1

ELEMENTS FOR THE DEVELOPMENT OF STRATEGIES FOR COMPOUND LIBRARY ENHANCEMENT

EDGAR JACOBY

Janssen Research & Development, Beerse, Belgium

1.1 INTRODUCTION

The main purpose of a small molecule compound collection that is sometimes considered to constitute the crown jewels of a drug discovery organization is to supply the discovery pipeline with hit-to-lead compounds for today's and the future's portfolio of drug discovery programs and to provide tool compounds for the investigation of biological targets and pathways [1–7]. Independent of the followed discovery strategy relying on diversity or hypothesis-based screening, the automated access to high-quality compounds constitutes a key asset [8]. Accordingly, all major organizations, including the National Institutes of Health (NIH) and the European Union Innovative Medicines Initiative (EU IMI), have initiated over the last years dedicated compound collection enhancement projects [9]. In alignment with the general paradigm shift observed in drug discovery, going from quantity to quality, the fundamental principle aims to select both—at the chemical and the biological level—the best possible molecular starting points for lead discovery and development in the early drug discovery phases in order to reduce attrition at later preclinical and clinical stages.

To be successful on the long-term perspective, such design strategy addresses the known target space and tries to expand into nonprecedented areas of chemical and biological spaces using diversity principles [5, 6]. Directing the molecular properties toward the lead-like space is expected to improve overall success rates. The application of absorption, distribution, metabolism, excretion, and toxicity (ADMET) property models and rules of thumb aims to reduce the attrition risk and can be front-loaded into the design

Small Molecule Medicinal Chemistry: Strategies and Technologies, First Edition.
Edited by Werngard Czechtizky and Peter Hamley.

of the collection. On the other hand, a screening collection should allow for serendipitous discovery going in hand with diversity designs.

Drug discovery compound collections have evolved during recent history. Up to the early 1990s when drug discovery was mainly conduced in phenotypic *in vivo* screening of corporate medicinal chemistry compounds, the collections were limited to a few thousands of compounds that were carefully generated within the individual therapeutic programs. With the advances of molecular and cell biology and the advent of high-throughput chemistry and screening, the drug discovery world changed and compound collections were grown in the last 15 years to pass in a number of organizations beyond the one million number. Today, screening collections integrate design-focused and diversity-based compound sets from the synthetic and natural paradigms generated via corporate medicinal chemistry and combinatorial compound synthesis and external compound acquisition or merger projects [1–3]. The compound collections serve diverse screening paradigms, ranging from target-based to phenotypic-based screening, from biochemical to cell-based screening, and from focused hypothesis-based to diversity-based screening, opening a wide diversity of strategic choices for the future enhancement of the compound collection.

Herein, we review chemical, biological, and informatics elements for the development of strategies for compound library enhancement. The interdisciplinary nature of the library design activity is emphasized.

1.2 CHEMICAL SPACE FOR DRUG DISCOVERY

The chemical space is the ensemble of all possible molecules and comprises physically documented molecules available in the corporate and public databases as well as yet unknown, virtual molecules [10]. To delineate how many and which molecules populate unknown chemical space in total, Jean-Louis Reymond's group at the University of Berne performed a systematic computational enumeration and assembled the so-called chemical universe database—Figure 1.1 [10]. GDB-11 lists 26.4 million molecules of up to 11 atoms of C, N, O, and F, GDB-13 lists 977 million molecules up to 13 atoms of C, N, O, Cl, and S, and GDB-17 lists 166 billion molecules up to 17 atoms of C, N, O, S, and halogens [13]. The number of molecules enumerated in GDB increases exponentially with the number of atoms such that the database will become impracticably large as molecular size increases. For instance, extrapolation from the numbers in GDB-17 suggests that there would be approximately 10^{24} molecules up to 30 nonhydrogen atoms—typically, drug-sized molecules include up to 35 nonhydrogen atoms with molecular weight (MW) < 500 Da.

Within a drug discovery context, these astronomic numbers have to be placed in relation to the number of physically available chemicals and the actual number of around 1200 approved drugs satisfying stringent efficacy and safety criteria [14]. The Elsevier Medicinal Chemistry and Chemical Abstracts Service (CAS) Registry databases, which are up-to-date representatives of molecules described in the chemical literature, list, respectively, 5.5 and 74 million compounds [15, 16]. The eMolecules and ChemNavigator iResearch libraries, which are industry references for off-the-shelf compound acquisition, list, respectively, five and six million unique commercially available compounds [17, 18]. The screening collections of the major pharmaceutical companies include typically one to two million proprietary and nonproprietary compounds [7]. Given the practically infinite possibilities,

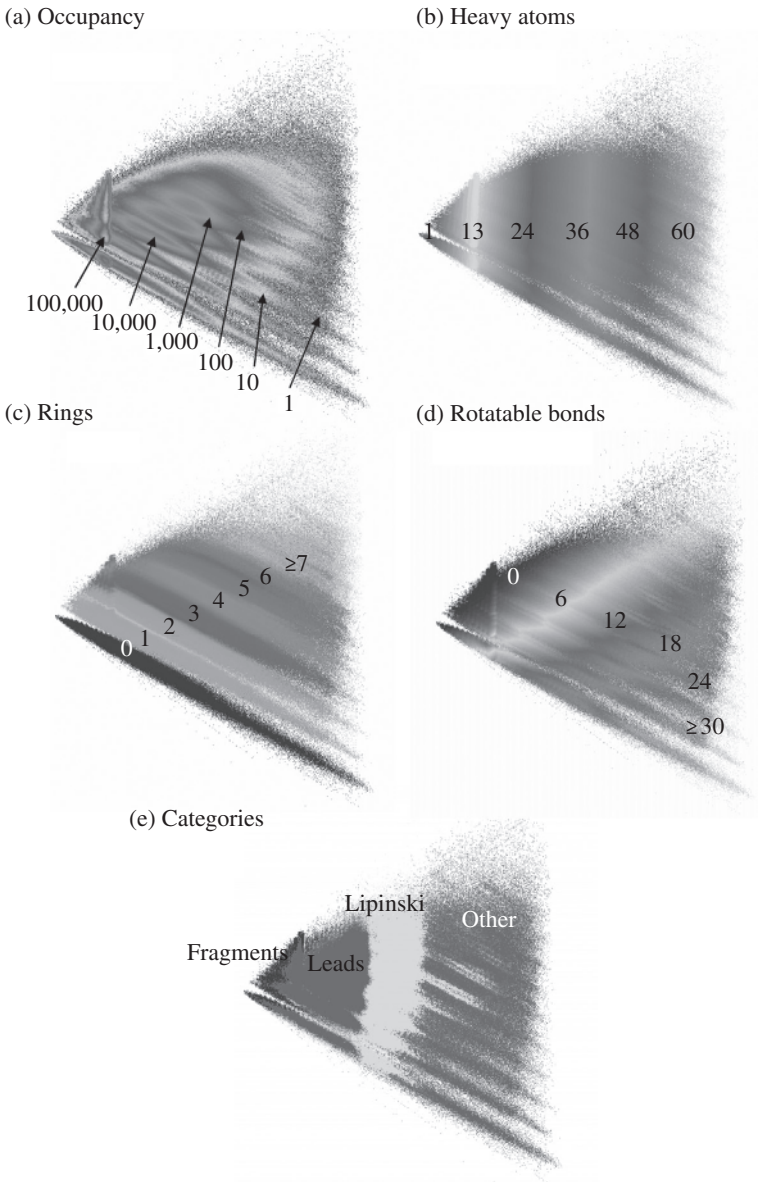

FIGURE 1.1 Example of visualization of chemical space via principal component analysis (PCA) [10–12]. Color-coded molecular quantum number (MQN) maps of the chemical space of PubChem compounds up to 60 heavy atoms and a subset of GDB-13 compounds in the (PC1, PC2) plane (total: 66,647,914 molecules). (a) Occupancy map color coded by the number of molecules per pixel. (b–d) Descriptor value maps color coded by the average descriptor value in each pixel. Saturation to gray is used to show standard deviation. (e) Category map for blue, fragments (rule of 3 (*vide infra*), 32.5 million compounds); green, lead-like (Teague's NOT rule of 3 (*vide infra*), 2.7 million compounds) (note: in total 12.2 million structures follow Teague's lead-likeness criteria); and cyan, Lipinski (rule of 5 (*vide infra*) NOT leads or rule of 3, 31.4 million compounds); and red, not matching any rule (1.6 million compounds). Color coding according to the majority category in each pixel except for leads (green), which were given priority to make them visible. Reprinted with permission from Ref. 10. © 2014, Pan Stanford Publishing. (*See insert for color representation of the figure.*)

the optimal size of a screening collection is frequently debated [19]. One estimate of the theoretically optimal size of a screening collection could be based on the size of the finite number of protein domains existing in the protein universe [5]. This number was recently estimated to be 1,500 domains and would translate to 15,000,000 compounds if one would design 10–20 chemotypes each of 500–1000 compounds to target each domain. A similar number can be reached if one would design 10–20 chemotypes each of 500–1000 for the estimated 600–1500 disease relevant druggable protein targets [20].

Tools to visualize, navigate, and select within the chemical space are essential chemoinformatic objectives for the design of the screening collection [21–23]. For every newly added compound, novelty needs to be checked at the individual compound and scaffold level. There are a number of commercial and proprietary informatics solutions that allow to store and search by chemical substructure and similarity chemical spaces in a robust and interactive fast manner. In 2001, Oprea and Gottfries pioneered the chemical global positioning system (ChemGPS) method to visualize chemical space [24]. The ChemGPS drug space map coordinates are *t*-scores extracted via PCA from 72 descriptors that evaluate a total set of 423 reference structures. Global ChemGPS scores describe well the latent structures extracted with PCA for a set of 8599 monocarboxylates, a set of 45 heteroaromatic compounds, and for 87 alpha-amino acids. ChemGPS positions novel structures in drug space via PCA-score prediction, providing a unique mapping and prediction device for the drug-like chemical space. ChemGPS scores are comparable across a large number of chemicals and do not change as new structures are predicted, making this tool a well-suited reference system for comparing multiple libraries and for keeping track of previously explored regions of the chemical space. The method was later on expanded to the chemical space for natural products and resulted in the ChemGPS-NP visualization and prediction system, which is publicly available on the web ChemGPS-NP(Web) (http://chemgps.bmc.uu.se) [25, 26]. ChemGPS-NP(Web) can assist in compound selection and prioritization, property description and interpretation, cluster analysis and neighborhood mapping, as well as comparison and characterization of large compound data sets. Schuffenhauer et al. introduced scaffold tree to analyze the scaffold diversity of natural products [27]. The method is a hierarchical classification of chemical scaffolds that form the leaf nodes in the hierarchy trees. By an iterative removal of rings, scaffolds forming the higher levels in the hierarchy tree are obtained. Prioritization rules ensure that less characteristic, peripheral rings are removed first. All scaffolds in the hierarchy tree are well-defined chemical entities making the classification chemically intuitive. The scaffold tree classification procedure handles robustly synthetic structures and natural products. In the design of new screening collections, the scaffold tree method is invaluable. Integrated with a chemically aware visualization tool like Tibco Spotfire, it allows the immediate assessment of the abundance of a given chemical scaffold within the existing collection and the candidate collection to integrate [28]. Within a collaboration between the Max Planck Institute for Molecular Physiology and Novartis, the method was integrated in a structural classification of natural products (SCONP) to chart the known chemical space explored by nature [29]. SCONP arranges the scaffolds of the natural products in a treelike fashion and provides a viable analysis- and hypothesis-generating tool for the design of natural product-derived compound collections. The Waldmann group developed the method further into Scaffold Hunter, an interactive computer-based tool for navigation in chemical space that fosters intuitive recognition of complex structural relationships associated with bioactivity [30, 31]. The program reads

compound structures and bioactivity data, generates compound scaffolds, correlates them in a hierarchical treelike arrangement, and annotates them with bioactivity.

In a need to enable navigation and selection with chemical space, researchers at Janssen presented library enhancement through the wisdom of crowds [32]. Compounds of interest are clustered together with the in-house collection using a fingerprint-based clustering algorithm that emphasizes common substructures and works with large data sets. Clusters populated exclusively by external compounds are identified as "diversity holes," and representative members of these clusters are presented to the global corporate medicinal chemistry community, who are asked to specify which ones they like, dislike, or are indifferent by using a simple point-and-click interface. The resulting votes are then used to rank the clusters from the most to the least desirable and to prioritize which ones should be targeted for acquisition.

Hypothesis-based selection in chemical space is supported by different types of virtual screening technologies depending on the size of the considered physical or virtual compound libraries. ChemNavigator offers, for instance, a comprehensive set of virtual screening services called 3DPL™ to select from their iResearch Library [18]. Researchers at Boehringer Ingelheim run virtual screening in a huge collection of virtual combinatorial libraries that led recently to the identification of two new structural classes of GPR119 agonists [33, 34]. Their virtual library called Boehringer Ingelheim Comprehensive Library of Accessible Innovative Molecules (BICLAIM) is based on combinatorial reactions and stored in a feature trees (FTrees) fragment space. The virtual chemical space contains about 1,600 scaffolds and 30,000 reagents encoding about 5×10^{11} theoretically chemically accessible molecules. The chemical universe database GDB-17 of 166.4 billion molecules can be virtually screened using a hashed fingerprint derived from the 42 integer MQN molecular descriptors [12]. An MQN-searchable 50 million subset of GDB-17 is publicly available at http://www.gdb.unibe.ch.

1.3 MOLECULAR PROPERTIES FOR DRUG DISCOVERY

Given that the size of the chemical space is virtually infinite, the art of library design lies in parts in the selection of the appropriate molecular property spaces. Medicinal computational chemists developed over the last decade a number of statistical analyses and ADMET models that are easily applicable upfront compound synthesis and are intended to reduce attrition at various stages [35–39]. The simplest models include substructure filters for potentially problematic chemical functionalities. The rapid elimination of swill (REOS) filters published by Vertex flag false positives in screening due to assay interference and reactivity or compounds with poor ADMET properties [40]. The pan-assay interference compounds (PAINS) filters identify frequent hitter in HTS [41, 42]. The analysis of Thorne et al. on typical screening technology-related assay artifacts provides a further guide to eliminate undesirable compound classes [43]. Among the molecular properties that are essential to small molecules are the water solubility and membrane permeability that form the basis of the two-dimensional biopharmaceutics classification system for drug developability [44]. The two properties are dependent in the sense that for specific oral dosing regime, a minimum equilibrium solubility level is required given the compound permeability class. They are not only important for late drug developability but also for early drug discovery. A compound has to be sufficiently soluble to enable

a dose–response-dependent readout. In addition, the compound has to have the appropriate permeability properties to reach its site of action within a cell or tissue. It is thus logical that a number of models focus on these properties.

Besides cheminformatics software like ACD/Labs Percepta Platform [45], Schrodinger's QikProp [46], or Simulation Plus' ADMET Predictor [47], which are based on advanced quantitative structure–property relationship (QSPR) modeling methods, there are a number of simpler heuristic-based models that have the advantage of being easily interpretable by the medicinal chemist. Chris Lipinski's pioneering work on the rule of 5, for instance, is derived from a quest for experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings [48]. In the discovery setting, the rule of 5 predicts that poor absorption or permeation is more likely when there are more than 5 H-bond donors and 10 H-bond acceptors, the MW is greater than 500 Da, and the calculated LogP (CLogP) is greater than 5.

As the rule of 5 is, for instance, not able to statistically discriminate drugs from nondrugs, Lipinksi's thinking was initially highly controverted in an era where combinatorial chemistry aimed to deliver compounds that were easily synthesized and purified with having higher MW and LogP. Later, Lipinski's thinking influenced an entire school of thought around small molecules for drug discovery [49, 50].

The analysis provided by Wenlock showed that MW and ClogP distributions move through the phases of drug development and that the property distributions approach those of marketed oral drugs—Figure 1.2 [51]. Given that the early combinatorial chemistry was not successful, it is not surprising that over time a quantity to quality shift developed. The observation that medicinal chemists focus on potency during early lead optimization by making compounds even bigger and lipophilic led Oprea's group at AstraZeneca to introduce the concept of lead-likeness [52]. Larger size and lipophilicity drive also compound promiscuity and potential off-target effects [53–55]. Increasing MW and CLogP is an easy way to reach the common nanomolar potency. This tendency is, however, counterproductive for ADMET and moves the properties further away from historical drug space.

Over the time, Oprea refined his analysis of lead–drug pairs and recommended that lead-likeness libraries should have the following as characteristics: $MW < 460$ Da, $-4 < CLogP < 4.2$, water solubility $LogS > -5$, number of rotatable bonds less than 10, number of rings less than 4, number of H-bond donors less than 5, and number of H-bond acceptors less than 9 [56]. These differences compared to drugs are thus subtle, and as concluded by Proudfoot, successful and timely drug discovery campaigns require high-quality lead structures, and these lead structures may need to be much more drug-like than is commonly accepted [57]. Similar conclusions were derived by Oprea when analyzing more recently chemical probes and leads [58]. The field of fragment-based drug discovery takes the concept of having small molecule starting points further. Following an analysis done at Astex, fragment libraries are often designed by applying a rule of 3 in which MW is less than 300 Da, the number of hydrogen-bond donors is less than or equal to 3, the number of hydrogen-bond acceptors is less than or equal to 3, and ClogP is less than or equal to 3. In addition, the analysis suggested that the number of rotatable bonds (NROT) (≤ 3) and topological polar surface area (tPSA) (≤ 60 Å^2) might also be useful criteria for fragment selection [59]. The Astex scientists argue that ADMET properties can be better controlled during optimization when starting with a fragment compound compared to a larger compound.

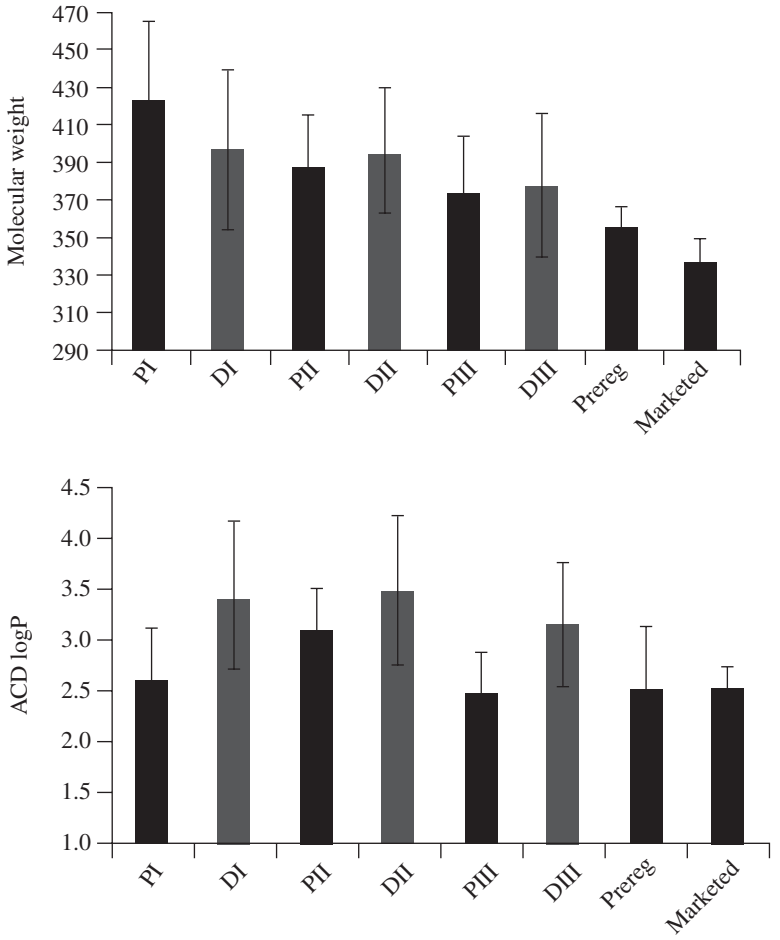

FIGURE 1.2 Analysis by Wenlock et al. of the evolution of molecular property distributions with progressing through development stages [51]. Phase I (PI); discontinued phase I (DI); phase II (PII); discontinued phase II (DII); phase III (PIII); discontinued phase III; preregistration (Prereg); marketed oral drugs. The analysis shows that the mean molecular weight of orally administered drugs in development decreases on passing through each of the different clinical phases and gradually converges toward the mean molecular weight of marketed oral drugs. It is also clear that the most lipophilic compounds are being discontinued from development. Reprinted with permission from Ref. 51. © 2003, American Chemical Society.

The observation that smaller and less lipophilic starting points are better prompted researchers at GSK to propose the concept of lead-oriented synthesis (LOS), which aims for compounds with LogP values in the range −1 to 3 and MW in the range of 200–350 Da [60]. The authors emphasize the need to access to novel synthesis methodologies given that the current array chemistry has an unintentional bias toward the synthesis of less drug-like molecules.

Further analyses of computed and experimental physicochemical properties of drug compounds lead to the conclusion that the property spaces depend on the target class and

therapeutic indication [35, 61]. For a given therapeutic indication, the site of *in vivo* action might require due to specific barriers the active compound to occupy quite specific property spaces like those illustrated, for instance, by the central nervous system (CNS) property space or the antibacterial property spaces [62, 63].

Gleeson provided a set of simple, consistent structure–property rules of thumb determined from an analysis of a number of key ADMET assays run within GSK: solubility, permeability, bioavailability, volume of distribution, plasma protein binding, CNS penetration, brain tissue binding, P-gp efflux, hERG inhibition, and cytochrome P450 1A2/2C9/2C19/2D6/3A4 inhibition [64, 65]. The rules have again been formulated using molecular properties that chemists intuitively know how to alter in a molecule, namely, MW, LogP, and ionization state. This study reemphasizes again the need to focus on a lower MW and LogP area of physicochemical property space to obtain improved ADMET parameters.

To assess the use of this knowledge in reducing the likelihood of compound-related attrition, the molecular properties of compounds acting at specific drug targets described in patents from leading pharmaceutical companies during the 2000–2010 period were analyzed by Leeson and St-Gallay [66]. The authors conclude that a substantial sector of the pharmaceutical industry has not modified its drug design practices and is according to them producing compounds with suboptimal physicochemical profiles.

The Golden Triangle is a visualization tool developed at Pfizer from *in vitro* permeability, *in vitro* clearance, and computational data designed to aid medicinal chemists in achieving metabolically stable, permeable, and potent drug candidates [67]. Classifying compounds as permeable and stable and plotting MW versus octanol–buffer (pH 7.4) distribution coefficients (LogD) or estimated octanol–buffer (pH 7.4) distribution coefficients (eLogD) reveal useful trends. The Golden Triangle is defined by an apex of MW 450 Da and a base of MW 200 Da, and a logD range of −2 to +5. 25% of the compounds in Golden Triangle has acceptable Caco-2 permeability and microsomal stability versus only 3% for compounds outside the Golden Triangle.

The analysis by Hill and Young of the relationship between hydrophobicity and approximately 100 k measured kinetic solubility values showed that better solubility predictions are obtained by taking ACD clogD(pH 7.4) values together with the number of aromatic rings in a given molecule—Figure 1.3 [68]. The Solubility Forecast Index (SFI = clogD(pH 7.4) + #Ar) was proposed as a simple, yet effective, guide to predicting solubility.

Pfizer provided with the 3/75 rule an example of how physicochemical drug properties are associated with *in vivo* toxicity [69]. From a data set consisting of animal *in vivo* toleration studies on 245 preclinical Pfizer compounds across a broad swath of chemical space, an increased likelihood of toxic events across a wide range of types of toxicity is observed for less polar, more lipophilic compounds. Compounds with CLogP < 3 and a tPSA > 75 Å^2 show a clear correlation of lower odds of promiscuity and toxicity.

Strict property-based assessment of drug-likeness has been recently criticized as being too blunt an instrument that affords only a yes–no answer. The quantitative estimate of drug-likeness (QED) has been introduced to overcome such limitations by characterizing how well physicochemical properties of a candidate compound match the property distributions of marketed oral drugs [70]. Ritchie and MacDonald showed that drugs with high QED scores exhibit higher absorption and bioavailability, are administered at lower doses, and have fewer drug–drug interaction warnings, P-glycoprotein interactions, and absorption issues due to a food effect. By contrast, the high-scoring drugs exhibit similar behavior to low-scoring drugs with respect to free fraction in plasma, extent of gut-wall

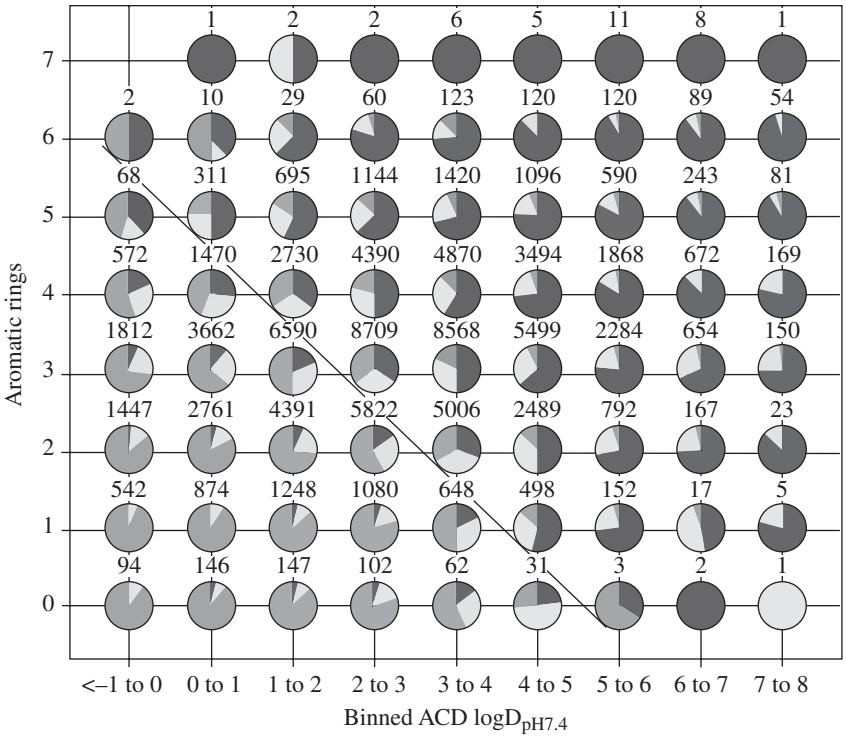

FIGURE 1.3 Trelis plot of Hill and Young showing the distribution of water solubility as a function of computed LogD and # of aromatic rings [68]. Solubility classes—green, high (>200 μM); yellow, medium (30–200 μM); and red, low (<30 μM). The number above the pie charts corresponds to the number of compounds analyzed for each bin. Reprinted with permission from Ref. 68. © 2010, Elsevier. (*See insert for color representation of the figure.*)

metabolism, first-pass hepatic extraction, elimination half-life, clearance, volume of distribution, and frequency of dosing [71].

1.4 MAJOR COMPOUND CLASSES

Natural products, known bioactives, peptides, heterocycles, and DOS libraries, constitute the prevalent compound classes represented in screening collections and are reviewed in this section [4]. For obvious reasons, natural principles play a predominant role in the history of drug discovery. Diverse classes of natural products including carbohydrates, steroids, fatty acids, polyketides, peptides, terpenoids, flavonoids, alkaloids, and many other products were isolated initially from herbs and later from various micro and higher organisms for structure and activity characterization [72, 73]. Natural products are a major source of innovative tool compounds for the elucidation of signaling pathways and new medicines for most indications, such as lipid disorders, cancer, infectious diseases, and immunomodulation. Between 1981 and 2002, 5% of the around 870 new chemical entities approved by the US Food and Drug Administration (FDA) were natural products, and another 23% were molecules derived from natural products [74].

Natural products offer a wealth of new structures far beyond the classical repertoire of synthetic compounds. The current most comprehensive summary on the chemical and biological information of around 230,000 isolated natural products is provided in the Chapman & Hall Dictionary of Natural Products (DNP) database [75].

A number of studies have investigated the structural characteristics of natural products compared to synthetic organic compounds [76–79]. Natural products often contain a greater proportion of oxygen than nitrogen heteroatoms. Typically, the natural products have a higher number of stereocenters, a higher density of functionalization and pharmacophore sites, a higher number of rings, and more skeletal diversity. Natural products exemplify macro- and polycyclic scaffolds beyond the imagination of the classical synthetic medicinal chemist. Conversely, examples also exist of very simple natural product structures with biological activity. The structural repertoire can be extended by genomic approaches to natural products. Approaches based on genome sequence information and subsequent annotation of biosynthetic pathways are emerging technologies [80]. Tang and Khosla described the potential of combinatorial biosynthesis of "unnatural" natural products via the genetic engineering of the biosynthetic pathways of polyketides [81].

Natural products were excluded from Lipinski's rule of 5 observation. Despite the fact that the distribution profiles of natural products are indeed broader compared to synthetic compounds, their fraction with two or more rule of 5 violations is equal to that of synthetic drugs. One interpretation of this finding might be that evolutionary optimization has coded in addition to these essential properties other biocharacteristics that still need to be deciphered. An analysis by Ganesan showed that those natural products that violated the rule of 5 have higher MW, more rotatable bonds, and more stereocenters; however, they remain largely compliant in terms of logP and H-bond donors, highlighting the importance of these two metrics in predicting bioavailability [82]. Natural products have learned to maintain low hydrophobicity and intermolecular H-bond donating potential when it needs to make biologically active compounds with high MW and large numbers of rotatable bonds. In addition, natural products are more likely than purely synthetic compounds to resemble biosynthetic intermediates or endogenous metabolites and hence take advantage of active transport mechanisms. Conversely, a number of marketed natural product-based drugs are not orally available, but uniquely address a number of therapeutic applications.

One key dilemma for natural products drug discovery is that although the primary HTS hit rates in the micromolar affinity range are 5–10 times higher than the hit rates for synthetic compounds, the take-up rate of the compounds by chemists for follow-up lead optimization is significantly lower [1]. This finding is most probably due to the higher structural complexity and challenges related to the chemical structure elucidation and synthesis. A promising trend to broaden the scope of natural products is given by making small combinatorial libraries from natural products and natural product-like scaffolds. A systematic extension of such libraries based on protein structure similarity clustering (PSSC) was proposed by the Waldmann group [83]. The principles of this approach consider the domain organization and conservation of proteins and the corresponding needs for conservatisms of the architectures and interaction modes of their ligands.

Primary metabolites and marketed drugs form additional sets of biologically relevant and validated compounds that form an essential component of a comprehensive screening collection [4].

Primary metabolites, which are key intermediates of cellular metabolisms and which interact with key enzymes and cellular regulatory receptor systems, are systematically included in deorphanization libraries of orphan targets. The CheBI database organizes the relevant chemical and biological information [84]. Hits from such libraries allow the elucidation of the functional relevance of a new potential target protein.

Marketed drugs and derivative libraries are an important and invaluable compound source and provide the basis for the selective optimization of side activities (SOSA) approach [85]. The SOSA approach consists of testing old drugs on new pharmacological targets. The aim is to subject to pharmacological screening a limited number of drug molecules that are structurally and therapeutically very diverse and that have known safety and bioavailability in humans, thereby potentially shortening the time and the cost needed for hit optimization. Since bioavailability and toxicity studies have already been performed for those drugs and since they have proven their usefulness in human therapy, all hits are *per definition* drug-like. In the second stage, the hits are optimized by means of traditional, parallel, or combinatorial chemistry in order to increase the affinity for the new target and decrease the affinity for the other targets. The objective is to prepare analogues of the hit molecule in order to transform the observed "side activity" into the main effect and to strongly reduce or abolish the initial pharmacological activity.

Peptide–protein molecular interactions are the most ubiquitous mode for controlling and modulating cellular function, intercellular communication, and signal transduction pathways [86]. Peptides are key components of chemogenomics discovery libraries and are especially useful for the characterization of orphan targets. A number of successful deorphanizations, especially in the GPCR field, are based on peptides, resulting in new drug discovery projects. New peptides for such libraries are discovered using HPLC fractionations of tissue extracts together with random or designed peptide libraries based on the bioinformatics analysis of putatively secreted peptides and protein hormones defined in the genome [87].

Limiting factors of peptide-based drugs are directed by the number of amide bonds that determine properties like a high tPSA, a low membrane permeability, and a potentially high proteolytic degradation, resulting in quite poor ADME properties [88]. Mainly because of these reasons, robust strategies for the design of peptide mimetics have been successfully developed [89]. Oral delivery of therapeutic peptides is still a challenge. A number of factors including high proteolytic activity and low pH conditions of the gastrointestinal tract act as major barriers in the successful delivery of intact peptide to the targeted site. Low permeability of peptides across the intestinal barrier is also a factor adding to the low bioavailability. Nanocarrier-based delivery presents an appropriate choice of drug carriers owing to their property to protect proteins from degradation by the low pH conditions in the stomach or by the proteolytic enzymes in the gastrointestinal tract [90]. Recently, cell-penetrating peptides (CPPs) such as HIV-1 Tat, penetratin, and oligoarginine are considered as a useful tool for the intracellular delivery of therapeutic macromolecules [91]. CPPs are likely to become powerful tools for overcoming the low permeability of therapeutic peptides through the intestinal membrane, the major barrier to their oral delivery. Peptide-derived (natural and nonnatural amino acids) macrocycles are a relatively new trend in drug discovery [92–95]. Macrocycles are conformationally constrained molecules that can fix the bioactive conformation. Macrocycles come in different flavors and can't be lumped into one class because they cover a wide range of different structural classes and different MW. A stapled peptide is very different from a large cyclic

peptide, which is very different from a synthetic macrocycle, which again is very different from a natural product macrocycle [95]. Heterocycles form historically the most prevalent class of drug molecules. They cover a diverse set of ring systems with various types of heteroatoms and have been extensively patented. The quest for new rings was systematically investigated *in silico*. Researchers at UCB generated a complete list of 24,847 ring systems called virtual exploratory heterocyclic library (VEHICLe) [96]. Searching literature and compound databases, using this list as substructure queries, identified only 1701 as synthesized. Using a carefully validated machine learning approach, it was possible to estimate that the number of unpublished, but synthetically tractable, VEHICLe rings could be over 3000. This analysis also shows that the rate of publication of novel examples to be as low as 5–10 per year. Corroboratively, Ertl and coworkers at Novartis showed that bioactive molecules only contain a relatively limited number of unique ring types [97]. To identify those ring properties and structural characteristics that are necessary for biological activity, a large virtual library of nearly 600,000 heteroaromatic scaffolds was created and characterized by calculated properties. Using a self-organizing neural network, the scaffolds were clustered and showed that bioactivity is very sparsely distributed within the scaffold property and structural space, forming only several relatively small, well-defined "bioactivity islands." Such analyses provide a fresh stimulus to creative organic chemists by highlighting a small set of apparently simple ring systems that are predicted to be tractable but are unconquered. A recent trend in heterocyclic chemistry is to increase the ratio of sp^3-hybridized carbon atoms (Fsp^3) yielding more saturated ring systems. Lovering et al. showed that both complexity (as measured by Fsp^3) and the presence of chiral centers correlate with success as compounds transition from discovery, through clinical testing, to drugs. In an attempt to explain these observations, it was demonstrated that saturation correlates with solubility [98]. Within the same perspective, Ishikawa and Hashimoto provided examples how the breaking of molecular symmetry and planarity is effective to improve solubility despite increasing hydrophobicity [99]. The impact of carboaromatic, heteroaromatic, carboaliphatic, and heteroaliphatic ring counts and fused aromatic ring count on several developability measures (solubility, lipophilicity, protein binding, P450 inhibition, and hERG binding) was recently reviewed [100]. Increasing ring counts have detrimental effects on developability in the order carboaromatics $\gg$ heteroaromatics $>$ carboaliphatics $>$ heteroaliphatics, with heteroaliphatics exerting a beneficial effect in many cases. Increasing aromatic ring count exerts effects on several developability parameters that are lipophilicity and size independent, and fused aromatic systems have a beneficial effect relative to their nonfused counterparts.

The metabolism of heterocycles can result in challenges for the optimization of pharmacokinetics/pharmacodynamics (PK/PD) profiles of the compounds. Recently, systematic mitigating strategies for heterocycle metabolism have been established by St. Jean and Fotsch allowing the selection of improved building blocks for library design [101].

Diversity-oriented synthesis (DOS), as opposed to the traditional target-oriented synthesis (TOS) chemistry approach, was introduced by the Schreiber group for forward chemical genetic screening in order to mimic the structural complexity and the skeletal and stereochemical diversity of natural products [102]. Conversely to a convergent synthesis strategy resulting from the logic of retrosynthetic analysis of the target molecules, DOS, in the ideal state, allows the application of a diverse set of reagents and structural transformations on each synthesis intermediate; this results in diverging synthesis pathways that

create a broad diversity of target molecules with different scaffolds. DOS compounds clearly share a number of characteristics with natural products including most notably the scaffold diversity and stereochemical complexity. The question remains, however, whether these products of pure chemist imagination capture the evolutionary advantages of natural products and natural product-based compounds. The DOS planning strategy allows, by enumeration over a larger number of steps, the genesis of truly novel structures that by itself is an innovative concept. In practice, DOS combinatorial libraries focus to leverage information about existing biologically active molecules in order to address the biologically relevant regions of chemical space. DOS libraries are not directed toward a single biological target and aim to provide diverse discovery libraries. DOS has increased the need for exceptionally efficient, stereoselective, and chemoselective reactions, including multicomponent reactions (MCR) that can be applied to a broad range of substrates.

A number of recent success stories prove that DOS compounds provide invaluable tools for target validation [103]. The validation of the ADMET and *in vivo* properties of these compounds and their value as therapeutics remains however to be proven. Comparable to natural products, as result of the structural complexity, a key challenge is expected in the lead optimization phase and for the industrial chemical development of the final compounds.

In a comparative analysis, Clemons et al. found that compounds from different sources (commercial, academic DOS, natural products) have different protein-binding behaviors against each of 100 diverse (sequence-unrelated) proteins [104]. These behaviors correlate with general trends in stereochemical and shape descriptors for these compound collections. Increasing the content of sp^3-hybridized and stereogenic atoms relative to compounds from commercial sources, which comprise the majority of current screening collections, improved binding selectivity and frequency.

1.5 CHEMICAL DESIGN APPROACHES TO EXPAND BIOACTIVE CHEMICAL SPACE

Systematic hypothesis-based expansion of the chemical space to reach a maximum of biological binding sites appears possible when conserved molecular recognition principles are the founding hypothesis for the design of the compounds. Such chemogenomics principles, including approaches focusing on target families, privileged scaffolds, protein secondary structure mimetics, cofactor mimetics, and BIOS libraries, were recently summarized by us [5]. To be broadly successful, these approaches are complemented by diversity-based principles like DOS, DNA-encoded libraries (DELs), and fragment-based approaches (FBS).

More than 50% of the marketed drugs target only four key gene families, including the rhodopsin-like GPCRs, nuclear receptors, ligand-gated ion channels, and voltage-gated ion channels [61, 105]. Historically, drug discovery has thus been focusing on a few "druggable" target families. The key design principles, focusing on similarities or differences in the physicochemistry of equivalent residues lining the binding site, can also rationalize the polypharmacology of many drugs. Because protein family-targeted library design requires extensive ligand-based or structure-based knowledge, it is not surprising that current design of chemical libraries directed to target classes focuses mainly on GPCRs, kinases, nuclear receptors, and more recently ion channels and epigenetic

targets. Today, protein family-targeted libraries with a large diversity of chemotypes are specifically designed toward subfamilies with conserved molecular recognition [106]. Various strategies have been applied to design GPCR [107, 108] and ion channel libraries [109], mostly based on ligand information captured in the form of molecular descriptors, pharmacophores, and substructures extracted from active reference compounds. In addition to these methods, the large amount of structure-based information available from X-ray analysis of ligand-target complexes makes structure-based design approaches feasible in the protein kinase [110, 111], protease [112], and nuclear receptor [113] classes. Impressive progress was reached within the last 5 years on new epigenetic targets like histone methyltransferases and bromodomains, thanks to the initiative of the structural genomics consortium (SGC) for chemical probes and structural biology [114–116]. The typical size of a protein target-focused library is 100–1000 compounds if it is centered around one single chemotype; the library size can grow to 10,000–20,000 compounds if it is oriented around multiple chemotypes.

De novo design approaches for target families were recently reported [117]. Automated design of ligands to polypharmacological profiles was demonstrated by the Hopkins group by the evolution of an approved acetylcholinesterase inhibitor drug into brain-penetrable ligands with either specific polypharmacology or exquisite selectivity profiles for biogenic G-protein-coupled receptors [118]. Overall, 800 ligand-target predictions of prospectively designed ligands were tested experimentally, of which 75% were confirmed to be correct. The approach can be a useful source of leads when multitarget profiles are required to achieve either selectivity over other drug targets or a desired polypharmacology. Hartenfeller at Novartis published the libDOGS methods in extension to DOGS developed by the Schneider group at ETH [119, 120]. The principle is a reaction-driven de novo design method for automated design of target family-oriented libraries. Focusing on hypotheses (suggested molecules) that can easily be validated via straightforward syntheses expected to allow testing more hypotheses compared to scenarios where complex molecules require more synthetic effort.

The goal of a protein family-targeted library is not to target a specific target exclusively, but to address by different library members different members of the target subfamilies. This coverage of a targeted library could until recently only be addressed experimentally by the analysis of hit rates. Such analyses showed that some designed libraries hit not only the primary target family but also other a priori nonrelated target families [1].

Biology-oriented synthesis (BIOS) was introduced by the Waldmann group [121–123]. BIOS centers on the generation of small compound libraries based on scaffolds of proven biological relevance. Library generation is focused on compound classes from the "biological relevant space," that is, the natural products and drugs. BIOS unifies the aforementioned SCONP and PSSC concepts (*vide supra*) that, respectively, allow navigation in the chemical and biological spaces.

The use of privileged substructures or molecular master keys is an accepted concept in medicinal chemistry. The privileged structure approach emphasizes molecular scaffolds or selected substructures that are able to provide high-affinity ligands (agonist or antagonists) for diverse receptors and originates from work at Merck Research Laboratories on the design of benzodiazepine-based CCK (Cholecystokinin) antagonists, where the previously known κ-opioid tifluadom was identified as a lead structure [124]. A number of recent literature reviews provide impressive reference repertoires of empirically derived privileged structures for various target families [125–127].

Protein–protein molecular interactions are the most ubiquitous mode for controlling and modulating cellular function, intercellular communication, and signal transduction pathways. Peptide and protein mimetic libraries including β-turn and α-helix mimetics are recognized of central importance in chemogenomics. The design of drug-like active β-turn mimetics based on organic drug-like scaffolds advanced to a quite routine methodology and a variety of approaches have been validated as recently summarized by Marshall et al. [128], including α,β-dehydroamino acids, proline and heterochiral *N*-methyl amino acids, or cyclic α-peptides. The work of Garland and Dean [129, 130], defining a set of triangular distance constraints that the substitution points of a scaffold have to satisfy in order to mimic the specific C_α atoms of the peptide template, provided a generalized frame for the design of novel β-turn mimetic scaffolds and was in combination with database searches successfully applied for the design of CCK, SST, and MC_4 antagonists [131].

Other protein–protein interactions (PPI) are via key α-helix motifs. The work of the groups of Hamilton [132] has established a solid foundation for the rational design of α-helix mimetics. 3.2′.2″-substituted terphenyl-derived motifs were among the first designed motifs and were shown to be able to mimic the side-chain positions i, $i+3$ or $i+4$, and $i+7$, which are on the same face of a α-helix. Besides these rationally designed helix mimetics, diversity-based high-throughput screening (HTS) and virtual screening have identified a number of scaffolds for a variety of targets that allow the correct spatial orientation of substituents for interaction with the protein target [133]. Combinatorial libraries around such scaffolds are an essential component of a chemogenomics discovery library. Morelli et al. analyzed the PPI chemical space together with current structural knowledge regarding both protein–protein and protein–inhibitor complexes summarized in the 2P2I database [134]. The statistical analysis of the 39 inhibitors present in $2P2I_{DB}$ enabled them to calculate the general characteristics of the PPI chemical space. Average values for the MW (547 ± 154 Da; thus, MW > 400 Da), ALogP (3.99 ± 2.37; thus, ALogP > 4), number of rings (4.44 ± 1.02; thus, #Rings > 4), and number of hydrogen-bond acceptors (6.62 ± 2.60; thus, #HBA > 4) define the generic profile of a PPI inhibitor compound that could be further derived into a more specific inhibitor. These chemical rule of 4 properties can be used to filter "in-house" databases and accelerate the process of hit identification by lowering both cost and time.

With the successful design of MDM2-P53 inhibitors, the Dömling group introduced the ANCHOR method whereby computational and MCR chemistry converge [135]. Their general workflow for the rapid generation of low-molecular-weight (ant)agonists of PPIs relies on a sequence of steps. First, the presence of a highly buried amino acid is identified. In the next step, a molecular anchor is generated by fragmentation of the candidate ligand. Virtual chemistry employing the anchor and based on MCR is then used for the enumeration of further candidate compounds that are docked into the binding site using the anchor as a constraint. Finally, chemical synthesis and biological screening of the most promising candidates are completed.

As successfully exemplified by the protein kinase ATP-binding site, cofactor-binding sites are evolutionary highly conserved binding sites that provide rich opportunities for enzyme-based drug discovery. The detailed comparative structural analysis of substrate- and cofactor-binding sites shows indeed that cofactor analogues open a very wide target window. Ji et al. examined around 2200 well-defined small molecule ligands and thousands of protein domains derived from a database of druggable binding sites and found that

cofactor molecules are the most prevalent bioligands in the protein fold universe and hence present a unique opportunity for chemogenomics systematization approaches [136].

Typical industrial screening libraries are of the size of approximately 1–2 Mio compounds, which is a drop in the ocean compared to the theoretical chemical space of drug-like molecules. FBS and DELs allow the access to a significantly larger chemical space compared to conventional screening libraries.

Considering the hypothetical case of a ligand comprising three fragments that bind to three subsites, if only 1000 fragments should be tested for each of the subsites, this would require the synthesis and testing of at least $1000 \times 1000 \times 1000 = 10^9$ molecules, without considering the number of possible linkers. In an FBS, it would only require the screening of 1000 compounds in this hypothetical case.

Key to the success of FBS, which can be performed by a variety of biophysical methods, is the quality of the FBS library, and different generations of designs have been introduced in the past [137–139]. More recently, the concepts of 3D fragments, fragments of known drugs, and fragments of natural products have been realized [140, 141]. Very important for the follow-up of the fragment hits is also the access to structural biology support providing high-resolution structures of the fragment-binding interactions and the three-dimensional building plan to expand the fragments into full-sized molecules [142, 143]. Typically, dedicated chemistry support is required, and a number of BioTech companies have focused on the FBS paradigm and delivered an impressive number of drug candidates [144, 145].

DELs provide access to 10^3–10^9 diverse compounds in a single test tube. The scientific approach was first described by Lerner and Brenner and consists of small molecules covalently attached to unique DNA sequences that serve as PCR amplifiable identification barcodes [146]. Selection offers large advantages over screening in terms of numbers, flexibility, convenience, and costs. A number of protocols have been designed [147]. The approaches enable the deconvolution of SAR information and families of actives are observed. One potential limitation is the restriction of chemical reactions to aqueous and DNA-compatible conditions. DELs have started to yield novel modulators of biological targets [148–150].

1.6 CONCLUSION

The discovery of new innovative small molecule drugs will continue to rely on the systematic exploration of new chemotypes. The enhancement of the corporate screening collection with high-quality chemical matter is a long-term continuous process of capital interest to every drug discovery organization. Quality library design is one of the most important scientific drug discovery competencies and is recognized as critical to the future productivity [151, 152]. Disruptive versus incremental innovation can be promoted by moving on purpose into new areas of the chemical space. As outlined in this article, there are a number of strategic options that will need to be combined in an appropriate manner. The optimal strategy for one organization might be very different from that of another organization. Even within one same organization, there might be different needs from the different disease areas, for instance, which will call for a mixed-strategy approach. A very important aspect is the time horizon on which the yield of the screening collection will be measured. Diverse libraries will require a longer readout time than

project-focused libraries. Compared to partial deck screening and screening deck turnover strategies, the full deck screening of a collection over a longer time period (~10 years) adds additional value in building a comprehensive experimental structure–activity relationship (SAR) matrix, which will help in making informed decisions relevant for many projects and also help to build the basis for future knowledge-based library design and virtual screening approaches [153–155].

ACKNOWLEDGMENTS

Drs. D. Berthelot, C. Buyck, R. Desjarlais, C. Martinez-Lamenca, T. Mirzadegan, V. Pande, P. Ten Holte, G. Tresadern, and H. van Vlijmen (all Janssen associates) are acknowledged for various support and discussions. Figures 1.1, 1.2, and 1.3 were, respectively, reproduced with permission from references [10, 51, 68].

REFERENCES

[1] 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, Roth HJ (2005). Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Curr. Top. Med. Chem.* **5**:397–411.

[2] Renner S, Popov M, Schuffenhauer A, Roth HJ, Breitenstein W, Marzinzik A, Lewis I, Krastel P, Nigsch F, Jenkins J, Jacoby E (2011). Recent trends and observations in the design of high-quality screening collections. *Future Med. Chem.* **3**:751–766.

[3] Djuric SW, Akritopoulou-Zanze I, Cox PB, Galasinski S (2010). Compound collection enhancement and paradigms for high-throughput screening-an update. *Ann. Rep. Med. Chem.* **45**:409–428.

[4] Jacoby E, Schuffenhauer A, Azzaoui K, Popov M, Dressler S, Glick M, Jenkins J, Davies J, Roggo S. Small molecules for chemogenomics-based drug discovery. In: Jacoby E, editor. *Chemogenomics—knowledge-based approaches to drug discovery*. London: Imperial College Press; 2006. p 1–38.

[5] Jacoby E, Mozzarelli A (2009). Chemogenomic strategies to expand the bioactive chemical space. *Curr. Med. Chem.* **16**:4374–4381.

[6] Drewry DH, Macarron R (2010). Enhancements of screening collections to address areas of unmet medical need: an industry perspective. *Curr. Opin. Chem. Biol.* **14**:289–298.

[7] Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DV, Hertzberg RP, Janzen WP, Paslay JW, Schopfer U, Sittampalam GS (2011). Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **10**:188–195.

[8] Wigglesworth M, Wood T, editors. *Management of chemical and biological samples for screening applications*. Weinheim: Wiley-VCH Verlag GmbH & co. KGaA; 2012.

[9] Mullard A (2013). European lead factory opens for business. *Nat. Rev. Drug Discov.* **12**:173–175.

[10] Reymond JL, Ruddigkeit L, Awale M. Considerations on the drug-like chemical space. In: Jacoby E, editor. *Computational chemogenomics*. Singapore: Pan Stanford; 2013.

[11] Deursen RV, Blum LC, Reymond JL (2010). A searchable map of PubChem. *J. Chem. Inf. Model.* **50**:1924–1934.

[12] Ruddigkeit L, Blum LC, Reymond JL (2013). Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* **53**:56–65.

[13] Ruddigkeit L, van Deursen R, Blum LC, Reymond JL (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**:2864–2875.

[14] Overington JP, Al-Lazikani B, Hopkins AL (2006). How many drug targets are there? *Nat. Rev. Drug Discov.* **5**:993–996.

[15] http://www.elsevier.com/online-tools/reaxys/reaxys-medicinal-chemistry

[16] http://www.cas.org/content/chemical-substances

[17] http://www.emolecules.com/

[18] http://www.chemnavigator.com/

[19] Lipkin MJ, Stevens AP, Livingstone DJ, Harris CJ (2008). How large does a compound screening collection need to be? *Comb. Chem. High Throughput Screen.* **11**:482–493.

[20] Hopkins AL, Groom CR (2002). The druggable genome. *Nat. Rev. Drug Discov.* **1**:727–730.

[21] Lajiness M, Watson I (2008). Dissimilarity-based approaches to compound acquisition. *Curr. Opin. Chem. Biol.* **12**:366–371.

[22] Medina-Franco JL. Cheminformatic characterization of the chemical space and molecular diversity of compound libraries. In: Trabocchi A, editor. *Diversity-oriented synthesis: basics and applications in organic synthesis, drug discovery, and chemical biology*. Hoboken: John Wiley & Sons, Inc.; 2013. p 325–352.

[23] Gibbs A, Agrafiotis DK. Chemical diversity: definition and quantification. In: Bartlett PA, Entzeroth M, editors. *Exploiting chemical diversity for drug discovery*. Cambridge: Royal Society of Chemistry; 2006. p 137–162.

[24] Oprea TI, Gottfries J (2001). Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **3**:157–166.

[25] Larsson J, Gottfries J, Muresan S, Backlund A (2007). ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J. Nat. Prod.* **70**:789–794.

[26] Rosén J, Gottfries J, Muresan S, Backlund A, Oprea TI (2009). Novel chemical space exploration via natural products. *J. Med. Chem.* **52**:1953–1962.

[27] Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H (2007). The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **47**:47–58.

[28] http://spotfire.tibco.com/

[29] Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H (2005). Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U. S. A.* **102**:17272–17277.

[30] Wetzel S, Klein K, Renner S, Rauh D, Oprea TI, Mutzel P, Waldmann H (2009). Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.* **5**:581–583.

[31] Lachance H, Wetzel S, Kumar K, Waldmann H (2012). Charting, navigating, and populating natural product chemical space for drug discovery. *J. Med. Chem.* **55**:5989–6001.

[32] Hack MD, Rassokhin DN, Buyck C, Seierstad M, Skalkin A, ten Holte P, Jones TK, Mirzadegan T, Agrafiotis DK (2011). Library enhancement through the wisdom of crowds. *J. Chem. Inf. Model.* **51**:3275–3286.

[33] Lessel U, Wellenzohn B, Lilienthal M, Claussen H (2009). Searching fragment spaces with feature trees. *J. Chem. Inf. Model.* **49**:270–279.

[34] Wellenzohn B, Lessel U, Beller A, Isambert T, Hoenke C, Nosse B (2012). Identification of new potent GPR119 agonists by combining virtual screening and combinatorial chemistry. *J. Med. Chem.* **55**:11031–11041.

[35] Meanwell NA (2011). Improving drug candidates by design: a focus on physicochemical properties as a means of improving compound disposition and safety. *Chem. Res. Toxicol.* **24**:1420–1456.

[36] Gleeson MP, Montanari D (2012). Strategies for the generation, validation and application of in silico ADMET models in lead generation and optimization. *Expert Opin. Drug Metab. Toxicol.* **8**:1435–1446.

[37] Cox PB, Vasudevan A. Generating a high quality compound collection. In: Wigglesworth M, Wood T, editors. *Management of chemical and biological samples for screening applications*. Weinheim: Wiley-VCH Verlag GmbH & co. KGaA; 2012. p 9–22.

[38] Grime KH, Barton P, McGinnity DF (2013). Application of in silico, in vitro and preclinical pharmacokinetic data for the effective and efficient prediction of human pharmacokinetics. *Mol. Pharm.* **10**:1191–1206.

[39] Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H (2013). Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discov.* **12**:948–962.

[40] Walters WP, Namchuk M (2003). Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* **2**:259–266.

[41] Baell JB, Holloway GA (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**:2719–2740.

[42] Baell JB (2013). Broad coverage of commercially available lead-like screening space with fewer than 350,000 compounds. *J. Chem. Inf. Model.* **53**:39–55.

[43] Thorne N, Auld DS, Inglese J (2010). Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr. Opin. Chem. Biol.* **14**:315–324.

[44] Bergström CA, Andersson SB, Fagerberg JH, Ragnarsson G, Lindahl A (2013). Is the full potential of the biopharmaceutics classification system reached? *Eur. J. Pharm. Sci.* **57**:224–231.

[45] http://www.acdlabs.com/products/percepta/

[46] http://www.schrodinger.com/productpage/14/17/

[47] http://www.simulations-plus.com/Products.aspx?pID=13

[48] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**:3–25.

[49] Jarvis L (2007). Originator of 'the rule of five' reflects on 10 years of drug development under its influence. *Chem. Eng. News* **85**:18–18.

[50] Baell J, Congreve M, Leeson P, Abad-Zapatero C (2013). Ask the experts: past, present and future of the rule of five. *Future Med. Chem.* **5**:745–752.

[51] Wenlock MC, Austin RP, Barton P, Davis AM, Leeson PD (2003). A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **46**:1250–1256.

[52] Teague SJ, Davis AM, Leeson PD, Oprea T (1999). The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* **111**:3962–3967.

[53] Leeson PD, Springthorpe B (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **6**:881–890.

[54] Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, Bender A, Jenkins JL, Urban L (2007). Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2**:874–880.

[55] Gleeson MP, Hersey A, Montanari D, Overington J (2011). Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discov.* **10**:197–208.

[56] Opera TI, editor. Chemoinformatics in lead discovery. In: *Chemoinformatics in drug discovery*. Weinheim: Wiley-VCH; 2005. p 25–42.

[57] Proudfoot JR (2005). The evolution of synthetic oral drug properties. *Bioorg. Med. Chem. Lett.* **15**:1087–1090.

[58] Oprea TI, Allu TK, Fara DC, Rad RF, Ostopovici L, Bologa CG (2007). Lead-like, drug-like or "Pub-like": how different are they? *J. Comput. Aided Mol. Des.* **21**:113–119.

[59] Congreve M, Carr R, Murray C, Jhoti H (2003). A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* **8**:876–877.

[60] Nadin A, Hattotuwagama C, Churcher I (2012). Lead-oriented synthesis: a new opportunity for synthetic chemistry. *Angew. Chem. Int. Ed. Engl.* **51**:1114–1122.

[61] Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006). Global mapping of pharmacological space. *Nat. Biotechnol.* **24**, 805–815.

[62] Hitchcock SA, Pennington LD (2006). Structure-brain exposure relationships. *J. Med. Chem.* **49**:7559–7583.

[63] O'Shea R, Moser HE (2008). Physicochemical properties of antibacterial compounds: implications for drug discovery. *J. Med. Chem.* **51**:2871–2878.

[64] Gleeson MP (2008). Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **51**:817–834.

[65] Gleeson MP, Bravi G, Modi S, Lowe D (2009). ADMET rules of thumb II: a comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg. Med. Chem.* **17**:5906–5919.

[66] Leeson PD, St-Gallay SA (2011). The influence of the 'organizational factor' on compound quality in drug discovery. *Nat. Rev. Drug Discov.* **10**:749–765.

[67] Johnson TW, Dress KR, Edwards M (2009). Using the Golden Triangle to optimize clearance and oral absorption. *Bioorg. Med. Chem. Lett.* **19**:5560–5564.

[68] Hill AP, Young RJ (2010). Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. *Drug Discov. Today* **15**:648–655.

[69] Hughes JD, Blagg J, Price DA, Bailey S, DeCrescenzo GA, Devraj RV, Ellsworth E, Fobian YM, Gibbs ME, Gilles RW, Greene N, Huang E, Krieger-Burke T, Loesel J, Wager T, Whitely L, Zhang Y (2008). Physicochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* **18**:4872–4875.

[70] Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012). Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**:90–98.

[71] Ritchie TJ, Macdonald SJ (2014). How drug-like are 'ugly' drugs: do drug-likeness metrics predict ADME behaviour in humans? *Drug Discov. Today* **19**:489–495.

[72] Clardy J, Walsh C (2004). Lessons from natural molecules. *Nature* **432**:829–837.

[73] Koehn FE, Carter GT (2005). The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.* **4**:206–220.

[74] Newman DJ, Cragg GM, Snader KM (2003). Natural products as sources of new drugs over the period 1981–2002. *J. Nat. Prod.* **66**:1022–1037.

[75] http://dnp.chemnetbase.com/intro/index.jsp

[76] Henkel T, Brunne R, Müller H, Reichel F (1999). Statistical investigation of structural complementarity of natural products and synthetic compounds. *Angew. Chem. Int. Ed. Engl.* **38**:643–647.

[77] Feher M, Scmidt JM (2003). Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **43**:218–227.

[78] Lee ML, Schneider G (2001). Scaffold architecture and pharmacophoric properties of natural products and trade drugs: applications in the design of natural products based combinatorial libraries. *J. Comb. Chem.* **3**:284–289.

[79] Vasilevich NI, Kombarov RV, Genis DV, Kirpichenok MA (2012). Lessons from natural products chemistry can offer novel approaches for synthetic chemistry in drug discovery. *J. Med. Chem.* **55**:7003–7009.

[80] Schmitt EK, Moore CM, Krastel P, Petersen F (2011). Natural products as catalysts for innovation: a pharmaceutical industry perspective. *Curr. Opin. Chem. Biol.* **15**:497–504.

[81] Tang Y, Khosla C. Biosynthesis of "unnatural" natural products. In: Bartlett PA, Entzeroth M, editors. *Exploiting chemical diversity for drug discovery*. Cambridge: Royal Society of Chemistry; 2006. p 137–162.

[82] Ganesan A (2008). The impact of natural products upon modern drug discovery. *Curr. Opin. Chem. Biol.* **12**:306–317.

[83] Koch MA, Wittenberg LO, Basu S, Jeyaraj DA, Gourzoulidou E, Reinecke K, Odermatt A, Waldmann H (2005). Compound library development guided by protein structure similarity clustering and natural product structure. *Proc. Natl. Acad. Sci. U. S. A.* **101**:16721–16726.

[84] Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* **41**(Database issue):D456–D463.

[85] Wermuth CG (2006). Selective optimization of side activities: the SOSA approach. *Drug Discov. Today* **11**:160–164.

[86] Hruby VJ (2002). Designing peptide receptor agonists and antagonists. *Nat. Rev. Drug Discov.* **1**:847–858.

[87] Wise A, Jupe SC, Rees S (2004). The identification of ligands at orphan G-protein coupled receptors. *Annu. Rev. Pharmacol. Toxicol.* **44**:43–66.

[88] Craik DJ, Fairlie DP, Liras S, Price D (2013). The future of peptide-based drugs. *Chem. Biol. Drug Des.* **81**:136–147.

[89] Hummel G, Reineke U, Reimer U. Translating peptides into small molecules. In: Bartlett PA, Entzeroth M, editors. *Exploiting chemical diversity for drug discovery*. Cambridge: Royal Society of Chemistry; 2006. p 57–90.

[90] Gupta S, Jain A, Chakraborty M, Sahni JK, Ali J, Dang S (2013). Oral delivery of therapeutic proteins and peptides: a review on recent developments. *Drug Deliv.* **20**:237–246.

[91] Khafagy EL-S, Morishita M (2012). Oral biodrug delivery using cell-penetrating peptide. *Adv. Drug Deliv. Rev.* **64**:531–539.

[92] Driggers EM, Hale SP, Lee J, Terrett NK (2008). The exploration of macrocycles for drug discovery—an underexploited structural class. *Nat. Rev. Drug Discov.* **7**:608–624.

[93] Mallinson J, Collins I (2012). Macrocycles in new drug discovery. *Future Med. Chem.* **4**:1409–1438.

[94] Giordanetto F, Kihlberg J (2014). Macrocyclic drugs and clinical candidates: what can medicinal chemists learn from their properties? *J. Med. Chem.* **57**:278–295.

[95] http://www.nature.com/scibx/collections/macrocycles

[96] Pitt WR, Parry DM, Perry BG, Groom CR (2009). Heteroaromatic rings of the future. *J. Med. Chem.* **52**:2952–2963.

[97] Ertl P, Jelfs S, Mühlbacher J, Schuffenhauer A, Selzer P (2006). Quest for the rings. In silico exploration of ring universe to identify novel bioactive heteroaromatic scaffolds. *J. Med. Chem.* **49**:4568–4573.

[98] Lovering F, Bikker J, Humblet C (2009). Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52**:6752–6756.

[99] Ishikawa M, Hashimoto Y (2011). Improvement in aqueous solubility in small molecule drug discovery programs by disruption of molecular planarity and symmetry. *J. Med. Chem.* **54**:1539–1554.

[100] Ritchie TJ, Macdonald SJ, Young RJ, Pickett SD (2011). The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types. *Drug Discov. Today* **16**:164–171.

[101] St Jean DJ Jr, Fotsch C (2012). Mitigating heterocycle metabolism in drug discovery. *J. Med. Chem.* **55**:6002–6020.

[102] Schreiber SL (2000). Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**:1964–1969.

[103] Trabocchi A, editor. *Diversity-oriented synthesis: basics and applications in organic synthesis, drug discovery, and chemical biology*. Hoboken: John Wiley & Sons, Inc.; 2013.

[104] Clemons PA, Bodycombe NE, Carrinski HA, Wilson JA, Shamji AF, Wagner BK, Koehler AN, Schreiber SL (2010). Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U. S. A.* **107**:18787–18792.

[105] Agarwal P, Sanseau P, Cardon LR (2013). Novelty in the target landscape of the pharmaceutical industry. *Nat. Rev. Drug Discov.* **12**:575–576.

[106] Lackey KE, editor. *Gene family targeted molecular design*. Hoboken: John Wiley & Sons, Inc.; 2008.

[107] Savchuk NP, Tkachenko SE, Balakin KV. Strategies for the design of pGPCR-targeted libraries. In: Rognan D, editor. *Methods and principles in medicinal chemistry, 30 (Ligand design for G protein-coupled receptors)*. Weinheim: Wiley-VCH; 2006. pp. 137–164.

[108] Jacoby E (2006). Designing compound libraries targeting GPCRs. *Ernst Schering Found. Symp. Proc.* **2**:93–103.

[109] Baringhaus KH, Hessler G. A chemical genomics approach for ion channel modulators. In: Kubinyi H, Müller G, editors. *Methods and principles in medicinal chemistry, 22 (Chemogenomics in drug discovery)*. Weinheim: Wiley-VCH; 2004. p 221–242.

[110] Harris CJ, Stevens AP (2006). Chemogenomics: structuring the drug discovery process to gene families. *Drug Discov. Today* **11**:880–888.

[111] Liao JJ (2007). Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors. *J. Med. Chem.* **50**:409–424.

[112] Sedrani R, Hommel U, Eder J. Protease-directed drug discovery. In: Lackey KE, editor. *Gene-family targeted molecular design*. Hoboken: John Wiley & Sons, Inc.; 2008. p 159–197.

[113] Moore JT, Collins JL, Pearce KH. The nuclear receptor superfamily and drug discovery. In: Schreiber SL, Kapoor TM, Wess G, editors. *Chemical biology—from small molecules to systems biology and drug design*. Weinheim: Wiley-VCH; 2007. p 891–932.

[114] Edwards AM, Bountra C, Kerr DJ, Willson TM (2009). Open access chemical and clinical probes to support drug discovery. *Nat. Chem. Biol.* **5**:436–440.

[115] Sippl W, Jung M, editors. *Epigenetic targets in drug discovery*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA; 2009.

[116] Helin K, Dhanak D (2013). Chromatin proteins and modifications as drug targets. *Nature* **502**:480–488.

[117] Schneider G, editor. *De novo molecular design*. Weinheim: Wiley-VCH Verlag GmbH & Co. KgaA; 2013.

[118] Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang XP, Norval S, Sassano MF, Shin AI, Webster LA, Simeons FR, Stojanovski L, Prat A, Seidah NG, Constam DB, Bickerton GR, Read KD, Wetsel WC, Gilbert IH, Roth BL, Hopkins AL (2012). Automated design of ligands to polypharmacological profiles. *Nature* **492**:215–220.

[119] Hartenfeller M, Renner S, Jacoby E. Reaction-driven de novo design: a keystone for automated design of target family-oriented libraries. In: Schneider G, editor. *De novo molecular design*. Weinheim: Wiley-VCH Verlag GmbH & Co. KgaA; 2013. p 245–263.

[120] Hartenfeller M, Eberle M, Meier P, Nieto-Oberhuber C, Altmann KH, Schneider G, Jacoby E, Renner S (2012). Probing the bioactivity-relevant chemical space of robust reactions and common molecular building blocks. *J. Chem. Inf. Model.* **52**:1167–1178.

[121] Wilk W, Zimmermann TJ, Kaiser M, Waldmann H (2010). Principles, implementation, and application of biology-oriented synthesis (BIOS). *Biol. Chem.* **391**:491–497.

[122] Wetzel S, Bon RS, Kumar K, Waldmann H (2011). Biology-oriented synthesis. *Angew. Chem. Int. Ed. Engl.* **50**:10800–10826.

[123] Zimmermann TJ, Roy S, Martinez NE, Ziegler S, Hedberg C, Waldmann H (2013). Biology-oriented synthesis of a tetrahydroisoquinoline-based compound collection targeting microtubule polymerization. *Chembiochem* **14**:295–300.

[124] Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL, Lundell GF, Veber DF, Anderson PS, Chang RS (1988). Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **31**:2235–2246.

[125] Müller G. Target family-directed masterkeys in chemogenomics. In: Kubinyi H, Müller G, editors. *Methods and principles in medicinal chemistry, 22 (Chemogenomics in drug discovery)*. Weinheim: Wiley-VCH; 2004. p 7–41.

[126] Hu Y, Wassermann AM, Lounkine E, Bajorath J (2010). Systematic analysis of public domain compound potency data identifies selective molecular scaffolds across druggable target families. *J. Med. Chem.* **53**:752–758.

[127] Kombarov R, Altieri A, Genis G, Kirpichenok M, Kochubey V, Rakitina N, Titarenko Z (2010). BioCores: identification of a drug/natural product-based privileged structural motif for small-molecule lead discovery. *Mol. Divers.* **14**:193–200.

[128] Marshall GR, Kuster DJ, Che Y (2009). Chemogenomics with protein secondary-structure mimetics. *Methods Mol. Biol.* **575**:123–158.

[129] Garland SL, Dean PM (1999). Design criteria for molecular mimics of fragments of the beta-turn. 1. C alpha atom analysis. *J. Comput. Aided Mol. Des.* **13**:469–483.

[130] Garland SL, Dean PM (1999). Design criteria for molecular mimics of fragments of the beta-turn. 2. C alpha-C beta bond vector analysis. *J. Comput. Aided Mol. Des.* **13**:485–498.

[131] Webb TR, Jiang L, Sviridov S, Venegas RE, Vlaskina AV, McGrath D, Tucker J, Wang J, Deschenes A, Li R (2007). Application of a novel design paradigm to generate general nonpeptide combinatorial templates mimicking beta-turns: synthesis of ligands for melanocortin receptors. *J. Comb. Chem.* **9**:704–710.

[132] Cummings CG, Hamilton AD (2010). Disrupting protein-protein interactions with non-peptidic, small molecule alpha-helix mimetics. *Curr. Opin. Chem. Biol.* **14**:341–346.

[133] Fry D, Huang KS, Di Lello P, Mohr P, Müller K, So SS, Harada T, Stahl M, Vu B, Mauser H (2013). Design of libraries targeting protein-protein interfaces. *ChemMedChem* **8**:726–732.

[134] Morelli X, Bourgeas R, Roche P (2011). Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.* **15**:475–481.

[135] Czarna A, Beck B, Srivastava S, Popowicz GM, Wolf S, Huang Y, Bista M, Holak TA, Dömling A (2010). Robust generation of lead compounds for protein-protein interactions by computational and MCR chemistry: p53/Hdm2 antagonists. *Angew. Chem. Int. Ed. Engl.* **49**:5352–5356.

[136] Ji HF, Kong DX, Shen L, Chen LL, Ma BG, Zhang HY (2007). Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol.* **8**:R176.

[137] Schuffenhauer A, Ruedisser S, Marzinzik AL, Jahnke W, Blommers M, Selzer P, Jacoby E. (2005). Library design for fragment based screening. *Curr. Top. Med. Chem.* **5**:751–762.

[138] Siegal G, Ab E, Schultz J (2007). Integration of fragment screening and library design. *Drug Discov. Today* **12**:1032–1039.

[139] Chen IJ, Hubbard RE (2009). Lessons for fragment library design: analysis of output from multiple screening campaigns. *J. Comput. Aided Mol. Des.* **23**:603–629.

[140] Genis D, Kirpichenok M, Kombarov R (2012). A minimalist fragment approach for the design of natural-product-like synthetic scaffolds. *Drug Discov. Today* **17**:1170–1174.

[141] Over B, Wetzel S, Grütter C, Nakai Y, Renner S, Rauh D, Waldmann H (2013). Natural-product-derived fragments for fragment-based ligand discovery. *Nat. Chem.* **5**:21–28.

[142] Hubbard RE, Chen I, Davis B (2007). Informatics and modeling challenges in fragment-based drug discovery. *Curr. Opin. Drug Discov. Devel.* **10**:289–297.

[143] Murray CW, Blundell TL (2010). Structural biology in fragment-based drug design. *Curr. Opin. Struct. Biol.* **20**:497–507.

[144] Hajduk PJ, Greer J (2007). A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discov.* **6**:211–219.

[145] Baker M (2013). Fragment-based lead discovery grows up. *Nat. Rev. Drug Discov.* **12**:5–7.

[146] Brenner S, Lerner RA (1992). Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. U. S. A.* **89**:5381–5383.

[147] Clark MA (2010). Selecting chemicals: the emerging utility of DNA-encoded libraries. *Curr. Opin. Chem. Biol.* **14**:396–403.

[148] Mannocci L, Leimbacher M, Wichert M, Scheuermann J, Neri D (2011). 20 years of DNA-encoded chemical libraries. *Chem. Commun. (Camb.)* **47**:12747–12753.

[149] Kleiner RE, Dumelin CE, Liu DR (2011). Small-molecule discovery from DNA-encoded chemical libraries. *Chem. Soc. Rev.* **40**:5707–5717.

[150] Goodnow RA, editor. *A handbook for DNA-encoded chemistry: theory and applications for exploring chemical space and drug discovery*. Hoboken: John Wiley & Sons, Inc.; 2014.

[151] Hajduk PJ, Galloway WR, Spring DR (2011). Drug discovery: a question of library design. *Nature* **470**:42–43.

[152] Barker A, Kettle JG, Nowak T, Pease E (2013). Expanding medicinal chemistry space. *Drug Discov. Today* **18**:298–304.

[153] Zhou JZ. Chemical library design. In: Walker JM, editor. *Methods in molecular biology (Volume 685)*. New York: Humana Press, Springer; 2011.

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

[155] Sotriffer C. Virtual screening: principles, challenges, and practical guidelines. In: Mannhold R, Kubinyi H, Folkers G, editors. *Methods and principles in medicinal chemistry (Volume 48)*. Weinheim: VCH-Wiley; 2011.