H-bond donor (sum of NH and OH) | $\leq 3$ | $\leq 5$ |
| H-bond acceptor (sum of N and O) | $\leq 8$ | $\leq 10$ |
| Polar surface area (PSA) | $\leq 120\,\text{Å}^2$ | $\leq 150\,\text{Å}^2$ |
| Number of rotatable bonds | $\leq 8$ | $\leq 10$ |
| Structural filters | Reactive groups<br>Warhead-containing agents<br>Frequent hitters<br>Promiscuous inhibitors | |

that the properties of an optimized compound remain within druglike space, the criteria for *leadlikeness* have been more narrowly defined to accommodate the expected growth during drug optimization (Table 1.1). Complementary to the comparison of drug and lead pairs from historical data, Hann *et al.* analyzed in a more theoretical approach, using a simplified ligand–receptor interaction model, how the probability of finding a hit varies with the complexity of a molecule [16]. The model shows that the probability of observing a "useful interaction event" decreases when molecules become increasingly complex. This suggests that less complex molecules, in accordance with leadlike criteria, are more likely to turn into hits (albeit weaker) serving as common starting points for the successful discovery of drugs.

Another aspect underlining the importance of leadlike properties is associated with the fundamental shift in the screening paradigm in drug discovery from functional biological assays to biochemical assays. While biological assays measure a true biological activity, biochemical assays are designed to measure specific molecular interactions between a compound and its target. Biochemical assays are highly sensitive assays, well suited for screening compounds in a high-throughput fashion, but due to their artificial nature they are also susceptible to compound interference resulting in false positive hits. It has been suggested that compounds with leadlike properties also interact with their targets in a leadlike manner, that is, by noncovalent binding through hydrogen bonds, hydrophobic interactions, and monoionic bonding [17]. In general, such desirable interaction types result in reversible, time-independent, and competitive binding characteristics allowing the generation of meaningful structure–activity data. In contrast, nonleadlike compounds tend to bind to their target in nonleadlike ways, such as forming covalent, chelate, or polyionic bonds. Thus, nonleadlike compounds are more prone to generating artifact data in biochemical assays.

Among the well-known offenders with nonleadlike properties are protein-reactive compounds, warhead-containing agents, frequent hitters, and aggregator compounds (Table 1.1) [18]. Computationally, the elimination of reactive and warhead-containing compounds can be accomplished by applying various sets of substructure filters [17, 19]. Frequent hitters can be identified by statistical models or other virtual screening methods [20]. Aggregator compounds have been described as being promiscuous inhibitors by forming aggregates in solution, resulting in nonspecific binding and interference with the biochemical assay [21]. However, they are difficult to predict computationally and require additional biophysical methods (e.g., light scattering experiments) or modifications of the biochemical assay (e.g., addition of detergent or protein serum) to support their detection experimentally [22]. Exacerbating the problem, the interference of compounds in biochemical screens resulting in artifact data mostly depends on the individual assay conditions, which makes it difficult to develop generally applicable rules for detecting potential false positives across different assays.

The ultimate goal of identifying compounds with leadlike properties is to design high-quality screening libraries, whether it is for experimental or virtual screening purposes [23]. From a practical standpoint, it appears that leadlike criteria are more straightforward to implement by applying rules to filter out nonleadlike compounds, with the aim of enriching the compound collection with leadlike matter. In other words, one can agree on which compounds not to screen, but the question which compounds to screen often leads to lengthy debates among experienced medicinal chemists.

## 1.4 Diversity-Based Libraries

Since the advent of large-scale combinatorial chemistry in drug discovery coupled with high-speed parallel synthesis of thousands of compounds, the concept of *molecular diversity* has increasingly gained importance. When little or nothing is known about the biological target, it is often assumed that screening a compound library as diverse as possible maximizes the chance of finding active hits. Moreover, the continuous addition of compounds to the screening file, either from internal combinatorial library efforts or through purchase of external compound collections, is most valuable when the underlying overall diversity can be expanded. At the same time, there is an ever-growing pressure to reduce costs by decreasing the number of compounds that need to be screened while simultaneously maintaining diversity. Hence, well-defined strategies for the optimal design of *diversity-based libraries* are necessary.

### 1.4.1 Concepts of Molecular Diversity

The generally accepted understanding of molecular diversity is a quantitative description of dissimilarity between molecules in a given set of compounds. The exact interpretation of this concept, however, has created quite a heated debate in the scientific literature [24]. For example, Roth fervently advocated that *per se* “there is no such thing as diversity” [25]. Diversity of chemical structure does not necessarily imply diversity of biological activity. In order to be meaningful, diversity can only be applied within a frame of reference, that is, the biological assay. Hence, structural diversity of compounds should be interpreted only with respect to their relative effect in biological screens. Finding descriptors for biological activity is necessary to describe the diversity of biological activities for compounds present in a library. Unfortunately, it is often difficult or impossible to predict in advance which descriptors are most effective in a given situation. While it remains to be a matter of subjectivity what makes a compound set diverse and how to quantify diversity, or if one compound set is more diverse than another, the minimum value gained by a diversity application is the elimination of redundancy within a screening set. A diverse set of compounds should contain only nonredundant molecules that simultaneously span a wide range of properties covering the chemical space.

The basis of removing redundancy from a compound set is formed by the general belief that similar molecules typically exhibit similar biological activities. This concept has been defined as the *similarity property principle* or neighborhood behavior, and is the fundamental assumption behind all similarity and diversity applications [26]. Although generally accepted, one can quickly find arguments against this principle, as there are many examples described where subtle modifications of a compound can lead to dramatic changes in activity (activity cliffs, “magic methyl,” etc.) or major changes in the molecular structure not resulting in significant activity differences (flat SAR). From a statistical point of view, however, it has been demonstrated that a set of compounds similar to an active hit contains a higher number of actives compared to a random set, thus increasing the probability of finding actives [27]. Various groups have analyzed large activity data sets and came to the conclusion that on average there is a 30% chance that a compound within a certain similarity cutoff (Tanimoto coefficient $\geq$0.85 using Daylight fingerprints) to an active hit is itself active [28, 29]. The backside of this finding is that diversity methods selecting a representative compound within a subset of similar compounds incur a 70% chance of picking an inactive compound and excluding compounds that might have had activity. Exacerbating the effect, diversity selections often tend to more aggressively reduce the size of screening sets by loosening similarity criteria beyond the range where the similarity property principle is applicable. This might lead to a decreased coverage of biological space, limiting the chance of finding actives within the chosen subset.

### 1.4.2
### Descriptor-Based Diversity Selection

Various strategies for the design of diversity-based screening collections have been proposed. Before initiating the selection process, some more fundamental questions should be addressed. For instance, it is often unclear how large a screening library should be and how many cluster representatives need to be selected. Using fingerprints and default similarity cutoffs for clustering (see above) and assuming the presence of actives in a cluster, there is only a 30% probability of identifying an active hit when a single representative per cluster is chosen. The selection of five compounds per cluster increases the chance of finding actives to 80% [28]. This finding suggests the selection of multiple representatives per cluster to increase the likelihood of uncovering actives. However, this comes at the expense of including fewer clusters during the selection process.

A mathematical model was developed by Harper *et al.* to provide a more quantitative framework for assessing the optimal parameters of a screening collection and their effect on the probability of producing lead series in a given biological assay [30]. For each cluster in a screening collection, the percentage of compounds expected to hit the biological target, as well as the probability of an existing lead molecule in the cluster, is empirically estimated. According to the model, the expected number of lead series per screen ("lead discovery rate") increases linearly with the number of compounds in a screening library. However, the probability of finding one or more lead series in a given screen does not grow proportionally with the size of the library. For instance, it was estimated that an average hit rate of 1.2 leads per screen is required to find at least one lead on 70% of screens. To increase the proportion of screens identifying leads to 80% and beyond requires a sharp increase in the number of compounds to be screened. This result of diminishing returns has been experienced by many companies when their screening collections have dramatically increased in size, but it has not translated into a proportional increase in successful screening campaigns. One of the main conclusions from the analysis is that, in order to increase the chance of finding lead series, a screening library of a given size should contain as many diverse clusters as possible, ideally with only one or few representatives per cluster. Increasing the number of compounds per cluster at the cost of decreasing the number of clusters ultimately lowers the likelihood of finding leads.

In principle, there are three main steps required to carry out diversity-based subset selections: (1) the calculation of descriptors representing the compound structures, (2) a quantitative method to describe the similarity or dissimilarity of molecules in relationship to each other, and (3) selection methods to identify compounds based on their similarity or dissimilarity values that best represent the entire compound set. In the following, the three steps are described in more detail.

Numerous descriptors encoding molecular properties with varying degrees of information content and complexity have been developed [31]. The current version of the Dragon software alone calculates over 3200 molecular descriptors [32]. The many different representations can be classified according to the type of information they encode [4, 33]. Whole-molecule descriptors represent different properties of a molecule in a single number, such as molecular weight or calculated log *P*. Descriptors derived from 2D representations of molecules include topological indices, which describe a structure according to its size and shape by a single number, and fingerprint-based descriptors, characterizing molecules by their substructural features. Graph-based molecular descriptors attempt to reduce the molecular complexity while capturing the overall information content of the molecular topology and properties. Descriptors derived from the 3D structure of molecules consist of fingerprint-based descriptors and other more complex representations, encoding properties such as shape or pharmacophore information of a molecule.

In order to quantify the degree of similarity or dissimilarity between two compounds, various similarity coefficients have been developed for different applications, many of them widely used for chemical similarity searching [34]. Several groups compared the performance of different similarity coefficients in combination with various fingerprint types, and it was often found that the Tanimoto coefficient markedly outperformed other similarity measures, making it the similarity coefficient of choice for fingerprint-based similarity searching [35].

Methods for selecting diverse subsets from a compound collection include (1) dissimilarity-based compound selection, (2) clustering, (3) partitioning, and (4) the use of optimization approaches, and are discussed in the following. Dissimilarity-based methods involve the selection of compound sets that maximize the dissimilarity between pairs of molecules [36]. In an iterative fashion, those molecules from a compound collection that are mostly dissimilar to the already selected compounds from the subset are added to the subset. The MaxMin selection technique and the sphere exclusion algorithm are the preferred methods of choice among dissimilarity-based methods [37, 38]. Clustering methods involve the identification of groups of compounds such that compounds within a cluster are highly similar whereas compounds from different clusters are dissimilar. Choosing one or only few representatives per cluster, usually the cluster centroids, has been demonstrated to be the best strategy for designing a highly diverse subset to maximize the chances of hit identification. Many different clustering algorithms have been developed, and they can be divided into hierarchical and nonhierarchical methods [39]. Since clustering is based on relative similarities of molecules to each other and not on an absolute scale in chemical space, it is often difficult to compare two different data sets, which is required, for instance, when purchasing new compound collections. In contrast, partitioning or cell-based methods provide an absolute measure of compounds in terms of their location in chemical space, spanned by a predefined descriptor set [40]. A low-dimensional descriptor space is required, where descriptors are mapped onto each axis of the chemical space by binning (partitioning) the range of their values into a set of cells. Molecules that fall into the same cells can be considered similar, and a diverse subset of compounds is selected by taking one or a few representatives from each cell. Pearlman's well-known BCUT descriptors, typically mapped into a six-dimensional space, were developed for the use in partitioning-based approaches [41]. A chemical global positioning system, ChemGPS, was introduced to provide a low-dimensional chemical space as a frame of reference suitable for diversity analysis [42]. A set of 72 descriptors was condensed into a nine-dimensional space by means of principal component analysis. Finally, optimization-based approaches use genetic algorithms or simulated annealing to efficiently sample large chemical spaces [43, 44].

### 1.4.3 Scaffold-Based Diversity Selection

An alternative approach to describe the diversity of a compound collection has been realized by the classification of molecules according to their underlying scaffolds. Compared to methods using traditional descriptors such as fingerprints, scaffold classification methods provide a different view of comparing databases of compounds. Scaffold diversity and coverage, as well as over- or underrepresented regions of scaffold space, can be easily assessed across different data sets, such as publicly or commercially available screening collections [45]. Scaffold analysis is also applied to HTS data to retrieve more chemically intuitive clustering results [46]. Finally, classification of compounds according to their scaffolds can help identify privileged structures and serve as a starting point for designing scaffold-focused libraries (Section 1.5) [47, 48].

Although there is no exact definition for a molecular scaffold, it generally refers to a common structural core motif. Scaffolds often resemble the chemotypes of molecules, which medicinal chemists use to categorize compounds into chemical series. Bemis and Murcko have introduced the widely used classification of compounds according to their molecular framework [49]. The molecular framework of a compound, also referred to as “Murcko scaffold,” is formed by deleting all terminal acyclic side-chain atoms from the original molecule. In addition, all atom and bond types can be removed to arrive at the graph framework of the molecule. The removal of linker length and ring size information results in the reduced graph representation of the molecule. The feature tree descriptor used in FTrees is a popular example where compounds are described by a graph (tree) that represents each molecular fragment and functional group (feature) as a node and their connectivity as edges [50]. This reduces the molecular descriptor complexity while still maintaining the overall topology and property information, making this descriptor ideal for scaffold hopping searches [51]. In a related approach, “molecular equivalence indices” (MEQI) classify molecules with respect to a variety of structural features and topological shapes, which can be used to hierarchically classify compound sets into classes of chemotypes [52]. Recently, a hierarchical classification system, Scaffold Tree, has been described [53]. Each level of the hierarchy consists of well-defined chemical substructures by iteratively removing rings from the molecular framework. Prioritization rules ensure that peripheral rings are removed first to achieve unique classification trees. Besides the benefit of its visually intuitive presentation of the scaffold tree, potential applications of this method are the detection of potential chemical series from screening hits on the basis of their hierarchical classification and the retrosynthetic combinatorial analysis of library compounds to identify the scaffolds that have been most likely used. The idea of a hierarchical classification of scaffolds has been expanded to incorporate the biological space associated with the compounds. The program Scaffold Hunter has been developed both to analyze the complex relationship of structure and activity data and to identify scaffolds of compounds likely to contain the desired biological activity [54, 55]. Analogous to the Scaffold Tree approach, scaffolds are hierarchically organized, however, using activity data as the key selection criterion during the structural deconstruction and tree building process. Scaffolds that share activity with their neighboring scaffolds in the hierarchical tree but are not represented by compounds in the data set are identified. Such virtual scaffolds can serve as starting points for the discovery of new biologically relevant scaffolds.

### 1.4.4 Sources of Diversity

Besides the established sources of obtaining diversity, mainly from historic compound collections, publicly or commercially available compound libraries, and natural products, novel approaches toward expanding diversity have been described in the recent literature.

The systematic enumeration of all possible organic molecules of up to 11 atoms of C, N, O, and F, applying simple valence, chemical stability, and synthetic feasibility rules, has been reported [56]. A total of 26.4 million compounds were generated and collected in a chemical universe database (GDB-11). An extended version (GDB-13) has been published that contains 970 million molecules of up to 13 atoms of C, N, O, S, and Cl enumerated in a similar manner, making it the largest database of publicly available virtual molecules [5]. It contains a vast number of unexplored structures and provides a new source for design ideas to identify bioactive small molecules and scaffolds. The first successful application of the GDB discovering a novel class of NMDA glycine site inhibitors has been recently reported [57].

Bioactive molecules have been shown to contain only a limited number of unique ring systems. For that reason, in analogy to the chemical universe of the GDB, several groups have explored the ring universe to identify novel ring systems and heteroaromatic scaffolds. A comprehensive collection of more than 40 000 different rings extracted from the CAS registry has been classified into ring systems on the basis of their topology, and it was shown that the distribution of rings is not continuous but contains many significant voids [58]. A drug ring database containing ring systems from proprietary and commercial compound collections has been developed as a source for scaffold replacement design [59]. Generating a database of over 600 000 heteroaromatic ring scaffolds, the comparison to scaffolds associated with biological activity revealed that bioactive scaffolds are very sparsely distributed, forming well-defined "bioactivity islands" in virtual scaffold space [60]. It is, however, unclear if biological activity is truly limited to only such small region of ring space, or if most ring systems are simply not synthetically accessible and thus have never been prepared. To overcome this limitation, the future challenge is to actively develop novel synthetic routes to prepare molecules with so far unexplored ring systems. A "virtual exploratory heterocyclic library" (VEHICLe) of almost 25 000 ring systems was created, containing a complete enumerated set of heteroaromatic rings, with rings being removed that are likely to be synthetically unfeasible according to a set of empirical rules [61]. Interestingly, the authors find that only 1700 of them (7%) have been published, and of these only a small percentage is routinely used in the synthesis of druglike molecules. They highlight many simple and apparently tractable heterocycles that have not been described in the literature so far and put out a “challenge to creative organic chemists to either make them or explain why they cannot be made.”

It has been argued that the trend in drug discovery over the past decade toward achiral, aromatic compounds, presumably due to their amenability to high-throughput synthetic approaches, may have contributed to a higher failure rate of drug development candidates [62]. Concurrently, it has been reported that the complexity of a molecule is a key criterion determining the success of the drug candidate [63]. The increase in molecular complexity, measured as the extent of bond saturation and the number of chiral centers, has been demonstrated to correlate with an overall improved compound developability. Changes in molecular complexity affect the three-dimensional shape of a compound, which might lead to improved interactions with the target receptor. The resulting improved potency and selectivity profile ultimately increases the chance of a successful drug candidate. Although aromatic rings and achiral centers still dominantly define classical drug structures, this might suggest a trend away from flat aromatic structures toward more complex molecules.

In the recent past, natural products and natural product-like molecules that lie outside the range of traditional “rule of five” druglike space have gained renewed interest in drug discovery [64, 65]. Technological advances have enabled the combination of approaches that leverage the unique diversity of building blocks from natural product sources with the strength of combinatorial library design. The diversity-oriented synthesis (DOS) approach allows the rapid synthesis of chemical libraries containing structurally complex molecules with a range of scaffold variations and chiral centers, creating a broad distribution of diverse compounds capable of binding a range of biological targets [66]. The main emphasis of natural product-like drug discovery so far, however, is on the identification of novel tool compounds to probe the target of interest and support further pharmacological *in vitro* studies, not on the development of oral drugs [67, 68]. Finally, macrocyclic molecules (containing a ring of seven or more atoms) represent another emerging structural class outside of classical oral druglike space, with a strong potential for historically difficult targets such as protein–protein interactions [69]. Macrocycles are capable of forming high-affinity interactions with the shallow contact surfaces that are typical for interfaces involved in protein–protein interactions. Due to their intrinsic conformational constraint, they can position arrays of functional groups across a wide interaction area, without the penalty of introducing multiple rotatable bonds.

Virtual screening provides an excellent opportunity to explore large databases of virtual small molecules and ring systems as highlighted above, supporting the design of combinatorial libraries with novel scaffolds or ring systems, or can be employed for tasks such as bioisosteric replacement design and scaffold hopping. However, in order to increase the chance of successfully synthesizing molecules proposed by virtual screening methods, more effort has to be put into the development of predictive methods to account for chemical feasibility. Not only should it include if a particular compound can be synthesized but it should also include if it can be rapidly followed up (i.e., chemically enabled) with analogues during lead optimization in a medicinal chemistry campaign. Computational approaches to assess synthetic accessibility have been described in the literature, mainly based on retrosynthetic or complexity-based analysis of molecules [70, 71].

## 1.5 Focused Libraries

Random screening of compound libraries in a high-throughput fashion is the major source of finding new leads in drug discovery when little or nothing is known about a target. Modern HTS technologies can routinely screen millions of compounds in a few weeks. However, in certain screening paradigms this approach may not always be feasible. For instance, the assay format cannot be scaled up to HTS format or only a low-throughput cell-based assay is available. In this situation, it is necessary to reduce the number of compounds to be screened, and the selection of diversity-based compound subsets is a viable option (Section 1.4). At the beginning of a new drug discovery project, it is often likely that information already exists to jump-start the program. When sufficient knowledge about a drug target is available, the design of targeted or *focused libraries* is possible. Various computational methods can "focus" the selection of compounds toward individual targets or entire target classes. The growing amount of knowledge for many drug targets and known drugs has resulted in an increased number of publications that describe new methods for the design of target and target family-focused libraries. Consequently, the screening of moderately sized ($10^4$–$10^5$) focused libraries has emerged as a promising alternative and efficient approach for lead identification.

### 1.5.1 Concepts of Focused Design

In their seminal work, Hopkins and coworkers mapped the entire known pharmacological space (biologically relevant chemical space) on the basis of a comprehensive collection of literature data [8]. This compound-centric view allowed them to identify targets for which drugs and chemical tools have been discovered, and how they are related to each other. Not surprisingly, the most densely populated target families represent attractive gene families that are actively pursued in drug discovery, notably kinases, GPCRs, ion channels, proteases, phosphatases, and nuclear hormone receptors. These multimember gene families, which account for more than 50% of the known human druggable genome, are also the prime candidates for targeted library design efforts due to their family-wide similarities in molecular recognition or enzymatic mechanisms.

A certain amount of prior knowledge about members of the target families is required for the successful design of focused libraries. Increased efforts in collecting and annotating pharmacological data of molecules are undertaken (Section 1.7). The availability of an ever-growing number of protein structures in the Protein Data Bank (PDB) facilitates the generation of structure-based knowledge of target classes.

Depending on the type and amount of information available for a target family, one can apply ligand-based, structure-based, or chemogenomics approaches for the design of focused libraries, or ideally all methods can be combined in a complementary fashion [72]. Interestingly, while GPCRs represent the second largest gene family (after kinases), they almost completely lack any structural information due to their membrane-bound nature. Only recently, the first crystal structures of human GPCR targets ($\beta_2$ adrenergic, $A_{2a}$ adenosine receptors) have been solved. Hence, the design of GPCR focused libraries in the past was often limited to ligand-based methods. The development of refined homology models and prediction of ligand binding modes start to compensate for the lack of GPCR structures, making structure-based methods more amenable to focused library design [73].

### 1.5.2 Ligand-Based Focused Design

The use of simple molecular properties has been demonstrated to construct models that discriminate compounds belonging to different target families. Depending on their target family, bioactive molecules show differences in their physicochemical properties [74, 75]. These variations can be used either as simple descriptor-based guidelines or to develop predictive models that focus the composition of chemical libraries toward a particular target family. For example, a neural network model was constructed on the basis of a small set of physicochemical properties to categorize compounds as GPCR-like or non-GPCR-like [76]. According to the model, GPCR-like compounds tend to be less flexible, less polar, and more hydrophobic compared to non-GPCR-like molecules. Other classification methods have been applied as well, such as Bayesian models, recursive partitioning, self-organizing maps (SOM), and other machine learning systems, in combination with various 2D or 3D descriptors [77, 78]. BCUT descriptors not only are being used for diversity-based subset selections (Section 1.4) but have also been employed to design focused libraries [79]. Instead of selecting diverse representatives from each cell in the low-dimensional descriptor space, compounds from "promising cells" are chosen that contain known active ligands.

Following a pharmacophore approach, a consensus 3D pharmacophore fingerprint based on a set of known GPCR ligands was created, which was subsequently used for building GPCR focused libraries [80]. Using a four-point pharmacophore fingerprint as a measure for similarity, GPCR-specific pharmacophores were identified and applied toward the design of focused libraries [81]. Particularly for large target families where a wealth of data exists, such as kinases and GPCRs, known ligands can be used to define queries to search against compound collections using substructure and similarity searches, and identified molecules can be compiled into focused screening collections [82, 83].

The notion that compounds with a benzodiazepine core were active as ligands against a variety of GPCRs, such as central and peripheral benzodiazepine, kappa opioid, and CCK receptors, has led to the *privileged structure* concept [84]. The observed activity crossover of a single compound to multiple targets of the same gene family implies that an underlying molecular scaffold is “privileged” to lend its activity to more than one receptor. The privileged structure concept has appeared more frequently in the recent literature and is interpreted in two different flavors [48]. In a strict sense, privileged structures should be defined as substructures with a proven correlation to a particular target family, based on specifically defined key structural elements that account for a commonality in molecular recognition across the target family members. In a broader context, privileged structures are interpreted as substructures emerging in compounds that showed effects on multiple target proteins, irrespective of their underlying target families. It is unknown why these fragments bind with higher than average frequency to multiple targets, and often there is no clear dividing line between privileged structures and promiscuous binders. Nevertheless, an increasing number of scaffolds have been described as privileged structures, mainly for GPCR and kinase targets, including indoles, biphenyls, benzopyranes, aryl piperazines, or aminopyrimidines [85].

Several groups have carried out fragmentation analyses of compounds active against target families, and the observed frequently recurring fragments were used as input for the design of focused libraries. In one of the first studies, a “retrosynthetic combinatorial analysis procedure” (RECAP) was developed using a limited set of defined fragmentation rules to identify privileged substructural elements [86]. Novel kinase inhibitors were designed in a reconstructive approach by fragmentation of known inhibitors and *de novo* assembly of fragments based on predictive models. The authors noted that while this approach worked well in designing active molecules, it remains to be challenging to also find selective kinase inhibitors [87]. A similar “virtual fragment linking” approach validated across a variety of different target classes has also been reported [88]. A combination of virtual and experimental (NMR, biochemical) screening methods was employed to identify novel scaffolds for the design of kinase-targeted libraries [89].

These and other reports give evidence of frequently occurring substructural elements connected to activity across a diverse panel of proteins. However, it is still debatable if such privileged structures are truly privileged in a selective manner against specific target families [90]. For instance, it is possible that fragments occurring with high frequency are simply elements of druglike, easy to synthesize, and therefore overrepresented molecules in compound collections. More thorough analyses are required that study the selectivity of privileged structures against target families and consider only statistically significant fragments by normalizing their occurrences in active compounds versus mere incidence in the screening libraries.

### 1.5.3
### Structure-Based Focused Design

Compared to ligand-based approaches, the use of structure-based methods has received less attention in the past, mainly due to their limited applicability to target classes with available protein structures and also because some of the methods such as docking still involve computationally intensive processes. Nevertheless, for certain target classes, particularly for kinases, significant amount of structural information is available, making structure-based approaches a viable option for compound selection and focused library design. The publication of the first human GPCR crystal structures, as well as the development of more refined homology models in combination with site-directed mutagenesis data and ligand structure–activity relationships (SAR), has also advanced the design of structure-based focused libraries for GPCR targets.

Protein–ligand docking and pharmacophore searching are the two main structure-based techniques for searching a large number of compounds to design focused libraries. Fast high-throughput docking methods allow the evaluation of large virtual libraries in a protein binding pocket, followed by the selection of compounds according to their docking scores. Due to the limited accuracy of scoring functions in general, it is advisable to also maintain a certain amount of diversity within the selected subset to facilitate the detection of potential new binding modes not predicted by docking [91]. The docking of combinatorial libraries based on a selected number of scaffolds, linkers, and functional groups allowed the design and synthesis of a target-focused virtual library while optimizing size versus diversity [47]. For designing a kinase-focused library, various scaffolds were prioritized by docking small virtual libraries generated around each core. Scaffolds that led to consistently docked structures according to cluster analysis were selected [92]. By docking privileged fragments derived from fragmentation of known kinase inhibitors into their respective protein binding pockets, both ligand-based and structural information was combined to assemble novel scaffold libraries for a kinase-focused library [93].

### 1.5.4 Chemogenomics Approaches

A new strategy in drug discovery termed *chemogenomics* has been defined as the investigation of classes of compounds against families of functionally related proteins [94]. It provides a unique approach to organize targets according to their gene families and discover ligands for related targets in a systematic manner, thus enhancing the efficiency of the drug discovery process [95]. Analogous to the similarity property principle (i.e., similar chemical structures share similar biological activities), the underlying assumption of chemogenomics is that similar biological structures share similar ligands. Indeed, it has often been observed that compounds active against one protein also show activity against other members of the same gene family. Chemogenomics attempts to link chemical structures of bioactive molecules and their effect on entire targets classes. Ultimately, insights into this relationship enable the rational design of focused libraries against one or multiple gene family members. Since the success of identifying and exploiting these links depends on the availability of structural and biological data for a larger number of targets, chemogenomics methods have been applied largely to not only major gene families, GPCRs and kinases mainly, but also proteases, nuclear hormone receptors, and ion channels.

Due to the lack of protein structures, similarities among GPCR targets have been mainly explored by sequence-based methods. The conventional approach of mapping the homology between GPCRs in a phylogenetic tree analysis employs full protein sequences. More relevant from a drug design perspective, however, is the comparison of targets based on their ligand recognition. Several groups have identified putative ligand binding sites within the transmembrane region of GPCRs [96, 97]. The relevant amino acids defining the pockets (20–50 residues) can be represented by their physicochemical properties (charged, polar, hydrophobic, aromatic, etc.) and are converted into bit strings unique for each receptor. This enables a straightforward comparison of ligand binding sites of targets across the entire gene family. Similarly, "structural interaction fingerprints" (SIFt) were developed for kinase binding sites, based on fingerprints derived from physicochemical properties of selected binding site residues [98].

This approach can be applied to find novel sources of ligands for GPCRs without existing leads by first evaluating similar GPCRs for which ligands are known. For example, active compounds for the CRTH2 receptor were identified by screening ligands from the structurally related AT1 and AT2 receptors [99]. Notably, the close relationship between the CRTH2 and AT1/AT2 receptors was not revealed by their full-length sequence homology. Small molecule antagonists for SSTR5 were identified by starting with a known histamine H1 antagonist, after a chemogenomics analysis had shown a close similarity between SSTR5 and opioid, histamine, dopamine, and serotonin receptors [100]. The same strategy has also been proposed to identify ligands for several orphan GPCRs [97].

A more elaborate chemogenomics approach, "thematic analysis," has been developed by Biofocus to better characterize the binding regions across a range of different GPCRs [101]. The putative ligand interaction site is divided into separate microenvironments with well-defined properties for each subsite (themes) and paired with matching fragments (motifs) based on known ligand structures. Combined, these thematic fingerprints can be used to classify new GPCRs and to design libraries focused on particular GPCR subclasses. A similar concept termed "chemoprint" has been applied to GPCRs by systematically annotating key interaction pairs of ligand fragments and their putative protein binding residues [102]. It is worth noting that in combination with homology modeling, the authors have established a method for deriving sequence-based 3D pharmacophore models for a wide range of GPCR targets, useful for virtual screening and focused library design.

A ligand-centric view of chemogenomics is the classification of biological targets based on activity profiles of diverse ligands [103]. These affinity fingerprints can then serve as a measure of protein similarity. It has been demonstrated that clustering of kinases based on their ligand SAR is different from the sequence-based clustering [104]. Both ligand- and structure-based classifications are often complementary and provide alternative views of the same protein family. The BioPrint database from Cerep comprises activity profiles of chemical structures experimentally measured across a large number of targets. Using experimental binding data generated from 2000 druglike compounds on 40 GPCR targets, a global QSAR model employing pharmacophore features relevant to GPCR binding characteristics has been developed and applied to design GPCR focused libraries [105].

## 1.6 Virtual Combinatorial Libraries and Fragment Spaces

With today's computational resources it is usually not a problem to exhaustively search compounds from corporate collections, vendor libraries, or small combinatorial libraries, which typically range in the order of $10^5$–$10^7$ molecules. However, for large virtual combinatorial libraries and collections thereof, it becomes quickly unfeasible to enumerate all possible virtual molecules in advance due to combinatorial explosion. Consequently, there has been an increasing interest in computational methods to find alternative ways to systematically search large virtual combinatorial libraries, allowing a dramatic expansion of unexplored chemical space.

A solution to the problem is to encode molecules of virtual combinatorial libraries not as enumerated products, but rather to keep them in their unenumerated form as building blocks (fragments), together with linkage rules for how to combine them. The efficiency of searching virtual combinatorial libraries in their *fragment space* representation compared to their enumerated products is easily explained by the different numbers of molecules that have to be processed during a similarity search. For instance, if a two-component combinatorial library with 1000 building blocks each is searched in its enumerated form, one million product structures have to be compared to a given query. In contrast, only 2000 monomers have to be evaluated in their corresponding fragment space. Extending this comparison to large sets of combinatorial libraries, the number of products can by far exceed $10^{12}$ possible virtual products, making a systematic enumeration unfeasible, whereas searching in the equivalent fragment space keeps the number still at a manageable size.

A large fragment space encoding over $10^{13}$ possible product structures has been created at Pfizer by taking a collection of 358 combinatorial libraries based on proprietary validated reaction protocols [106]. This fragment space can be systematically explored with the similarity search program FTrees-FS (extension of FTrees) that does not require the upfront enumeration of product structures [50, 107]. The result is a list of virtual products similar to the search query, synthetically accessible by one or more of the reaction protocols stored in the fragment space. Grouping the virtual hits by their synthetic protocols allows a fast follow-up design of focused libraries. A similar fragment space based on feature tree descriptors spanning $10^{11}$ potential products was generated using proprietary reaction protocols [108]. The selection of diverse sets of input reagents enabled both diverse and focused decorations of the central scaffolds identified by the search. In a related approach, a "monomer-based similarity searching" (MoBSS) method has been developed using atom pair descriptors [109]. To avoid time-consuming product enumeration, the descriptors are generated from the monomers of virtual combinatorial libraries collected in a fragment space. Since atom pair descriptors computed from interatomic distances lend themselves to pairwise additivity, product atom pairs can be rapidly computed from those of the constituent monomers through an arithmetic manipulation.

Fragment spaces have also been assembled by rule-based fragmentation of drug molecules. To build such a space, the RECAP program has been applied to drugs from the World Drug Index (WDI) and other druglike molecule collections [86, 110]. Applying a set of retrosynthetic rules, the drug molecules are disconnected into their individual fragments, and together with their associated link types they are incorporated into the fragment space. The large variety of generated fragments combined with a countless number of possible recombinations based on a limited set of connection rules permits a high degree of diversity of virtual products. However, this comes at the expense of their synthetic feasibility, which in contrast to hits from combinatorial fragment spaces is often unknown [111]. Predictive methods to assess synthetic feasibility of virtual *de novo* molecules have been proposed [112].

The calculation of "basis products" is an efficient way of representing large combinatorial libraries by strategically selecting subsets of compounds without complete enumeration. They are derived by enumerating one building block at a time while holding all other reaction components fixed, and this is consecutively repeated for all other building blocks. In the case of a two-component reaction with 1000 monomers each, only 1999 basis products have to be enumerated ($A_1B_1$, $A_2B_1$, $A_3B_1$, …, $A_{1000}B_1$, $A_1B_2$, $A_1B_3$, …, $A_1B_{1000}$). This way, any virtual product in the combinatorial library is represented by a set of basis products (e.g., $A_3B_7$ represented by $A_3B_1$ and $A_1B_7$). Basis products can be used in the same way as enumerated combinatorial library products, such as property filtering or docking [113]. However, the underlying assumption and limitation of this approach is that properties or docking scores from the various building blocks behave in an additive manner. A large number of virtual combinatorial libraries were also efficiently encoded by Bayesian modeling [114]. Representative members from a large number of combinatorial libraries are used to derive a multicategory Bayesian model. The resulting "Bayesian idea generator" (BIG) allows to predict the likelihood of a given compound belonging to a certain combinatorial library. Several top-ranking libraries are suggested and prioritized according to their Bayesian score.

Virtual combinatorial libraries can also be explored by structure-based virtual screening methods without prior enumeration. Docking of large virtual combinatorial libraries is offered by many docking programs, such as DREAM++, FlexX$^c$, CombiDOCK, and CombiGlide, among others [115]. Typically, the library templates are placed in the binding pocket, and each subsite is individually probed by docking the various R-groups of the library. Compounds with the highest scoring R-group combinations are selected and synthesized as part of the focused library. Notably, the FlexNovo software is capable of directly accessing fragment spaces generated as described above [116].

## 1.7 Databases of Chemical and Biological Information

Numerous databases of chemical structures, biological targets, and bioactivity data relevant for drug discovery have emerged in recent years. They are a unique source both for generating new ideas to identify chemical matter and for providing information-rich content for chemical and biological targets. Besides traditional compound databases with a large number of diverse chemical structures, there is an increased interest in annotated compound libraries aiming at capturing information to establish relationships between chemical matter and their biological function [117]. All relevant databases have been extensively reviewed in the literature. In the following, representative public and commercial databases relevant for drug discovery are highlighted (Table 1.2).

ZINC is a free database of commercially available compounds that in its current version contains over 13 million purchasable compounds from vendor catalogs [118]. Multiple formats are provided to fit the needs for various virtual screening applications such as substructure and similarity searches, property filtering, and docking. Subsets of druglike, leadlike, and fragment compounds have been generated and can be accessed separately. Similarly, ChemDB is a chemical database of over 5 million small molecules collected from electronic catalogs of commercial vendors. In addition, computational reaction models enable searches through virtual chemical space by predicting hypothetical products synthetically accessible from building blocks contained in the database [119]. Web-based chemistry search engines are capable of mining a large number of molecules from various public data repositories, such as ChemSpider (20 million) and eMolecules (7 million), the latter also offering databases for download to be used for virtual screening.

PubChem is a component of the NIH Molecular Libraries Roadmap Initiative providing information on biological activities of small molecules. It is organized as three interconnected databases, containing chemical samples from a variety of sources (PubChem Substance), compound information related to substances such as physicochemical properties and descriptors for similarity searching (PubChem Compound), and bioactivity data of chemical substances (PubChem BioAssay). The dynamically growing primary databases contain over 61 million records of chemical substances, 25 million unique compound structures, and bioactivity data from more than 1600 assays. Data mining approaches to create representative subsets for virtual screening purposes [120] and cross-assay analyses of bioactivity data to study polypharmacology behavior in the PubChem collection [121] have been described. Other publicly available compound databases annotated with biological data are ChemBank with over 1.2 million chemicals [122] and DrugBank covering almost 4800 drugs [123].

Several commercial databases of annotated compound libraries exist, mostly compiled from literature and patent sources. The StARLITe database (now ChEMBL) is a large collection of chemicals mined from literature, including target and bioactivity information for 500 000 compounds. The WOMBAT (World of Molecular Bioactivity) database from Sunset Molecular contains 300 000 molecular entries associated with biological activities and target information [124]. Jubilant BioSys, GVK Bio, and Aureus Pharma are commercial providers of large target-centric compound databases, focusing mostly on large target classes such as kinases, GPCRs, nuclear hormone receptors, or ion channels. The databases integrate chemical structures with activity data and target information collected from literature and published patents.

Nowadays, more than 60% of new chemical substances entering the Chemical Abstract Service (CAS) registry are sourced from patents. Thus, in addition to

**Table 1.2** Representative databases of chemical and biological information relevant for drug discovery.

| Database | Provider | Coverage | Website |
|---|---|---|---|
| ChemSpider | RSC | 20 million chemicals | http://www.chemspider.com |
| eMolecules | eMolecules Inc. | 7 million chemicals | http://www.emolecules.com |
| ZINC 8 | UCSF | 13 million chemicals | http://zinc.docking.org |
| ChemDB | UC Irvine | 5 million chemicals | http://cdb.ics.uci.edu |
| ChemBank | Broad Institute | 1.2 million chemicals | http://chembank.broadinstitute.org |
| DrugBank | University Alberta | 4800 drugs; 2500 targets | http://www.drugbank.ca |
| PubChem | NIH Molecular Libraries Roadmap Initiative | 61 million substances; 25 million chemicals; 1600 bioassays | http://pubchem.ncbi.nlm.nih.gov |
| ChEMBL (StARLITe) | EMBL-EBI (Inpharmatica) | 500 000 chemicals; 5000 targets | http://www.ebi.ac.uk/chembldb |
| WOMBAT 2009.1 | Sunset Molecular | 300 000 chemicals; 2000 targets | http://www.sunsetmolecular.com |
| BioPrint | Cerep | 2500 drugs; 159 bioassays | http://www.cerep.fr |
| ChemBioBase | Jubilant Biosys | 2 million chemicals; 1500 targets | http://www.jubilantbiosys.com |
| MedChem | GVK Bio | 1 million chemicals; 5600 targets | http://www.gvkbio.com |
| Target inhibitor | GVK Bio | 3.5 million chemicals (patent and literature) | http://www.gvkbio.com |
| AurSCOPE | Aureus Pharma | 500 000 chemicals; 1.7 million activities | http://www.aureus-pharma.com |

capturing information from literature and available compound databases, the mining of chemical and biological data from the patent literature is attracting considerable attention. Not only does it provide a mostly untapped source of ideas for new lead generation but it also allows to identify regions in chemical space already investigated. The use of text analytics tools is an efficient way to mine the drug and patent literature [125]. Researchers at IBM have developed a system that enables the user to mine patents from the US Patent corpus [126]. A chemical annotator recognizes and extracts chemical entities in patent documents, and a name-to-structure converter generates molecular structures that are stored in a database for similarity searching. The authors were able to index 3.6 million unique chemical structures from 4.4 million patents. A similar patent database, SureChem from Reel Two, was created by extracting all chemical names from the full text of US, EPO, JP, and WO patents and contains 11 million unique chemical structures covering 18 million patents.

A limiting factor is that text analytics methods are largely confined to specific compounds exemplified in the patents, which are only a small portion of the theoretically possible chemical structures represented in the Markush claim. The improved access to searchable databases of Markush structures and the development of sophisticated chemoinformatics tools to efficiently mine and enumerate the potentially billions of claimed chemical structures are the next logical steps toward capturing the vast chemical space contained in the patent corpus [127, 128].

## 1.8 Conclusions and Outlook

Virtual screening increasingly impacts the hit finding process in drug discovery by preselecting compounds for biological evaluation. Due to a high false positive rate associated with most virtual screening methods, the selection of only few cherry-picked compounds to identify active molecules (“needle in the haystack”) is often less likely to be successful. It rather plays to its strength when virtual screening is applied in the context of narrowing down the number of compounds to be tested by enriching screening sets with drug- and leadlike compounds likely to be active, that is, molecules in biologically and medicinally relevant chemical space. Eliminating compounds with nondruglike and nonleadlike properties from a screening collection is often not considered as virtual screening, but it is a crucial factor contributing to the overall success of high-quality lead identification. Reducing the number of compounds for efficient biological testing can be accomplished by carefully applying diversity-based selection criteria. The design of focused libraries targeting a specific protein or protein family is a proven method of choice to increase the chances of finding active leads. The unique capability of virtual screening to search compounds in their virtual form not only allows access to the small fraction of chemical space represented by existing screening libraries but also allows to expand into other regions of chemical space (Figure 1.2). Virtual screening of compound collections from external sources (vendors, patents, and literature), the design of large virtual combinatorial libraries and their efficient representation as frag-

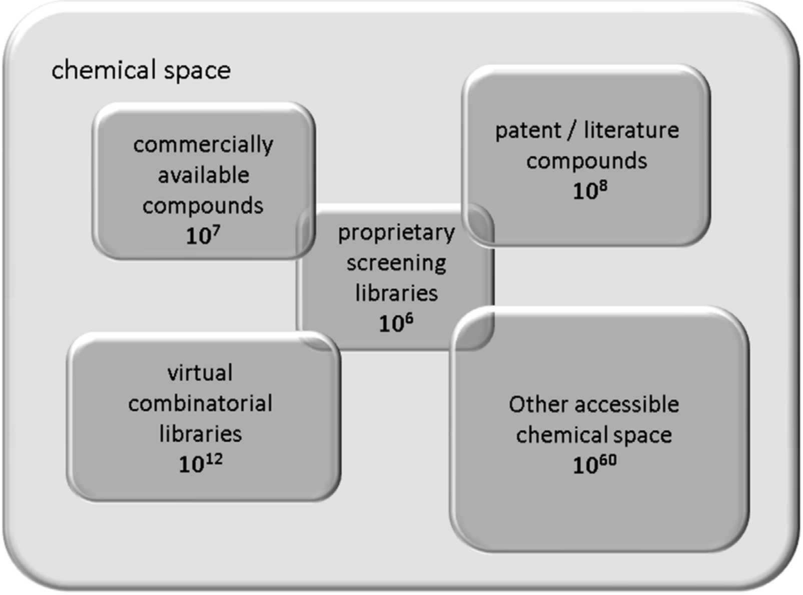
chemical space
commercially available compounds $10^7$
patent / literature compounds $10^8$
proprietary screening libraries $10^6$
virtual combinatorial libraries $10^{12}$
Other accessible chemical space $10^{60}$

**Figure 1.2** Virtual screening has the capability to expand searches outside of typical screening libraries (amenable to HTS) into new dimensions of chemical space. The various sources of accessible chemical spaces are described throughout the text. Typical or estimated numbers of compounds are highlighted in bold.

ment spaces, or recently emerging alternative sources of diverse chemical matter (*de novo* enumerated small molecules or ring scaffolds) offer plenty of opportunities. However, the future challenge remains to more reliably predict biological activity and chemical feasibility of compounds being proposed for synthesis. Ultimately, the goal of next-generation virtual screening methods is the identification and systematic exploration of truly synthetically accessible and biologically and medicinally relevant chemical space.

## 1.9 Glossary

| | |
|---|---|
| Chemical library | Collection of chemical compounds. |
| Chemical space | Collection of all possible meaningful compounds, typically restricted to small organic molecules. |
| Chemogenomics | Discovery and description of classes of compounds against families of functionally related proteins. |
| Combinatorial chemistry | Generation of large collections of compound libraries by systematic combination of smaller building blocks. Large virtual combinatorial libraries are often created in the form of fragment spaces. |

Diversity-based library
Compound library designed to create a maximally diverse collection of compounds to cover a broad range of protein classes, especially when they are unknown or difficult to target.

Druglikeness
Physicochemical properties to improve the likelihood of success in drug development.

Focused library
Compound library designed around selected, often privileged scaffolds tailored toward targeting specific protein families (GPCRs, kinases, etc.).

Fragment space
Chemical space representation where molecules are encoded as building blocks (fragments) and linkage rules.

Leadlikeness
Criteria for ideal lead molecules that serve as a basis for further chemical optimization in a medicinal chemistry campaign.

Molecular diversity
Quantitative description how different molecules are from each other in a compound collection.

Privileged structure
A single molecular framework or frequently occurring fragment able to provide ligands for multiple receptors, often within a protein family.

Similarity property principle
Fundamental assumption that similar compounds typically exhibit similar biological activity; also referred to as neighborhood behavior.

### Acknowledgment

The author would like to thank David Hepworth for fruitful discussions and critical reading of the manuscript.

## References

1 Lipinski, C. and Hopkins, A. (2004) Navigating chemical space for biology and medicine. *Nature*, **432**, 855–861.

2 Dobson, C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.

3 Medina-Franco, J.L., Martinez-Mayorga, K., Giulianotti, M.A., Houghten, R.A., and Pinilla, C. (2008) Visualization of the chemical space in drug discovery. *Current Computer-Aided Drug Design*, **4**, 322–333.

4 Gorse, A.D. (2006) Diversity in medicinal chemistry space. *Current Topics in Medicinal Chemistry*, **6**, 3–18.

5 Blum, L.C. and Reymond, J.L. (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, **131**, 8732–8733.

6 Bohacek, R.S., McMartin, C., and Guida, W.C. (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal Research Reviews*, **16**, 3–50.

7 Hert, J., Irwin, J.J., Laggner, C., Keiser, M.J., and Shoichet, B.K. (2009) Quantifying biogenic bias in screening libraries. *Nature Chemical Biology*, **5**, 479–483.

8 Paolini, G.V., Shapland, R.H., van Hoorn, W.P., Mason, J.S., and Hopkins, A.L. (2006) Global mapping of pharmacological space. *Nature Biotechnology*, **24**, 805–815.

9 Arkin, M.R. and Wells, J.A. (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Reviews. Drug Discovery*, **3**, 301–317.

10 Milne, G.M. (2003) Pharmaceutical productivity: the imperative for new paradigms. *Annual Reports in Medicinal Chemistry*, **38**, 383–396.

11 Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **23**, 3–25.

12 Leeson, P.D. and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews. Drug Discovery*, **6**, 881–890.

13 Vistoli, G., Pedretti, A., and Testa, B. (2008) Assessing drug-likeness: what are we missing? *Drug Discovery Today*, **13**, 285–294.

14 Teague, S.J., Davis, A.M., Leeson, P.D., and Oprea, T. (1999) The design of leadlike combinatorial libraries. *Angewandte Chemie. International Edition in English*, **38**, 3743–3748.

15 Keserü, G.M. and Makara, G.M. (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nature Reviews. Drug Discovery*, **8**, 203–212.

16 Hann, M.M., Leach, A.R., and Harper, G. (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of Chemical Information and Computer Sciences*, **41**, 856–864.

17 Rishton, G.M. (2003) Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today*, **8**, 86–96.

18 Rishton, G.M. (2008) Molecular diversity in the context of leadlikeness: compound properties that enable effective biochemical screening. *Current Opinion in Chemical Biology*, **12**, 340–351.

19 Rishton, G.M. (1997) Reactive compounds and *in vitro* false positives in HTS. *Drug Discovery Today*, **2**, 382–384.

20 Roche, O., Schneider, P., Zuegge, J., Guba, W., Kansy, M., Alanine, A., Bleicher, K., Danel, F., Gutknecht, E.M., Rogers-Evans, M., Neidhart, W., Stalder, H., Dillon, M., Sjogren, E., Fotouhi, N., Gillespie, P., Goodnow, R., Harris, W., Jones, P., Taniguchi, M., Tsujii, S., von der Saal, W., Zimmermann, G., and Schneider, G. (2002) Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *Journal of Medicinal Chemistry*, **45**, 137–142.

21 McGovern, S.L., Helfand, B.T., Feng, B., and Shoichet, B.K. (2003) A specific mechanism of nonspecific inhibition. *Journal of Medicinal Chemistry*, **46**, 4265–4272.

22 Feng, B.Y., Simeonov, A., Jadhav, A., Babaoglu, K., Inglese, J., Shoichet, B.K., and Austin, C.P. (2007) A high-throughput screen for aggregation-based inhibition in a large compound library. *Journal of Medicinal Chemistry*, **50**, 2385–2390.

23 Hann, M.M. and Oprea, T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology*, **8**, 255–263.

24 Martin, Y.C. (2001) Diverse viewpoints on computational aspects of molecular diversity. *Journal of Combinatorial Chemistry*, **3**, 231–250.

25 Roth, H.J. (2005) There is no such thing as 'diversity'! *Current Opinion in Chemical Biology*, **9**, 293–295.

26 Johnson, M.A. and Maggiora, G.M. (1990) *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, Inc., New York.

27 Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., and Weinberger, L.E. (1996) Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *Journal of Medicinal Chemistry*, **39**, 3049–3059.

28 Martin, Y.C., Kofron, J.L., and Traphagen, L.M. (2002) Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, **45**, 4350–4358.

29 Muchmore, S.W., Debe, D.A., Metz, J.T., Brown, S.P., Martin, Y.C., and Hajduk,

P.J. (2008) Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *Journal of Chemical Information and Modeling*, **48**, 941–948.

**30** Harper, G., Pickett, S.D., and Green, D.V. (2004) Design of a compound screening collection for use in high throughput screening. *Combinatorial Chemistry and High Throughput Screening*, **7**, 63–70.

**31** Todeschini, R. (2009) *Molecular Descriptors for Cheminformatics*, Wiley-VCH Verlag GmbH, Weinheim.

**32** Todeschini, R. (2009) Dragon descriptors. Talete srl, Milan, Italy.

**33** Gillet, V.J., Willett, P., John, B.T., and David, J.T. (2007) Compound selection using measures of similarity and dissimilarity, in *Comprehensive Medicinal Chemistry II*, Elsevier, Oxford, pp. 167–192.

**34** Holliday, J.D., Hu, C.Y., and Willett, P. (2002) Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Combinatorial Chemistry and High Throughput Screening*, **5**, 155–166.

**35** Chen, X. and Reynolds, C.H. (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of Chemical Information and Computer Sciences*, **42**, 1407–1414.

**36** Willett, P. (1999) Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *Journal of Computational Biology*, **6**, 447–457.

**37** Ashton, M., Barnard, J., Casset, F., Charlton, M., Downs, G., Gorse, D., Holliday, J., Lahana, R., and Willett, P. (2002) Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quantitative Structure–Activity Relationship*, **21**, 598–604.

**38** Gobbi, A. and Lee, M.L. (2003) DISE: directed sphere exclusion. *Journal of Chemical Information and Computer Sciences*, **43**, 317–323.

**39** Downs, G.M. and Barnard, J.M. (2003) Clustering methods and their uses in computational chemistry, in *Reviews in Computational Chemistry*, vol. **18** (eds B. Kenny and D.B.B. Lipkowitz), John Wiley & Sons, Inc., New York, pp. 1–40.

**40** Lewis, R.A., Mason, J.S., and McLay, I.M. (1997) Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. *Journal of Chemical Information and Computer Sciences*, **37**, 599–614.

**41** Pearlman, R.S. and Smith, K.M. (1999) Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information and Computer Sciences*, **39**, 28–35.

**42** Oprea, T.I. and Gottfries, J. (2001) Chemography: the art of navigating in chemical space. *Journal of Combinatorial Chemistry*, **3**, 157–166.

**43** Gillet, V.J., Willett, P., Bradshaw, J., and Green, D.V.S. (1998) Selecting combinatorial libraries to optimize diversity and physical properties. *Journal of Chemical Information and Computer Sciences*, **39**, 169–177.

**44** Agrafiotis, D.K. (2002) Multiobjective optimization of combinatorial libraries. *Journal of Computer-Aided Molecular Design*, **16**, 335–356.

**45** Krier, M., Bret, G., and Rognan, D. (2006) Assessing the scaffold diversity of screening libraries. *Journal of Chemical Information and Modeling*, **46**, 512–524.

**46** Harper, G. and Pickett, S.D. (2006) Methods for mining HTS data. *Drug Discovery Today*, **11**, 694–699.

**47** Krier, M., Araujo-Junior, J.X., Schmitt, M., Duranton, J., Justiano-Basaran, H., Lugnier, C., Bourguignon, J.J., and Rognan, D. (2005) Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor. *Journal of Medicinal Chemistry*, **48**, 3816–3822.

**48** Muller, G. (2003) Medicinal chemistry of target family-directed masterkeys. *Drug Discovery Today*, **8**, 681–691.

**49** Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, **39**, 2887–2893.

**50** Rarey, M. and Dixon, J.S. (1998) Feature trees: a new molecular similarity measure based on tree matching. *Journal of Computer-Aided Molecular Design*, **12**, 471–490.

**51** Martin, Y.C. and Muchmore, S. (2009) Beyond QSAR: lead hopping to different structures. *QSAR & Combinatorial Science*, **28**, 797–801.

**52** Xu, Y.J. and Johnson, M. (2002) Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *Journal of Chemical Information and Computer Sciences*, **42**, 912–926.

**53** Schuffenhauer, A., Ertl, P., Roggo, S., Wetzel, S., Koch, M.A., and Waldmann, H. (2007) The scaffold tree: visualization of the scaffold universe by hierarchical scaffold classification. *Journal of Chemical Information and Modeling*, **47**, 47–58.

**54** Wetzel, S., Klein, K., Renner, S., Rauh, D., Oprea, T.I., Mutzel, P., and Waldmann, H. (2009) Interactive exploration of chemical space with Scaffold Hunter. *Nature Chemical Biology*, **5**, 581–583.

**55** Renner, S., van Otterlo, W.A., Dominguez Seoane, M., Mocklinghoff, S., Hofmann, B., Wetzel, S., Schuffenhauer, A., Ertl, P., Oprea, T.I., Steinhilber, D., Brunsveld, L., Rauh, D., and Waldmann, H. (2009) Bioactivity-guided mapping and navigation of chemical space. *Nature Chemical Biology*, **5**, 585–592.

**56** Fink, T. and Reymond, J.L. (2007) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *Journal of Chemical Information and Modeling*, **47**, 342–353.

**57** Nguyen, K.T., Syed, S., Urwyler, S., Bertrand, S., Bertrand, D., and Reymond, J.L. (2008) Discovery of NMDA glycine site inhibitors from the chemical universe database GDB. *ChemMedChem*, **3**, 1520–1524.

**58** Lipkus, A.H. (2001) Exploring chemical rings in a simple topological-descriptor space. *Journal of Chemical Information and Computer Sciences*, **41**, 430–438.

**59** Lewell, X.Q., Jones, A.C., Bruce, C.L., Harper, G., Jones, M.M., McLay, I.M., and Bradshaw, J. (2003) Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *Journal of Medicinal Chemistry*, **46**, 3257–3274.

**60** Ertl, P., Jelfs, S., Muhlbacher, J., Schuffenhauer, A., and Selzer, P. (2006) Quest for the rings. *In silico* exploration of ring universe to identify novel bioactive heteroaromatic scaffolds. *Journal of Medicinal Chemistry*, **49**, 4568–4573.

**61** Pitt, W.R., Parry, D.M., Perry, B.G., and Groom, C.R. (2009) Heteroaromatic rings of the future. *Journal of Medicinal Chemistry*, **52**, 2952–2963.

**62** Ritchie, T.J. and Macdonald, S.J.F. (2009) The impact of aromatic ring count on compound developability: are too many aromatic rings a liability in drug design? *Drug Discovery Today*, **14**, 1011–1020.

**63** Lovering, F., Bikker, J., and Humblet, C. (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *Journal of Medicinal Chemistry*, **52**, 6752–6756.

**64** Koehn, F.E. and Carter, G.T. (2005) The evolving role of natural products in drug discovery. *Nature Reviews. Drug Discovery*, **4**, 206–220.

**65** Zhang, M.Q. and Wilkinson, B. (2007) Drug discovery beyond the 'rule-of-five'. *Current Opinion in Biotechnology*, **18**, 478–488.

**66** Arya, P., Joseph, R., Gan, Z., and Rakic, B. (2005) Exploring new chemical space by stereocontrolled diversity-oriented synthesis. *Chemistry and Biology*, **12**, 163–180.

**67** Schreiber, S.L. (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science*, **287**, 1964–1969.

**68** Tan, D.S. (2005) Diversity-oriented synthesis: exploring the intersections between chemistry and biology. *Nature Chemical Biology*, **1**, 74–84.

**69** Driggers, E.M., Hale, S.P., Lee, J., and Terrett, N.K. (2008) The exploration of macrocycles for drug discovery:

an underexploited structural class. *Nature Reviews. Drug Discovery*, **7**, 608–624.

**70** Baber, J.C. and Feher, M. (2004) Predicting synthetic accessibility: application in drug discovery and development. *Mini Reviews in Medicinal Chemistry*, **4**, 681–692.

**71** Ertl, P. and Schuffenhauer, A. (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, **1**, 8.

**72** Miller, J.L. (2006) Recent developments in focused library design: targeting gene-families. *Current Topics in Medicinal Chemistry*, **6**, 19–29.

**73** Gregori-Puigjane, E. and Mestres, J. (2008) Coverage and bias in chemical library design. *Current Opinion in Chemical Biology*, **12**, 359–365.

**74** Morphy, R. (2006) The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. *Journal of Medicinal Chemistry*, **49**, 2969–2978.

**75** Vieth, M. and Sutherland, J.J. (2006) Dependence of molecular properties on proteomic family for marketed oral drugs. *Journal of Medicinal Chemistry*, **49**, 3451–3453.

**76** Balakin, K.V., Tkachenko, S.E., Lang, S.A., Okun, I., Ivashchenko, A.A., and Savchuk, N.P. (2002) Property-based design of GPCR-targeted library. *Journal of Chemical Information and Computer Sciences*, **42**, 1332–1342.

**77** Rognan, D. (2007) Chemogenomic approaches to rational drug design. *British Journal of Pharmacology*, **152**, 38–52.

**78** Schneider, P., Tanrikulu, Y., and Schneider, G. (2009) Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. *Current Medicinal Chemistry*, **16**, 258–266.

**79** Lavrador, K., Murphy, B., Saunders, J., Struthers, S., Wang, X., and Williams, J. (2004) A screening library for peptide activated G-protein coupled receptors. 1. The test set. *Journal of Medicinal Chemistry*, **47**, 6864–6874.

**80** Lamb, M.L., Bradley, E.K., Beaton, G., Bondy, S.S., Castellino, A.J., Gibbons, P.A., Suto, M.J., and Grootenhuis, P.D. (2004) Design of a gene family screening library targeting G-protein coupled receptors. *Journal of Molecular Graphics and Modelling*, **23**, 15–21.

**81** Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C., and Labaudiniere, R.F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *Journal of Medicinal Chemistry*, **42**, 3251–3264.

**82** Sun, D., Chuaqui, C., Deng, Z., Bowes, S., Chin, D., Singh, J., Cullen, P., Hankins, G., Lee, W.C., Donnelly, J., Friedman, J., and Josiah, S. (2006) A kinase-focused compound collection: compilation and screening strategy. *Chemical Biology and Drug Design*, **67**, 385–394.

**83** Decornez, H., Gulyas-Forro, A., Papp, A., Szabo, M., Sarmay, G., Hajdu, I., Cseh, S., Dorman, G., and Kitchen, D.B. (2009) Design, selection, and evaluation of a general kinase-focused library. *ChemMedChem*, **4**, 1273–1278.

**84** Evans, B.E., Rittle, K.E., Bock, M.G., DiPardo, R.M., Freidinger, R.M., Whitter, W.L., Lundell, G.F., Veber, D.F., Anderson, P.S., Chang, R.S. *et al.* (1988) Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *Journal of Medicinal Chemistry*, **31**, 2235–2246.

**85** DeSimone, R.W., Currie, K.S., Mitchell, S.A., Darrow, J.W., and Pippin, D.A. (2004) Privileged structures: applications in drug discovery. *Combinatorial Chemistry and High Throughput Screening*, **7**, 473–494.

**86** Lewell, X.Q., Judd, D.B., Watson, S.P., and Hann, M.M. (1998) RECAP: retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of*

*Chemical Information and Computer Sciences*, **38**, 511–522.

87 Vieth, M., Erickson, J., Wang, J., Webster, Y., Mader, M., Higgs, R., and Watson, I. (2009) Kinase inhibitor data modeling and *de novo* inhibitor design with fragment approaches. *Journal of Medicinal Chemistry*, **52**, 6456–6466.

88 Crisman, T.J., Bender, A., Milik, M., Jenkins, J.L., Scheiber, J., Sukuru, S.C., Fejzo, J., Hommel, U., Davies, J.W., and Glick, M. (2008) "Virtual fragment linking": an approach to identify potent binders from low affinity fragment hits. *Journal of Medicinal Chemistry*, **51**, 2481–2491.

89 Akritopoulou-Zanze, I. and Hajduk, P.J. (2009) Kinase-targeted libraries: the design and synthesis of novel, potent, and selective kinase inhibitors. *Drug Discovery Today*, **14** 291–297.

90 Schnur, D.M., Hermsmeier, M.A., and Tebben, A.J. (2006) Are target-family-privileged substructures truly privileged? *Journal of Medicinal Chemistry*, **49**, 2000–2009.

91 Jacoby, E., Schuffenhauer, A., Popov, M., Azzaoui, K., Havill, B., Schopfer, U., Engeloch, C., Stanek, J., Acklin, P., Rigollier, P., Stoll, F., Koch, G., Meier, P., Orain, D., Giger, R., Hinrichs, J., Malagu, K., Zimmermann, J., and Roth, H.J. (2005) Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Current Topics in Medicinal Chemistry*, **5**, 397–411.

92 Lowrie, J.F., Delisle, R.K., Hobbs, D.W., and Diller, D.J. (2004) The different strategies for designing GPCR and kinase targeted libraries. *Combinatorial Chemistry and High Throughput Screening*, **7**, 495–510.

93 Aronov, A.M. and Bemis, G.W. (2004) A minimalist approach to fragment-based ligand design using common rings and linkers: application to kinase inhibitors. *Proteins*, **57**, 36–50.

94 Kubinyi, H. and Müller, G. (2005) Chemogenomics in drug discovery: a medicinal chemistry perspective, in *Methods and Principles in Medicinal Chemistry* (eds R. Mannhold, H., Kubinyi, and G. Folkers) Wiley-VCH Verlag GmbH, Weinheim.

95 Harris, C.J. and Stevens, A.P. (2006) Chemogenomics: structuring the drug discovery process to gene families. *Drug Discovery Today*, **11**, 880–888.

96 Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E. (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of Chemical Information and Computer Sciences*, **43**, 391–405.

97 Surgand, J.S., Rodrigo, J., Kellenberger, E., and Rognan, D. (2006) A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins*, **62**, 509–538.

98 Deng, Z., Chuaqui, C., and Singh, J. (2004) Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein–ligand binding interactions. *Journal of Medicinal Chemistry*, **47**, 337–344.

99 Frimurer, T.M., Ulven, T., Elling, C.E., Gerlach, L.O., Kostenis, E., and Hogberg, T. (2005) A physicogenetic method to assign ligand-binding relationships between 7TM receptors. *Bioorganic and Medicinal Chemistry Letters*, **15**, 3707–3712.

100 Martin, R.E., Green, L.G., Guba, W., Kratochwil, N., and Christ, A. (2007) Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: a chemogenomics approach. *Journal of Medicinal Chemistry*, **50**, 6291–6294.

101 Crossley, R. (2004) The design of screening libraries targeted at G-protein coupled receptors. *Current Topics in Medicinal Chemistry*, **4**, 581–588.

102 Klabunde, T., Giegerich, C., and Evers, A. (2009) Sequence-derived three-dimensional pharmacophore models for G-protein-coupled receptors and their application in virtual screening. *Journal of Medicinal Chemistry*, **52**, 2923–2932.

103 Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., and Shoichet, B.K. (2007) Relating protein

pharmacology by ligand chemistry. *Nature Biotechnology*, **25**, 197–206.

**104** Vieth, M., Sutherland, J.J., Robertson, D.H., and Campbell, R.M. (2005) Kinomics: characterizing the therapeutically validated kinase space. *Drug Discovery Today*, **10**, 839–846.

**105** Gozalbes, R., Rolland, C., Nicolaï, E., Paugam, M.-F., Coussy, L., Horvath, D., Barbosa, F., Mao, B., Revah, F., and Froloff, N. (2005) QSAR strategy and experimental validation for the development of a GPCR focused library. *QSAR & Combinatorial Science*, **24**, 508–516.

**106** Boehm, M., Wu, T.Y., Claussen, H., and Lemmen, C. (2008) Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *Journal of Medicinal Chemistry*, **51**, 2468–2480.

**107** Rarey, M. and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design*, **15**, 497–520.

**108** Lessel, U., Wellenzohn, B., Lilienthal, M., and Claussen, H. (2009) Searching fragment spaces with feature trees. *Journal of Chemical Information and Modeling*, **49**, 270–279.

**109** Yu, N. and Bakken, G.A. (2009) Efficient exploration of large combinatorial chemistry spaces by monomer-based similarity searching. *Journal of Chemical Information and Modeling*, **49**, 745–755.

**110** Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. (2008) On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, **3**, 1503–1507.

**111** Mauser, H. and Stahl, M. (2007) Chemical fragment spaces for *de novo* design. *Journal of Chemical Information and Modeling*, **47**, 318–324.

**112** Schurer, S.C., Tyagi, P., and Muskal, S.M. (2005) Prospective exploration of synthetically feasible, medicinally relevant chemical space. *Journal of Chemical Information and Modeling*, **45**, 239–248.

**113** Zhou, J.Z., Shi, S., Na, J., Peng, Z., and Thacher, T. (2009) Combinatorial library-based design with Basis Products. *Journal of Computer-Aided Molecular Design*, **23** (10), 725–736.

**114** van Hoorn, W.P. and Bell, A.S. (2009) Searching chemical space with the Bayesian Idea Generator. *Journal of Chemical Information and Modeling*, **49**, 2211–2220.

**115** Ghosh, S., Nie, A., An, J., and Huang, Z. (2006) Structure-based virtual screening of chemical libraries for drug discovery. *Current Opinion in Chemical Biology*, **10**, 194–202.

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

**117** Oprea, T.I. and Tropsha, A. (2006) Target, chemical and bioactivity databases: integration is key. *Drug Discovery Today: Technologies*, **3**, 357–365.

**118** Irwin, J.J. and Shoichet, B.K. (2005) ZINC: a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, **45**, 177–182.

**119** Chen, J.H., Linstead, E., Swamidass, S.J., Wang, D., and Baldi, P. (2007) ChemDB update: full-text search and virtual chemical space. *Bioinformatics*, **23**, 2348–2351.

**120** Xie, X.Q. and Chen, J.Z. (2008) Data mining a small molecule drug screening representative subset from NIH PubChem. *Journal of Chemical Information and Modeling*, **48**, 465–475.

**121** Chen, B., Wild, D., and Guha, R. (2009) PubChem as a source of polypharmacology. *Journal of Chemical Information and Modeling*, **49**, 2044–2055.

**122** Seiler, K.P., George, G.A., Happ, M.P., Bodycombe, N.E., Carrinski, H.A., Norton, S., Brudz, S., Sullivan, J.P., Muhlich, J., Serrano, M., Ferraiolo, P., Tolliday, N.J., Schreiber, S.L., and Clemons, P.A. (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Research*, **36**, D351–D359.

**123** Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006) DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Research*, **34**, D668–D672.

**124** Olah, M., Mracec, M., Ostopovici, L., Rad, R., Bora, A., Hadaruga, N., Olah, I., Banda, M., Simon, Z., Mracec, M., and Oprea, T.I. (2005) WOMBAT: World of Molecular Bioactivity, in *Chemoinformatics in Drug Discovery* (ed. T.I. Oprea), Wiley-VCH Verlag GmbH, Weinheim, pp. 221–239.

**125** Banville, D.L. (2006) Mining chemical structural information from the drug literature. *Drug Discovery Today*, **11**, 35–42.

**126** Rhodes, J., Boyer, S., Kreulen, J., Chen, Y., and Ordonez, P. (2007) Mining patents using molecular similarity search. Proceedings of the Pacific Symposium on Biocomputing, pp. 304–315.

**127** Barnard, J.M. and Wright, P.M. (2009) Towards in-house searching of Markush structures from patents. *World Patent Information*, **31**, 97–103.

**128** Fliri, A., Moysan, E., Benichou, P., and Nolte, M. (2009) Methods for processing generic chemical structure representations. Patent WO09051741.