

# Lead Discovery and Lead Modification

## Chapter Outline

<b>2.1. Lead Discovery</b>	<b>20</b>		
2.1.1. General Considerations	20	2.2.5.1. Electronic Effects: The Hammett Equation	72
2.1.2. Sources of Lead Compounds	20	2.2.5.2. Lipophilicity Effects	74
2.1.2.1. Endogenous Ligands	20	2.2.5.2.1. Importance of Lipophilicity	74
2.1.2.2. Other Known Ligands	23	2.2.5.2.2. Measurement of Lipophilicities	74
2.1.2.3. Screening of Compounds	24	2.2.5.2.3. Computer Automation of log <i>P</i> Determination	78
2.1.2.3.1. Sources of Compounds for Screening	26	2.2.5.2.4. Membrane Lipophilicity	79
2.1.2.3.1.1. Natural Products	26	2.2.5.3. Balancing Potency of Ionizable Compounds with Lipophilicity and Oral Bioavailability	79
2.1.2.3.1.2. Medicinal Chemistry Collections and Other “Handcrafted” Compounds	27	2.2.5.4. Properties that Influence Ability to Cross the Blood–Brain Barrier	81
2.1.2.3.1.3. High-Throughput Organic Synthesis	27	2.2.5.5. Correlation of Lipophilicity with Promiscuity and Toxicity	82
2.1.2.3.1.3.1. Solid-Phase Library Synthesis	27	2.2.6. Computational Methods in Lead Modification	83
2.1.2.3.1.3.2. Solution-Phase Library Synthesis	30	2.2.6.1. Overview	83
2.1.2.3.1.3.3. Evolution of HTOS	31	2.2.6.2. Quantitative Structure–Activity Relationships (QSARs)	83
2.1.2.3.2. Drug-Like, Lead-Like, and Other Desirable Properties of Compounds for Screening	32	2.2.6.2.1. Historical Overview. Steric Effects: The Taft Equation and Other Equations	83
2.1.2.3.3. Random Screening	36	2.2.6.2.2. Methods Used to Correlate Physicochemical Parameters with Biological Activity	84
2.1.2.3.4. Targeted (or Focused) Screening, Virtual Screening, and Computational Methods in Lead Discovery	36	2.2.6.2.2.1. Hansch Analysis: A Linear Multiple Regression Analysis	84
2.1.2.3.4.1. Virtual Screening Database	37	2.2.6.2.2.2. Manual Stepwise Methods: Topliss Operational Schemes and Others	85
2.1.2.3.4.2. Virtual Screening Hypothesis	37	2.2.6.2.2.3. Batch Selection Methods: Batchwise Topliss Operational Scheme, Cluster Analysis, and Others	87
2.1.2.3.5. Hit-To-Lead Process	43	2.2.6.2.2.4. Free and Wilson or de Novo Method	88
2.1.2.3.6. Fragment-based Lead Discovery	45	2.2.6.2.2.5. Computational Methods for ADME Descriptors	89
<b>2.2. Lead Modification</b>	<b>54</b>	2.2.6.3. Scaffold Hopping	89
2.2.1. Identification of the Active Part: The Pharmacophore	55	2.2.6.4. Molecular Graphics-Based Lead Modification	90
2.2.2. Functional Group Modification	57	2.2.7. Epilogue	93
2.2.3. Structure–Activity Relationships	57	<b>2.3. General References</b>	<b>95</b>
2.2.4. Structure Modifications to Increase Potency, Therapeutic Index, and ADME Properties	59	<b>2.4. Problems</b>	<b>102</b>
2.2.4.1. Homologation	60	<b>References</b>	<b>106</b>
2.2.4.2. Chain Branching	61		
2.2.4.3. Bioisosterism	62		
2.2.4.4. Conformational Constraints and Ring-Chain Transformations	66		
2.2.4.5. Peptidomimetics	68		
2.2.5. Structure Modifications to Increase Oral Bioavailability and Membrane Permeability	72		

## 2.1. LEAD DISCOVERY

### 2.1.1. General Considerations

As discussed in the drug discovery overview in Chapter 1, identification of suitable lead compounds provides starting points for lead optimization, during which leads are modified to achieve requisite potency and selectivity, as well as absorption, distribution, metabolism, and excretion (ADME), and intellectual property (patent) position. Given the hurdles often presented by these multiple and diverse objectives, identification of the best lead compounds can be a critical factor to the overall success of a drug discovery program. The approach to lead identification taken in a given drug discovery program will usually take into account any known *ligand* (a smaller molecule that binds to a receptor) for the target. At one extreme, if there are already marketed drugs for a particular target, these may serve as lead compounds; however, in this case, establishing a suitable intellectual property position may be the greatest challenge. On the other hand, whereas the *endogenous ligand* (the molecule that binds to a biological target in an organism and is believed to be responsible for the native activity of the target) has provided good lead structures for many programs, the endogenous ligand for a new biological target may not be well characterized, or the only known ligand may not be attractive as a lead compound. For example, if an endogenous ligand is a complex molecule that is not readily amenable to synthetic modification or has some other undesirable properties that are not reasonably addressable, it may not be attractive as a lead, and other approaches to lead discovery must be considered. In the next few sections, we will first provide additional examples of endogenous or other known ligands as lead compounds to complement the examples given in Chapter 1, and then we will turn to a more detailed discussion of alternative approaches to lead discovery.

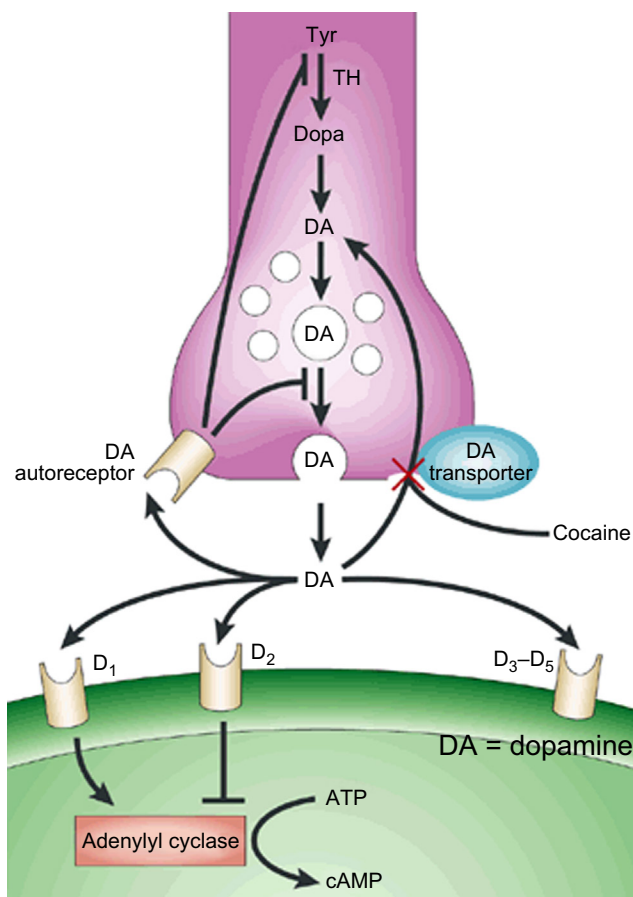
### 2.1.2. Sources of Lead Compounds

Lead compounds can be acquired from a variety of sources: endogenous ligands, e.g., substrates for enzymes and transporters or agonists for receptors; other known ligands, including marketed drugs, compounds isolated in drug metabolism studies, and compounds used in clinical trials; and through screening of compounds, including natural products and other chemical libraries, either at random or in a targeted approach.

#### 2.1.2.1. Endogenous Ligands

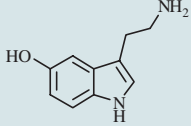
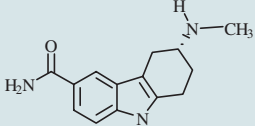
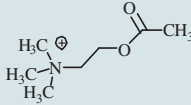
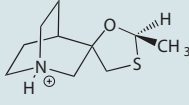
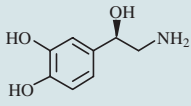
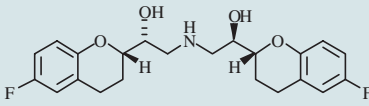
Rational approaches are important routes to lead discovery. The first step is to identify the cause for the disease state. Many diseases, or at least the symptoms of diseases, arise from an imbalance (either excess or deficiency) of a particular chemical in the body, from the invasion of a foreign organism, or from aberrant cell growth. As will be discussed

in later chapters, the effects of the imbalance can be corrected by antagonism or agonism of a receptor (see Chapter 3) or by inhibition of a particular enzyme (see Chapter 5); interference with deoxyribonucleic acid (DNA) biosynthesis or function (see Chapter 6) is another important approach to treating diseases arising from microorganisms or aberrant cell growth. Once the relevant biochemical system is identified, initial lead compounds become the endogenous receptor ligands or enzyme substrates. In Chapter 1, the example of dopamine as a lead compound for the discovery of rotigotine (**1.28**) was presented. Dopamine is the endogenous ligand for dopamine receptors, including the D<sub>3</sub> receptor, which is the target of rotigotine. Dopamine is one of a number of important *neurotransmitters*, substances released by nerve cells (*neurons*) that interact with receptors on the surface of nearby neurons to propagate a nerve signal (Figure 2.1). Endogenous neurotransmitters have served as lead compounds for many important drugs. Table 2.1 shows



**FIGURE 2.1** Depiction of dopamine (DA) in its role as a neurotransmitter. DA is released by a neuron prior to interacting with dopamine receptors (D<sub>1</sub>–D<sub>5</sub>) on the surface of another nearby neuron. Also shown is the dopamine transporter, which terminates the action of dopamine by transporting the released neurotransmitter from the synaptic cleft back into the presynaptic neuron. Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Drug Discovery* (Kreek, M. J.; LaForge, K. S.; Butelman, E. *Pharmacotherapy of addictions*. *Nat. Rev. Drug Discov.* 2002, 1, 710–726) Copyright 2002.

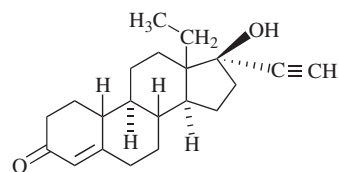
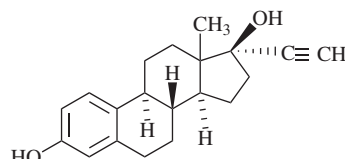
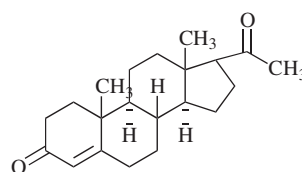
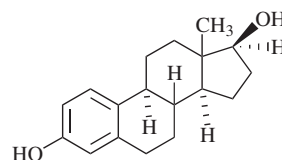
**TABLE 2.1** Examples of Endogenous Neurotransmitter Ligands That Have Served as Lead Compounds for Drug Discovery

Endogenous Ligand	Marketed Drug
 Serotonin	 Frovatriptan (antimigraine)
 Acetylcholine	 Cevimeline (dry mouth treatment)
 Norepinephrine	 Nebivolol (antihypertensive)

examples of the drugs that evolved from the structures of the endogenous neurotransmitters serotonin, acetylcholine, and norepinephrine.

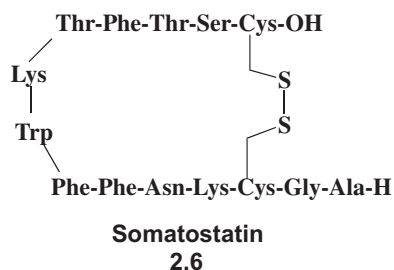
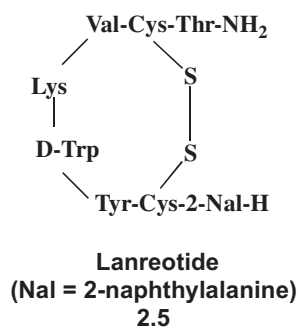
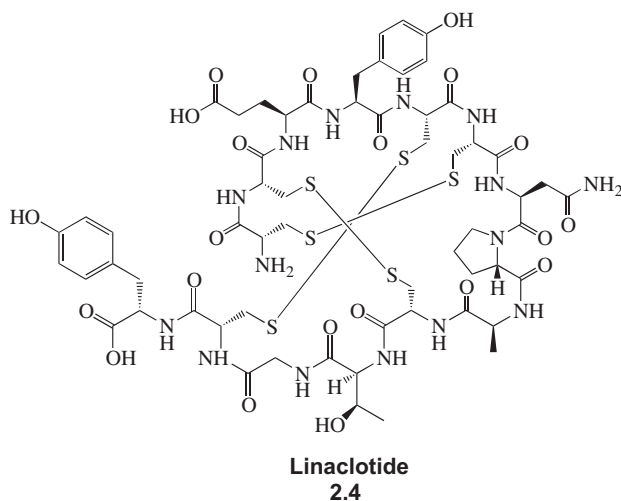
Hormones are another important class of endogenous substances that have served as lead compounds for drug discovery. Like neurotransmitters, hormones are released from cells and interact with receptors on the surface of other cells. However, whereas receptors for neurotransmitters are close to the site of neurotransmitter release, hormone receptors can be at quite some distance from the site of hormone release, so hormones have to travel to their site of action through the bloodstream. Steroids are one important class of hormones; lead compounds for the contraceptives (+)-norgestrel (**2.1**, Ovral) and 17 $\alpha$ -ethynyl estradiol (**2.2**, Activella) were the steroidal hormones progesterone (**2.3a**) and 17 $\beta$ -estradiol (**2.3b**), respectively. The endogenous steroid hormones (**2.3a** and **2.3b**) show weak and short-lasting effects, whereas oral contraceptives (**2.1** and **2.2**) exert strong progestational activity of long duration.

Peptides constitute another broad class of hormones. Peptides, like proteins, consist of a sequence of amino acid residues, but are smaller than proteins (in the range of two to approximately 100 amino acids). Most peptides have low stability in plasma as a result of the ubiquitous presence of *peptidases* (enzymes that catalyze hydrolysis of peptides into smaller peptides or constituent amino acids). Moreover, peptides usually cannot be delivered orally because of low permeability across gut membranes (as a result of their charge and polarity) and because of instability to gut peptidases. However, incorporation of disulfide bonds to cross-link a peptide can

**Norgestrel**  
**2.1****17 $\alpha$ -Ethynyl estradiol**  
**2.2****2.3a****2.3b**

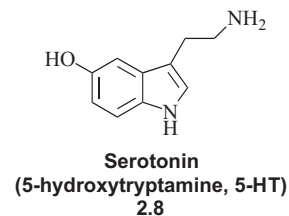
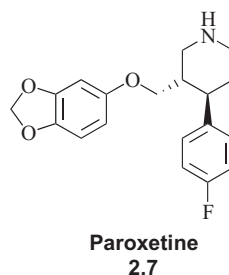
confer enzymatic stability, e.g., linaclotide (**2.4**, Linzess) used to treat bowel diseases. Considerable effort has been devoted to the goal of using natural peptides as lead compounds for the discovery of derivatives with improved properties. One successful drug that resulted from these endeavors is lanreotide (**2.5**, Somatuline),<sup>[1]</sup> a long-acting analog of the peptide hormone somatostatin (**2.6**), which is administered by injection to treat *acromegaly* (thickening of skin and enlargement of hands and feet from overproduction of growth hormone).

The discussion of endogenous ligands so far has focused on leads for drugs designed to interact with receptor targets. Endogenous ligands for other types of drug targets, including transporters and enzymes, have also served as valuable starting points for drugs. As mentioned in Chapter 1, transporters are proteins that help transport substances across cell membranes. One important class of transporters is responsible for neurotransmitter reuptake.<sup>[2]</sup> As illustrated in Figure 2.1 for the neurotransmitter dopamine, after dopamine is released into the synaptic cleft, excess neurotransmitter is transported back into the neuron that released it (the presynaptic neuron) by specific transporter proteins, which serves to deactivate the signal

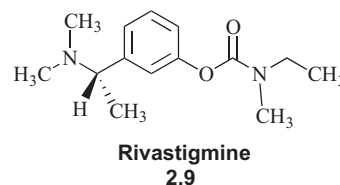


carried by the neurotransmitter. Therefore, an inhibitor of a neurotransmitter reuptake transporter would have the effect of prolonging the action of the neurotransmitter. Cocaine exerts its effects by inhibiting the dopamine reuptake transporter. Inhibitors of the reuptake transporters for other important neurotransmitters, such as norepinephrine and serotonin, comprise important classes of antidepressant drugs. The leads for many of these reuptake inhibitors were

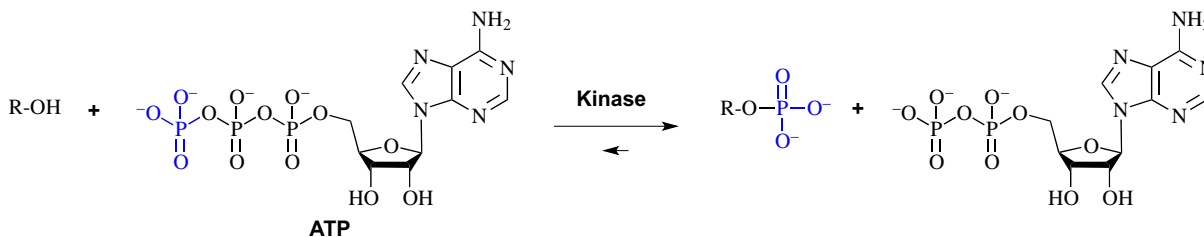
the transporter ligands, that is, norepinephrine or serotonin. Paroxetine (**2.7**, Paxil) is an example of a selective serotonin reuptake inhibitor marketed as an antidepressant drug with considerable structural resemblance to serotonin (**2.8**). Transporters of glucose have recently been targeted for the treatment of type 2 diabetes.<sup>[3]</sup>



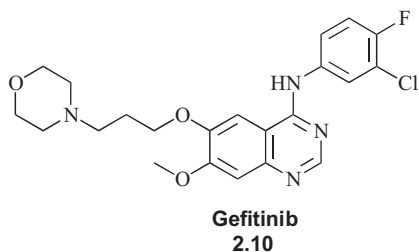
Similarly, an important source of leads for the design of enzyme inhibitors can be the corresponding enzyme substrate. For example, rivastigmine (**2.9**, Exelon) is an acetyl cholinesterase inhibitor prescribed as a treatment for dementia, for which the ultimate starting point was acetylcholine (Table 2.1), although in actuality, the evolution of rivastigmine occurred across several generations of drugs (you are probably thinking it is hard to see how this structure could come from acetylcholine, but that is how lead optimization evolves new structures).



Another example of using an enzyme substrate as a lead for drug discovery is in the design of kinase inhibitors. Kinases catalyze the transfer of the terminal phosphate group of adenosine triphosphate (ATP) and related molecules usually to the hydroxyl group of another molecule (Scheme 2.1), for example, to the hydroxyl group on the tyrosine residue of a substrate protein (protein tyrosine kinase). Thus, kinases have two substrates, ATP (the phosphate donor) and the phosphate acceptor. Many kinase inhibitors were ultimately designed based on the structure of ATP, for example, gefitinib (**2.10**, Iressa), which is used for the treatment of lung cancer.



**SCHEME 2.1** Reaction catalyzed by the kinase class of enzymes. Kinases catalyze the transfer of the terminal phosphate group of ATP or related molecules acceptor to the group of a substrate, in this case, an alcohol (ROH).



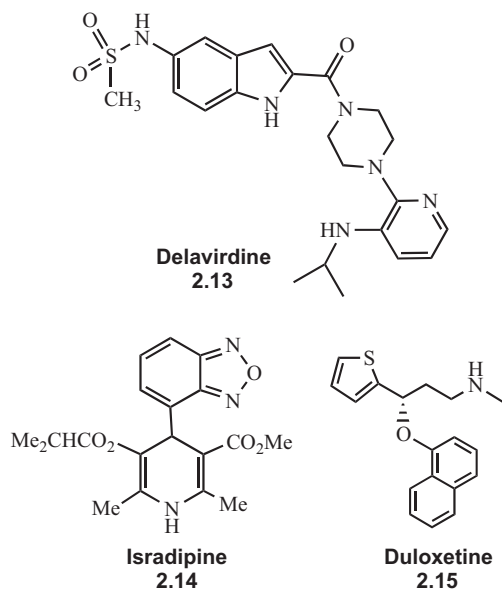
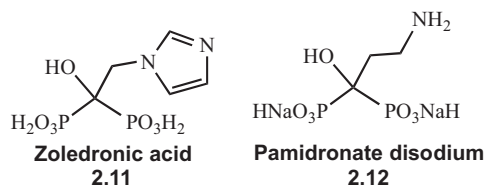
Currently, rational approaches to drug discovery are most relevant to the earlier stages of the process, most notably including target identification, lead discovery, and optimization of molecular interactions with the target during lead optimization. Later stages of drug discovery presently remain much more empirical owing to the difficulties in accurately predicting toxicities, anticipating transport properties, accurately predicting the full range of ADME properties of a drug, and numerous other factors. However, active ongoing research is attempting to increase the degree of rationality even for these complex facets of drug behavior. In addition to rational approaches, particularly when no target protein is known or little structural information is available for rational design, other less rational approaches can be taken to get a starting point for lead discovery using other known ligands or screening approaches.

### 2.1.2.2. Other Known Ligands

In Chapter 1, the example of using the plant alkaloid cytosine (**1.29**) as the starting point for discovery of the smoking cessation agent varenicline (**1.31**, Chantix) was described. Another variant of using a known ligand as a starting point is the use of an established drug as a lead toward development of the next generation of compounds.<sup>[4]</sup> One example is diazepam (**1.17**, Valium), as described in Chapter 1, Section 1.2.3, which was derived from the marketed drug Librium (**1.13**) and is almost 10 times more potent than the lead. Another example is zoledronic acid (**2.11**, Zometa), which is used to treat *osteoporosis* (loss of bone density) and *hypercalcemia*, a condition resulting in high blood calcium levels due to cancer, and to delay bone complications resulting from multiple myeloma and bone metastases. This is a second-generation drug derived from pamidronate disodium (**2.12**, Aredia), also used for treating hypercalcemia from malignancy.

Known drugs can also be *repurposed* (the identification and development of new uses for existing or abandoned drugs; also called *repositioned*) for a completely different indication.<sup>[5]</sup> The advantage of a repurposed drug is that the cost to bring it to market is diminished because the safety and pharmacokinetic profiles have already been established for its original indication. A *library* (a collection of compounds) of 3665 Food and Drug Administration (FDA)-approved and investigational drugs was tested for activity against hundreds of targets, from which 23 new drug–target relationships were confirmed.<sup>[6]</sup> For example, the reverse transcriptase inhibitor and acquired immune deficiency syndrome (AIDS) drug

delavirdine (**2.13**, Rescriptor) was found to antagonize the histamine H<sub>4</sub> receptor, which is a target for the potential treatment of asthma and allergies. Isradipine (**2.14**, Dynacirc), an antihypertensive drug, is in clinical trials as a treatment for Parkinson's disease.<sup>[7]</sup> The antidepressant drug duloxetine (**2.15**, Cymbalta) has been approved to treat chronic lower back pain.<sup>[8]</sup> A common dilemma to the repurposing of marketed drugs is that if the repurposed drug is used directly for a new indication, then only a new *method of use patent* (a patent that covers the new use for the molecule) application can be filed; however, it is best to own the rights to a molecule for *any purpose (composition of matter patent)*, which an altered structure would allow. Therefore, using a known drug as a lead to discover a novel compound could warrant independent patent protection for the new structure. An important advantage to repurposed drugs is that whereas only 10% of new drugs in Phase I clinical trials and 50% of Phase III drugs make it to the market, the rates for repurposed drugs are 25 and 65%, respectively.



Other sources of lead compounds, as described in Chapter 1, Sections 1.2.4 and 1.2.5, are metabolism studies and clinical trials. The cases cited in those sections involved the identification of new drugs from metabolism or from the clinic, some with novel indications; however, it is possible that the metabolite from a drug metabolism study or a compound in a clinical trial might act as a lead compound for a new indication requiring modification to enhance its potency or diminish undesirable properties.



### 2.1.2.3. Screening of Compounds

Endogenous or other ligands may not be known for a target of interest. Alternatively, known ligands for a target may not be well suited as starting points for discovery of drugs that will ultimately possess the desired properties. For example, many endogenous ligands are large proteins, which are not usually good leads when the goal is to discover an orally administered drug. For these reasons, screening for leads has played a central role in drug discovery for decades, although technological advances in the past 20 years have markedly changed how these screens are conducted, as discussed below.

The first requirement for a screening approach is to have a means to assay compounds for a particular biological activity, so that researchers will know when a compound is active. *Bioassay* (or *screen*) is a means of determining in a biological system, relative to a control compound, if a compound has the desired activity, and if so, what the relative potency of the compound is. Note the distinction between the terms activity and potency. *Activity* is the particular biological or pharmacological effect (for example, antibacterial activity or anticonvulsant activity); *potency* is the strength of that effect.

Until the late 1980s many screening efforts were conducted using whole animals or whole organisms, for example, screening for antiepileptic activity by assessing the ability of a compound to prevent an induced seizure in a mouse or rat, or for antibacterial activity by measuring the effect of test compounds on the growth of bacterial cultures in glass dishes. Especially when screening in whole animals, efforts have often been hampered by the comparatively large quantities of test compound required and by the fact that the results depended on other factors apart from the inherent potency of the compound at its intended target (*pharmacodynamics*), for example, the ability of the compound to be absorbed, distributed, metabolized, and excreted (*pharmacokinetics*). Thus, in general, *in vitro* tests have fewer confounding factors and are also quicker and less expensive to perform. The downside to this approach, however, is that you may identify a very potent compound for a target, but it may not have the ability to be absorbed or is rapidly metabolized. This more rapid screening method then requires additional studies of pharmacokinetics once the appropriate pharmacodynamics has been established. Pharmacokinetic aspects are discussed further throughout the chapter.

An exciting approach for screening compounds that might interact with an enzyme in a metabolic pathway was demonstrated by Wong, Pompliano, and coworkers for the discovery of lead compounds that block bacterial cell wall biosynthesis (as potential antibacterial agents).<sup>[9]</sup> Conditions were found to reconstitute all six enzymes in the cell wall biosynthetic pathway so that incubation with the substrate for the first enzyme led to the formation of the product of the last enzyme in the pathway. Then by screening compounds and looking for the buildup of an intermediate it was possible to identify compounds that blocked the pathway (and prevented the

formation of the bacterial cell wall) and also which enzyme was blocked (the buildup of an intermediate meant that the enzyme that acted on that intermediate was blocked).

Compound screening also can be carried out by electrospray ionization mass spectrometry (MS)<sup>[10]</sup> (the technique for which John Fenn received the Nobel Prize in 2002) and by nuclear magnetic resonance (NMR) spectrometry (the technique for which Richard Ernst and Kurt Wüthrich received Nobel Prizes in 1991 and 2002, respectively).<sup>[11]</sup> Tightly bound noncovalent complexes of compounds with a macromolecule (such as a receptor or enzyme) can be observed in the mass spectrum. The affinity of the ligand can be measured by varying the collision energy and determining at what energy the complex dissociates. This method also can be used to screen mixtures of compounds, provided they have different molecular masses and/or charges, so that *m/z* for each complex with the biomolecule can be separated in the mass spectrometer. By varying the collision energy, it is possible to determine which test molecules bind to the biomolecule best. The <sup>1</sup>H NMR method exploits changes in either relaxation rates or diffusion rates of small molecules when they bind to a macromolecule. This method can also be used to screen mixtures of compounds to determine the ones that bind best.

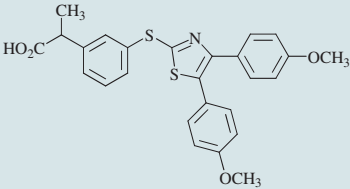
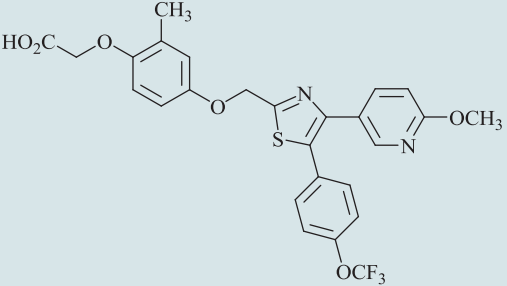
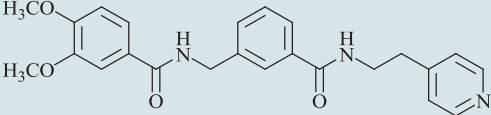
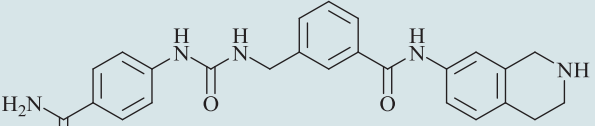
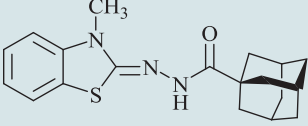
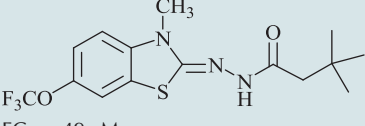
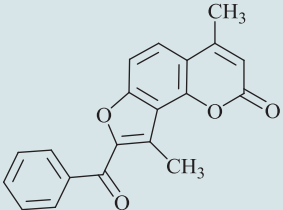
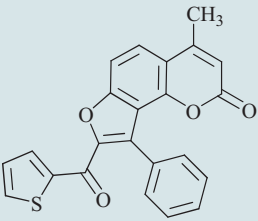
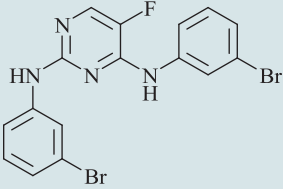
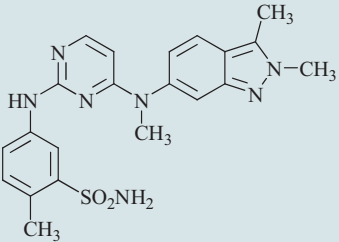
*High-throughput screening* (HTS),<sup>[12]</sup> from which greater than two-thirds of drug discovery projects now originate,<sup>[13]</sup> was initially developed in the late 1980s employing very rapid and sensitive *in vitro* screens, which could be carried out robotically. According to Drews,<sup>[14]</sup> the number of compounds assayed in a large pharmaceutical company in the early 1990s was about 200,000 a year, which rose to 5–6 million during the mid-1990s, and by the end of the 1990s it was >50 million! HTS can be carried out robotically in 1536- or 3456-well titer plates on small (submicrogram) amounts of compound (dissolved in submicroliter volumes). With these ultrahigh throughput screening approaches of the early part of the twenty-first century,<sup>[15]</sup> it is possible to screen 100,000 compounds in a day! In 2010, an HTS method using *drop-based microfluidics* (the ability to manipulate tiny volumes of liquid) was reported that allowed a 1000 times faster screening (10 million reactions per hour) with 10<sup>-7</sup> times the reagent volume and at one-millionth the cost of conventional techniques.<sup>[16]</sup> In this technique, drops of aqueous fluid dispersed in fluorocarbon oil replace the microtiter plates, which allows analysis and compound sorting in picoliter volume reactions while reagents flow through channels. A silicone sheet of lenses can be used to cover the microfluidic arrays, allowing fluorescence measurements of 62 different output channels simultaneously and analysis of 200,000 drops per second.<sup>[17]</sup> Therefore, screening compounds is no longer the slow step in the lead discovery process!

Because of the ease of screening vast numbers of compounds, early in the application of HTS, every compound in the company library, regardless of its properties, was screened. By the early part of the first decade of the twenty-first century, because an increase in the number of useful

lead compounds was not forthcoming despite the huge rise in the application of screening, it was realized that the physicochemical properties of molecules were key for screening compounds.<sup>[18]</sup> Therefore, additional considerations for HTS became the sources and selection of compounds to be screened and the development of effective methods for processing and utilizing the screening data that were generated.

Medicinal chemists have an important role in these activities, which we discuss in more detail in the next several sections. A keyword search for “high-throughput screening” in the *Journal of Medicinal Chemistry* website (<http://pubs.acs.org/journal/jmcmr>) readily retrieves a multitude of examples in which HTS played a central role in lead discovery. Representative examples are shown in Table 2.2, together with

**TABLE 2.2** Examples of Hits from HTS and Analogs Resulting from Subsequent Optimization Efforts

Biological Target HTS Hit	Representative Structure after Initial or Full Optimization
Peroxisome proliferator-activated receptor (PPAR) $\delta$ (target class: nuclear hormone receptor) $EC_{50} = 3200 \text{ nM}$ 	 $EC_{50} = 17 \text{ nM}$
Rho kinase (target class: enzyme) $IC_{50} = 2300 \text{ nM}$ 	 $IC_{50} = 4 \text{ nM}$
KCNQ2/Q3 potassium channels (target class: ion channel) $EC_{50} = 27 \text{ nM}$ 	 $EC_{50} = 49 \text{ nM}$ Significantly increased oral efficacy
Influenza A (H1N1) virus $IC_{50} = 4500 \text{ nM}$ 	 $IC_{50} = 70 \text{ nM}$
Vascular endothelial growth factor receptor 2 kinase domain (target class: enzyme) $IC_{50} \sim 400 \text{ nM}$ 	 $IC_{50} = 30 \text{ nM}$ Marketed drug (pazopanib)

structures of products from subsequent lead optimization activities.<sup>[19]</sup> See Section 2.2 for what properties need to be considered prior to and during the lead optimization process.

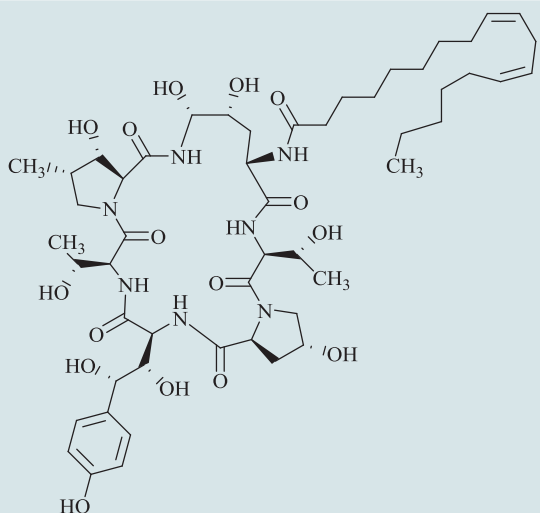
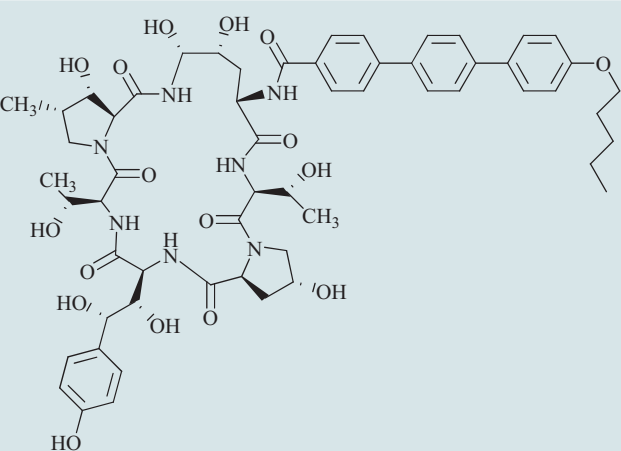
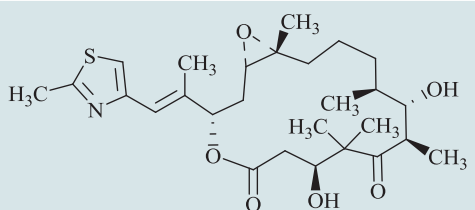
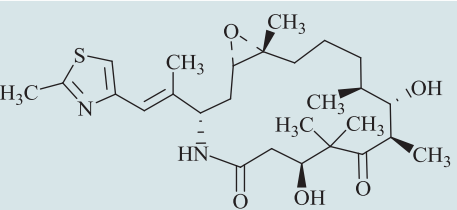
### 2.1.2.3.1. Sources of Compounds for Screening

As stated above, besides a high-throughput assay, an essential second requirement for HTS is a large number of suitable compounds for screening. In the following several subsections, we discuss the most common sources of compounds for HTS. The criteria for selecting compounds to be added to a general screening collection and for improving the selection of specific compounds for a given screen have evolved considerably over the past decade. An important goal of an organization that conducts many HTS campaigns across a variety of types of biological targets will be to construct a screening library of structurally diverse compounds. The assumption is that structurally similar compounds will have similar biological activities, and conversely, that structurally diverse collections will show divergent biological activities. In general, this is the case; however, such generalizations should be made with caution, since Dixon and Villar showed that a protein can bind a set of structurally diverse molecules with similar potent binding affinities, and analogs closely related to these compounds can exhibit very weak binding.<sup>[20]</sup>

**2.1.2.3.1.1. Natural Products** Nature is still an excellent source of drug precursors, or in some cases, of actual drugs. Although endogenous ligands discussed earlier are technically also natural products, the present category is intended to encompass products from nonmammalian natural sources, for example, plants, marine organisms, bacteria, and fungi. Nearly half of the new drugs approved between 1994 and 2007 are based on natural products, including 13 natural product-related drugs approved from 2005 to 2007.<sup>[21]</sup> More than 60% of the anticancer and anti-infective agents that went on the market between 1981 and 2006 were of natural product origin or derived from natural products; if biologicals, for example, antibodies and genetically engineered proteins, and vaccines are ignored, then the percentage increases to 73%.<sup>[22]</sup> This may be a result of the inherent nature of these secondary metabolites as a means of defense for their producing organisms; for example, a fungal natural product that inhibits cell replication may be produced by the fungus to act on potential invading organisms such as bacteria or other fungi.<sup>[23]</sup> Table 2.3 shows two examples of recently approved drugs that were derived from natural product lead compounds<sup>[24]</sup>; many others are currently in various stages of clinical development.

It has been suggested that small molecule natural products tend to target essential proteins of genes from organisms

**TABLE 2.3** Examples of Natural Product Lead Compounds and Marketed Drugs Derived from Them

Natural Product Lead Compound	Marketed Drug
 <p>Echinocandin B (a fungal metabolite)</p>	 <p>Anidulafungin (antifungal)</p>
 <p>Epothilone B (from bacterial fermentation)</p>	 <p>Ixabepilone (anti-cancer)</p>



with which they coevolved, rather than those involved in human disease, and the reverse is true of synthetic drugs.<sup>[25]</sup> According to this hypothesis, natural products should be important molecules to combat microorganisms or aberrant (tumor) cell growth, but they should not be expected to be effective for other diseases, such as central nervous system (CNS) or cardiovascular diseases. However, genomes and biological pathways can be conserved across a variety of organisms. Furthermore, evolution over billions of years has produced these natural products to bind to specific regions in targets, and these binding regions can be very similar in targets for human disease as well as in microorganisms.

Because natural products often have the ability to cross biological barriers and penetrate cells, they often have desirable pharmacokinetic properties, which makes them good starting points for lead discovery. In fact, several structural neighbors of active natural products were shown to retain the same activity as the natural product.<sup>[26]</sup> One measure of the potential oral bioavailability of a compound is a set of guidelines called the *Rule of 5* (see Section 2.1.2.3.2). About 60% of the 126,140 natural products in the *Dictionary of Natural Products* had no violations of these guidelines, and many natural products remain bioavailable despite violating these rules.<sup>[27]</sup> This supports natural products as being an important source of lead compounds.

Frequently, screening of natural products has been done on semipurified extracts of sources such as plant materials, marine organisms, or fermentation broths. A significant challenge in screening of natural products in this way is that when activity is found, there is still considerable work to be done to isolate the active component and determine its structure. When HTS of chemical libraries started, such slower, more tedious screening methods were often put aside. However, because of the earlier success with natural product screening, the natural product approach has begun to return to the drug discovery process.

**2.1.2.3.1.2. Medicinal Chemistry Collections and Other “Handcrafted” Compounds** Many large, established pharmaceutical companies have been synthesizing compounds in one-at-a-time fashion for decades as part of their overall drug discovery efforts. In most cases, these institutions have had long-standing compound inventory management systems, such that samples of compounds prepared many years ago are still available for screening. One advantage of using these compounds for screening is that they are frequently close analogs of compounds that progressed substantially through the drug discovery process and thus have a reasonable probability of possessing biological activity and drug-like properties. One disadvantage, though, is that these compounds may be structurally biased toward the limited proteins that these companies have targeted over the years. Large companies may possess up to several million compounds in their corporate compound collections; however, most companies have substantially trimmed their collections

used for screening, leaving only compounds that have good drug-like properties for lead discovery (see Section 2.1.2.3.2).

Another source of handcrafted compounds is samples from academic or nonpharmaceutical synthetic laboratories. Some businesses have been established to purchase such samples and market them to drug discovery organizations.

**2.1.2.3.1.3. High-Throughput Organic Synthesis** To provide the large number of compounds needed to feed ultrahigh throughput screening operations, enormous efforts during the 1990s turned toward developing methods for high-throughput organic synthesis (HTOS). HTOS had its origins in the techniques of *solid-phase synthesis* (synthesis carried out on a polymer support, which makes removal of excess reagents and by-products from the desired product easier), and many drug discovery organizations established internal HTOS groups to supply compounds for screening using solid-phase chemistry. Millions of compounds were synthesized for HTS campaigns using these HTOS methods. The synthesis of large numbers of related compounds has now declined substantially in favor of smaller sets,<sup>[28]</sup> and this evolution has been accompanied by a dramatic shift of emphasis from solid-phase methods back to solution-phase chemistry. One approach taken to create more diversity in chemical libraries called *diversity-oriented synthesis*, the synthesis of numerous diverse scaffolds from a common intermediate, has had limited success.<sup>[29]</sup> Below we briefly review key aspects of the HTOS approach of the 1990s and early 2000s and its relationship to HTS during these years, because some of the lessons learned during this period serve as key concepts in the present practices of lead discovery.

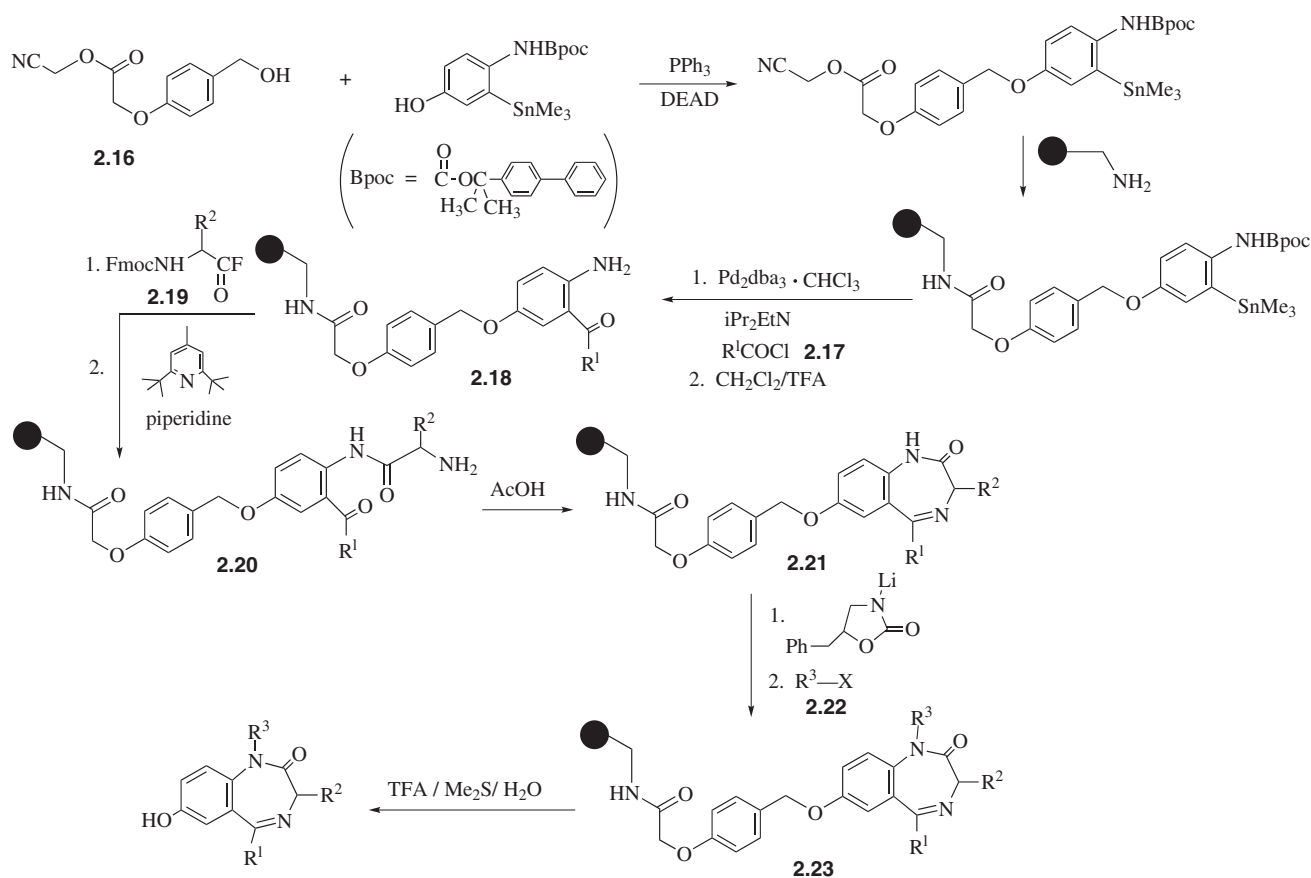
**2.1.2.3.1.3.1. Solid-Phase Library Synthesis** The most widely practiced methods in the early application of HTOS centered on the simultaneous synthesis of large collections (*libraries*) of compounds using solid-phase synthesis techniques. The synthesis of large numbers of compounds generally relied on a *combinatorial* strategy, that is, the practice of combining each member of one set of building blocks (i.e., reactants) with each member of one or more additional sets of building blocks (see examples below).<sup>[30]</sup> The beginnings of combinatorial chemistry are attributed to Furka,<sup>[31]</sup> with applications in peptide synthesis by Geysen and coworkers<sup>[32]</sup> and by Houghten.<sup>[33]</sup> These initial efforts in peptide library synthesis were followed by the synthesis of peptoids by Zuckermann and coworkers<sup>[34]</sup> and of small molecule nonpeptide libraries by Ellman and coworkers<sup>[35]</sup> and Terrett and coworkers.<sup>[36]</sup>

The efficiency of HTOS in producing large numbers of compounds relies, among other factors, on the ability to conduct reactions on multiple different (albeit often related) reactants in parallel. Solid-phase synthesis<sup>[37]</sup> is carried out by covalently attaching the starting material to a polymeric solid support and conducting a sequence of reactions while the corresponding intermediates and product remain attached to the solid phase, ultimately followed by a cleavage step to release the product into solution. Classically, functionalized

polystyrene beads (polystyrene resin) were used as the solid support, although many other polymeric materials have since been developed expressly for the purpose of increasing the versatility of the solid-phase methodology. To minimize unreacted starting material, excess reagents are usually used, which are then easily removed along with any solution-phase by-products by filtration and repeated washing of the solid-phase material. This type of reaction workup is well suited to parallel processing and automation, accounting for its initial broad implementation for synthesis of large libraries. Somewhat less well advertised during the early hype of solid-phase combinatorial chemistry was the fact that side reactions can and do occur during solid-phase synthesis just as they do in solution, and the resulting polymer-bound side products are retained as impurities throughout the solid-phase process. Monitoring reactions on solid phase is not as straightforward as it is for solution-phase reactions; it requires either specialized methods such as Fourier transform infrared spectroscopy or separate cleavage of an aliquot of a polymer-bound intermediate to release it into solution so it can be analyzed by conventional methods such as thin-layer chromatography or high-performance liquid chromatography (HPLC). Nevertheless, since the early days of solid-phase peptide synthesis (the Merrifield synthesis<sup>[38]</sup>) carried out through sequential amide

couplings and amine deprotections, a remarkably wide variety of reactions have been adapted to solid-phase methods.<sup>[39]</sup>

An early example of using solid-phase methodology to synthesize a nonpeptide library was the preparation of benzodiazepines as shown in **Scheme 2.2**.<sup>[40]</sup> Key reactions on solid phase include a Stille coupling to form ketone **2.18**, an amide coupling followed by an *N*-deprotection to form aminoketone **2.20** (note that by-products from Fmoc cleavage are soluble and thus readily removed), acid-promoted intramolecular imine formation to give polymer-bound benzodiazepine **2.21**, and an *N*-alkylation to form the polymer-bound version (**2.23**) of the final product. The *p*-alkoxybenzyl linker **2.16** serves two purposes: (1) the *p*-alkoxy substituent promotes the release of the final product from the polymer under acid conditions and (2) it acts as a spacer, moving the sites of the reactions in the synthetic sequence away from the surface of the resin to avoid steric hindrance to reaction and to facilitate access to the reaction sites by reactants in solution. In this solid-phase synthesis, there are three *diversity elements* ( $R^1$ ,  $R^2$ , and  $R^3$ ), which are correspondingly introduced by three sets of building blocks (also known as *monomers*), namely, a set of acid chlorides **2.17**, a set of Fmoc-protected amino acids **2.19**, and a set of alkylating agents **2.22**. The theoretical



**SCHEME 2.2** Solid-phase synthesis of a library of 7-hydroxybenzodiazepines

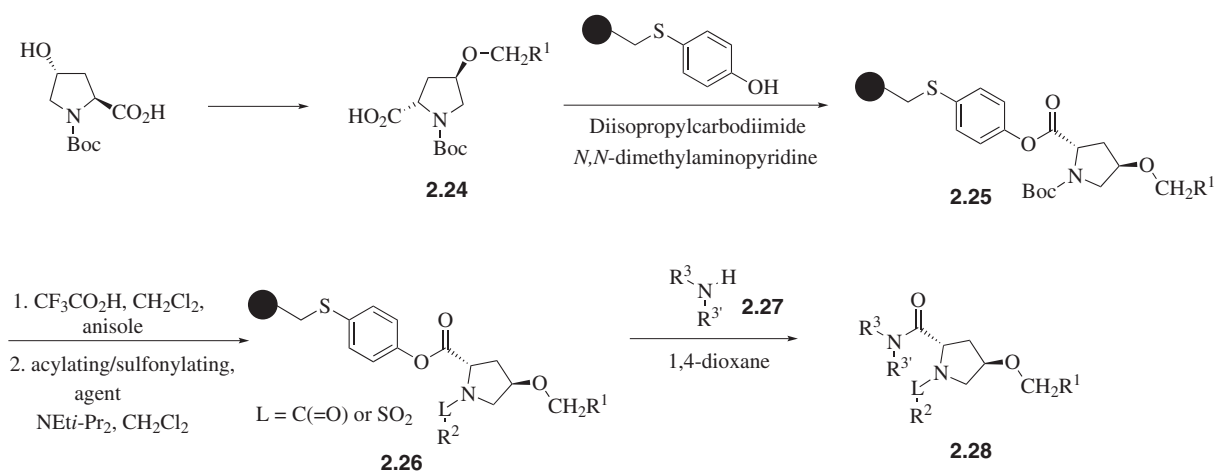
number of products equals the *product* of the number of each type of building block used; for example, 10 of each type of building block in **Scheme 2.2** would theoretically afford 1000 ( $10 \times 10 \times 10$ ) final products. Alternatively, 10  $R^1$  building blocks, 20  $R^2$  building blocks, and 50  $R^3$  building blocks would theoretically afford 10,000 products ( $10 \times 20 \times 50$ ). This comparison underscores the combinatorial power of combinatorial chemistry (in the above examples, a total of 30 monomers ( $10 + 10 + 10$ ) leads to 1000 different products, whereas adding only 50 monomers leads to an additional 9000 products!). It should be noted that all final products from **Scheme 2.2** have a hydroxyl substituent on the benzo portion of the benzodiazepine; this is an artifact that was required for linkage to the solid phase via spacer **2.16**. Accordingly, the products of this work are technically a library of 7-hydroxybenzodiazepines.

The efficiencies inherent in conducting many reactions simultaneously in separate reaction vessels (termed in *parallel*<sup>[41]</sup>) on solid phase include efficient use of time, simplified workups (filtration and washing), and no need to perform chromatography, recrystallization, or distillation of intermediates (not because the intermediates are necessarily highly pure, but because these techniques are not applicable to polymer-bound intermediates). Since it is generally not practical to obtain and critically assess NMR spectra or elemental analysis data on so many final compounds, these steps are usually bypassed in favor of HPLC and MS as the sole methods for final compound analysis.

As an example, the chemistry in **Scheme 2.3** was used to synthesize over 17,000 discrete compounds in parallel.<sup>[42]</sup> First, multiple Boc-4-alkoxyproline derivatives **2.24** were prepared in solution using a modified Williamson reaction at the 4-hydroxyl group, and the products were then coupled to polymer-bound phenolic hydroxyl groups to give polymer-bound activated esters **2.25**. A test for free phenolic hydroxyl groups on the polymer using  $\text{FeCl}_3$ /

pyridine qualitatively showed that most of the free sites had been acylated, and the gain in resin weight was consistent with this conclusion. Acid-mediated cleavage of the Boc protecting group of **2.25** followed by functionalization of the resulting secondary amine with diverse reagents gave diverse resin-bound products **2.26**. In this library synthesis, the primary and secondary amines (**2.27**) that provide the final diversity element also cleave the products from the solid phase via reaction with the activated ester linkage to result in product amides **2.28** in solution. The final products need to be separated from the excess amine reactants. This can be accomplished by filtering the reaction mixtures through diatomaceous earth (Celite<sup>®</sup>) impregnated with aqueous acid, effectively sequestering the excess basic amines (**2.27**) onto the diatomaceous earth while the neutral library products (**2.28**) pass through with the filtrate. This procedure demonstrates the feasibility of performing solution phase workups in a parallel fashion, foreshadowing the ultimate emergence of *solution-phase parallel synthesis* as the dominant HTOS method (next section).

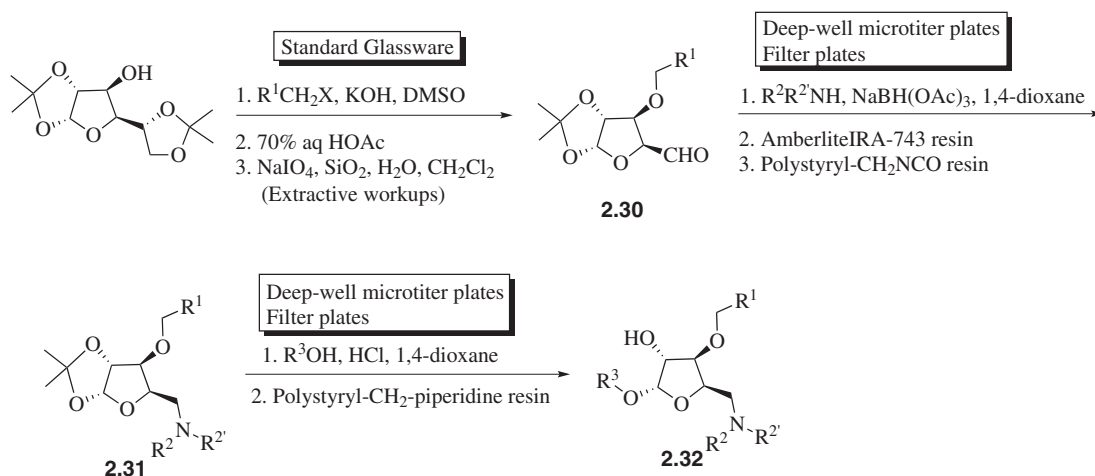
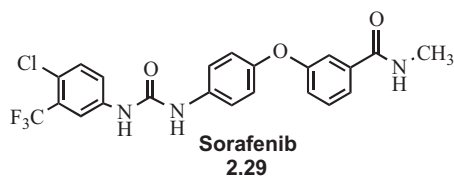
The foregoing library synthesis is an example of *parallel synthesis*. In contrast, a special variant of solid-phase combinatorial synthesis called *mix and split synthesis* (also known as *split and pool synthesis*) should be mentioned.<sup>[43]</sup> This technique is applicable to making very large libraries ( $10^4$ – $10^6$  compounds) as a collection of polymer beads, each containing, in principle, one library member, i.e., one bead, one compound. An important consideration is that for the one bead, one compound result to hold, each synthetic step must proceed reproducibly with very high conversion, even higher than in the synthesis of discrete compounds, to a single product.<sup>[44]</sup> Each bead carries only about 100–500 pmol of product, and special methods must be employed to determine which product is on a given bead. For simple compounds, mass spectrometric methods can be used,<sup>[45]</sup> but this is not applicable if the library



**SCHEME 2.3** Solid-phase synthesis of a library of 4-alkoxyproline derivatives

contains many thousands or millions of members that may not be pure or are isomeric with other library members. In that case, encoding methods need to be utilized. Although the structure of the actual compound might not be directly elucidated, the structure of certain tag molecules attached to the polymer that encode the structure can be determined.<sup>[46]</sup> One important approach that involves the attachment of unique arrays of readily analyzable, chemically inert, small molecule tags to each bead in a split synthesis was reported by Still and coworkers.<sup>[47]</sup> In this method, groups of tags are attached to a bead at each combinatorial step in a split synthesis, which create a record of the building blocks used in that step. At the end of the synthesis, the tags are removed and analyzed, which decodes the structure of the compound attached to that bead. Ideal encoding tags must survive organic synthesis conditions, not interfere with screening assays, be readily decoded without ambiguity, and encode large numbers of compounds; the test compound and the encoding tag must be able to be packed into a very small volume.

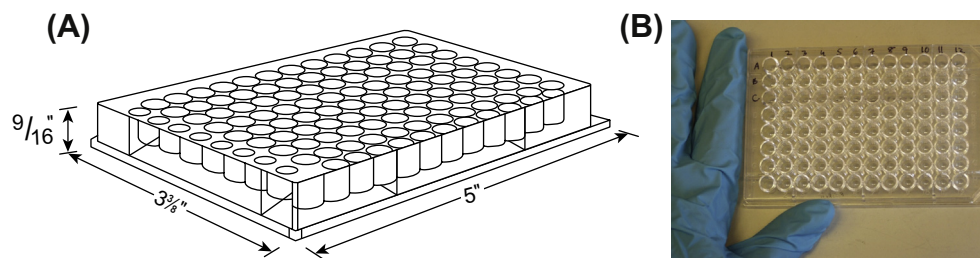
Although combinatorial chemistry was a common approach for about 15 years (from the late 1980s to the early 2000s), only one new de novo drug is believed to have resulted from this massive effort, namely, the anti-tumor drug sorafenib (**2.29**, Nexavar).<sup>[48]</sup> As will be discussed in more detail in Section 2.1.2.3.1.3.3, since about 2003–2005, solid-phase methods have been much less frequently used for HTOS than the solution-phase methods described in the next section.



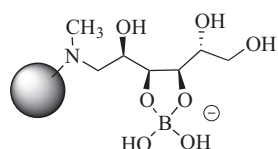
**SCHEME 2.4** Solution-phase synthesis of a library of furanose derivatives

**2.1.2.3.1.3.2. Solution-Phase Library Synthesis** Parallel library synthesis of up to a few thousand compounds at a time can frequently be carried out entirely by solution-phase parallel methods<sup>[49]</sup>; Scheme 2.4 summarizes the methods used to prepare a several thousand-member library in solution phase.<sup>[50]</sup> This library is derived from D-glucose, so it could be characterized as being derived from a natural product. In the first step, the free hydroxyl group of diacetone D-glucose is alkylated with different alkyl halides to form a series of ethers varied at R<sup>1</sup>. These intermediates are then selectively hydrolyzed (aq. HOAc) to the corresponding 1,2-diols, which are oxidatively cleaved with periodate to form aldehydes **2.30**. In this solution-phase library example, the subsequent reactions are run in parallel in microtiter plates (Figure 2.2), which facilitates convenient tracking of the individual reactions using plate positions in place of physical labels on reaction flasks. Thus, each aldehyde (**2.30**) is added to multiple wells of a microtiter plate and treated with different secondary amines under reductive amination conditions (NaBH(OAc)<sub>3</sub>) to give aminomethyl derivatives **2.31**. Workup can be accomplished sequentially using two different *solid-phase scavenger resins* (a polymer-supported molecule that can react with excess reagents in solution, thereby removing them from solution), followed by filtration. Thus, after completion of the reductive amination reactions, the mixtures are first treated with Amberlite IRA743 resin to scavenge borate anion (derived from NaBH(OAc)<sub>3</sub>). This scavenging agent contains polymer-bound *N*-methylglucosamine, which chelates with borate anion and is highly effective for removing borate from solution (Figure 2.3).<sup>[51]</sup> The Amberlite scavenger resin is removed by filtration using a 96-well filter plate (Figure 2.4; you can use an eight-channel pipettor to transfer contents of the microtiter plate eight wells at a time to the filter plate, which has various sorbents or filters, collecting the filtrate in another microtiter plate). The filtrates are treated with a polystyrene-bound isocyanate, which reacts with the excess secondary amine used in each

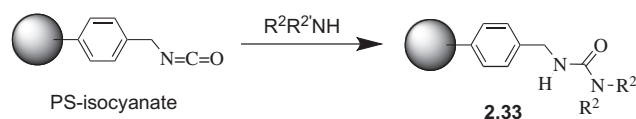




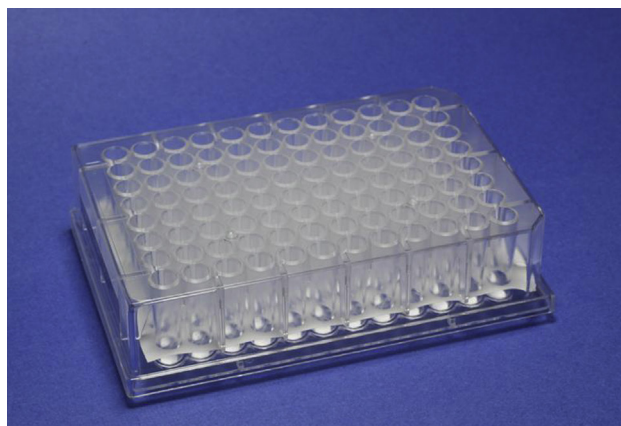
**FIGURE 2.2** (A) Schematic of a typical 96-well microtiter plate. (Reprinted with permission from Custom Biogenic Systems (<http://www.biomedical-marketing.com/CBS/MicrotiterCRacks.html>).) (B) Picture of a 96-well microtiter plate taken by Jeffrey M. Vinocur, 4/21/06, published on Wikipedia Commons ([http://commons.wikimedia.org/wiki/File:Microtiter\\_plate.JPG](http://commons.wikimedia.org/wiki/File:Microtiter_plate.JPG))



**FIGURE 2.3** Product of polymer-bound *N*-methylglucosamine with borate anion



**SCHEME 2.5** Use of a solid-phase scavenger in solution-phase synthesis. In this example, a polymer-bound isocyanate is used to scavenge excess primary or secondary amine from a solution by forming the corresponding polymer-bound urea.



**FIGURE 2.4** Image of 96-well filter plates. Reprinted with permission from Norgen Biotek Corp.

reaction, to form polymer-bound urea **2.33** (Scheme 2.5), effectively removing the amine from solution. The mixtures are again filtered (filter plate) to remove the polymeric scavenger. In the preceding step, 1,4-dioxane (freezing point 12 °C) is used as the reaction and rinse solvent. After the second filtration, the filtrates are frozen on Dry Ice, and the solvents are removed by sublimation under vacuum (called *lyophilization*). Introduction of a third point of diversity is effected by treatment of products **2.31** with an alcohol in the presence of hydrogen chloride to form hydroxyl ethers **2.32**, followed by evaporation of volatile components under vacuum. The resulting residues are dissolved in 1,4-dioxane/THF and treated with polystyrene-bound piperidine to remove residual HCl; omitting removal of residual HCl leads to poor stability of the products to storage and moisture. Finally, the products are frozen and lyophilized to afford library products as residues in the wells of the 96-well plates. These compounds often are then purified by reverse-phase liquid

chromatography. It is important to point out that for each step in the sequence, it is necessary to first evaluate a number of conditions to identify those conditions that give the highest purity of products across a number of representative building blocks. Therefore, although library production is rapid once the conditions are worked out, the myriad of process development trials must be factored in when assessing the overall efficiency gained by parallel synthesis.

Many of the techniques illustrated in the above example have gained considerable use in the parallel synthesis of smaller libraries as well, many of which may have only one or two points of diversity. Use of two points of diversity can reasonably support the synthesis of a library containing more than a 1000 compounds, for example, a 20 × 96 array (1920 compounds). When large libraries of analogs are needed, developmental work is often done in-house; then the library production can be outsourced to lower the cost of generating the library and to free up the time of the in-house chemist for new design and developmental studies.

**2.1.2.3.1.3.3. Evolution of HTOS** The use of solid-phase methods to synthesize large combinatorial libraries was in widespread practice during the 1990s and the early 2000s, but is currently not favored. Although obtaining large numbers of compounds for HTS was the initial driver for the technology, some investigators began to question whether the effort to collect and analyze HTS data on thousands, much less tens of thousands or millions, of compounds that are necessarily related by virtue of their common method of synthesis was truly an efficient use of resources. The structural diversity is limited in many cases not only by the fundamental chemistry used to prepare a library but also by the fact that diversity in commercially available building blocks did not always translate to a high level of diversity in the corresponding



substituents of the final products. This is because the building blocks that were *successfully* incorporated into final products were more frequently those with simpler, less reactive functionality (like substituted phenyl compared to a heterocycle). Furthermore, the large numbers of compounds generated usually precluded individual purification and weighing of final products; therefore, the screening samples were usually of only approximate purity and concentration. Moreover, although the incorporation of three or more diversity elements in a library contributed greatly to combinatorial power and the number of compounds in the library, this also tended to yield compounds of molecular weight (MW) higher than that of most orally active drugs (see Section 2.1.2.3.2). Because of this observation, several groups began to define what properties a compound should possess to make it drug-like or lead-like. Among the several properties considered, MW less than about 500 Da and CLog $P$  (a term related to lipophilicity of the compound; see Sections 2.2.5.2.2 and 2.2.5.2.3) less than 5 emerged as central criteria. Many of the libraries most amenable to large-scale synthesis by solid-phase combinatorial methods did not meet either of these criteria for a significant proportion of library members. For example, consider a library with a scaffold having a MW of 149 (see Scheme 2.4, 2.32, where R<sup>1</sup>CH<sub>2</sub>, R<sup>2</sup>, R<sup>2'</sup>, R<sup>3</sup> all = H) and incorporating three diversity elements; the average contribution of the diversity elements to the MW of a given product must be <117 to keep the MW of the product molecule under 500.

Consequently, several significant changes to the common practice of HTOS began to evolve, including the synthesis of fewer compounds per library and the decision to purify final products, for example, by preparative reverse-phase HPLC. Once a final purification step was incorporated into the process, there developed a tendency to work on a larger scale to make up for mechanical purification losses. The prospect of obtaining a larger quantity of each purified product inspired a desire to store some of the material as dry solid, enabling more extensive follow-up studies in case interesting biological activity could be identified. It then became difficult for solid-phase synthesis to be applicable to these new objectives because the reaction scale is limited by the amount of solid support that could fit into reaction vessels of manageable size.

Although solid-phase methodology offers a strong advantage when the objective is to synthesize very large numbers of unpurified compounds in limited quantities and with a distinct tendency toward high MWs, the disadvantages of each of these characteristics led to the decline of its use in lead discovery. The synthesis of smaller libraries of compounds in larger quantities is usually well accommodated by parallel solution-phase chemistry, and its inherently greater flexibility with respect to scale, variety of reaction conditions accommodated, ability to analyze reaction mixtures, and option to purify intermediates made it

the method of choice for high-throughput synthesis of lead discovery libraries. Moreover, solution-phase parallel synthesis using scavenger resins, disposable reaction vessels, specialized liquid transfer methods, automated purification, and other tools is applicable not only to the preparation of libraries for lead discovery but also to the downstream medicinal chemistry objectives, for example, during hit-to-lead (see Section 2.1.2.3.5) or lead modification activities (Section 2.2.).<sup>[52]</sup> In these latter contexts, it is most common to prepare libraries of only about 10–200 compounds.

### 2.1.2.3.2. Drug-Like, Lead-Like, and Other Desirable Properties of Compounds for Screening

As discussed in Chapter 1, lead compounds often require optimization with respect to not only their activity against a biological target but also a number of pharmacokinetic parameters, including ADME characteristics. If these properties could be predicted from the structure of a compound, then they could be taken into account at an early stage, even including the design and selection of compounds for a screening collection. Lipinski<sup>[53]</sup> proposed *the Rule of 5* as a guide to predict oral bioavailability. On the basis of a large database of known drugs, the Rule of 5 states that it is highly likely (>90% probability) that compounds with two or more of the following characteristics will have *poor* oral absorption and/or distribution properties:

- The MW is >500
- The log  $P$  is >5 (log  $P$  is a measure of the lipophilicity, discussed in Section 2.2.5.2.2); conveniently, the value can be predicted computationally, as described in Section 2.2.5.2.3.
- There are more than 5 H-bond donors (expressed as the sum of OH and NH groups)
- There are more than 10 H-bond acceptors (expressed as the sum of N and O atoms)

In 2006, it was determined that 885 (74%) of all small molecule drugs pass the Rule of 5; 159 of the orally administered small molecules fail at least one of the Rule of 5 parameters.<sup>[54]</sup>

Gleeson compared results of about 10 ADME assays with many compounds from GlaxoSmithKline and found that MW (<400), log  $P$  (<4), and ionization state are the most important molecular properties that affect ADME parameters.<sup>[55]</sup> To get a drug across the blood–brain barrier, the upper limits really should be 3 H-bond donors and 6 H-bond acceptors.<sup>[56]</sup> Some drugs, for example, certain antibiotics, antifungal drugs, vitamins, and cardiac glycosides, have active transporters to carry them across membranes, so lipophilicity is less relevant in those cases. Because active transporters allow molecules with poor physicochemical parameters to cross membranes readily, it is possible to design compounds with groups that are recognized by

one of these transporters to aid in their bioavailability.<sup>[57]</sup> In the absence of a transporter, it is useful to understand what properties of a molecule promote good oral bioavailability (oral bioavailability is usually expressed as a percent; 100% bioavailable means that all the administered drug reached the systemic blood circulation).

In contrast to the Rule of 5, Veber and coworkers<sup>[58]</sup> measured the oral bioavailability of 1100 drug candidates and found that reduced molecular flexibility, as determined by the number of rotatable bonds (10 or fewer), and low polar surface area (PSA, the sum of surfaces of polar atoms, usually oxygens, nitrogens, and attached hydrogens, in a molecule) favored good oral bioavailability. The three-dimensional (3D)-PSA can be readily calculated and is referred to as the *topological polar surface area* (TPSA).<sup>[59]</sup> Veber and coworkers determined that a  $PSA \leq 140 \text{ \AA}^2$  (for intestinal absorption;  $\leq 70 \text{ \AA}^2$  to cross the blood–brain barrier<sup>[60]</sup>) or a total hydrogen bond count ( $\leq$  a total of 12 donors and acceptors) are important predictors of good oral bioavailability independent of MW. Both the number of rotatable bonds and hydrogen bond count tend to increase with MW, which may explain Lipinski's first rule. Lower PSA was found to correlate better with increased membrane permeation than did higher lipophilicity. The charge on molecules at physiological pH affects the PSA range that is important.<sup>[61]</sup> The fraction of anions with  $>10\%$  F (F is the symbol for oral bioavailability) falls from 85% when the PSA is  $\leq 75 \text{ \AA}^2$  to 56% when  $75 \text{ \AA}^2 < PSA < 150 \text{ \AA}^2$ . For neutral, zwitterionic, and cationic compounds that pass the Rule of 5, 55% have  $>10\%$  F, but for those that fail the Rule of 5, only 17% have  $>10\%$  F. A group at AstraZeneca found that two physicochemical properties unrelated to molecular size or lipophilicity, but related to molecular topology, namely, the fraction of the molecular framework ( $f_{MF}$ ) and the fraction of  $sp^3$ -hybridized carbon atoms ( $F_{sp^3}$ ) are important to ADME and toxicity.<sup>[62]</sup> The  $f_{MF}$  refers to the size of the molecule without side chains (the core ring structure) relative to its overall size (or the number of heavy atoms in the molecular framework divided by the total number of heavy atoms in the molecule)<sup>[63]</sup>;  $F_{sp^3}$  is the number of  $sp^3$ -hybridized carbon atoms divided by the total number of carbon atoms.<sup>[64]</sup> Aqueous solubility, Caco-2 permeability, plasma protein binding, human ether à go-go-related gene (hERG; see Section 2.1.2.3.5) potassium channel inhibition, and cytochrome P450 (CYP3A4) inhibition are all influenced by molecular topology, some favorably and others unfavorably by increased  $f_{MF}$  and  $F_{sp^3}$ . Important considerations for assessing potential oral bioavailability of compounds were assembled in the form of a road map for oral bioavailability with emphasis on absorption (permeability and solubility) and metabolism properties.<sup>[65]</sup> Analogously, a group at Pfizer used six physicochemical parameters to construct a drug likeness algorithm for CNS drugs and applied it to marketed CNS drugs, CNS candidate

compounds, and a diverse set of compounds.<sup>[66]</sup> This CNS multiparameter optimization algorithm showed that 74% of the marketed CNS drugs received a high score ( $\geq 4$  out of 6). Of the compounds with a score  $>5$ , 91–96% displayed high passive permeability into the CNS, low efflux liability (ejection from the CNS), favorable metabolic stability, and high cellular viability.

Compounds that meet the Lipinski or Veber criteria are frequently referred to as *drug-like molecules*. However, the physicochemical properties of marketed orally administered drugs are generally more conservative than these rules allow compared to nonorally administered or nonmarketed drugs, e.g., lower MW, fewer H-bond donors and acceptors, and rotatable bonds.<sup>[67]</sup> Over the years, certain physicochemical properties of oral drugs change and others do not. Up through 2003 (the time frame of the Veber study), mean values of lipophilicity, PSA, and H-bond donor count were the same, which implies that they are the most important properties of oral drugs; however, MW, numbers of O and N atoms, H-bond acceptors, rotatable bonds, and number of rings increased between 1983 and 2002 (13–29%).<sup>[68]</sup> Fewer than 5% of marketed oral drugs have more than 4 H-bond donors; only 2% have a combination of MW  $> 500$  and  $>3$  H-bond donors. The balance between polar and nonpolar properties seems to be quite important for oral drugs.

Ajay and coworkers proposed that *drug-likeness* is a possible inherent property of some molecules,<sup>[69]</sup> and this property could determine which molecules should be selected for screening. They used a set of one-dimensional and two-dimensional (2D) parameters in their computation and were able to predict correctly over 90% of the compounds in the Comprehensive Medicinal Chemistry (CMC) database.<sup>[70]</sup> Another computational approach to differentiate drug-like and nondrug-like molecules using a scoring scheme was developed,<sup>[71]</sup> which was able to classify correctly 83% of the compounds in the Available Chemicals Directory (ACD)<sup>[72]</sup> and 77% of the compounds in the World Drug Index.<sup>[73]</sup> A variety of other approaches have been taken to identify drug-like molecules.<sup>[74]</sup>

It is now a common practice to bias screening collections in favor of drug-like molecules, particularly when the ultimate objective is development of orally bioavailable drugs.<sup>[75]</sup> Teague and coworkers<sup>[76]</sup> have taken the concept a step further to describe *lead-like molecules*. These authors note that during lead optimization, an increase in MW by up to 200 Da and increase of  $CLog P$  by up to 4 units frequently occur. Therefore, in order for an optimized compound to stay within, or close to, drug-like parameters, a lead compound should have a MW of 100–350 Da and a  $CLog P$  value of 1–3, and the authors propose that screening collections should be more heavily populated with compounds possessing these lead-like properties. As already noted, in the parallel synthesis of compounds for screening libraries, the more

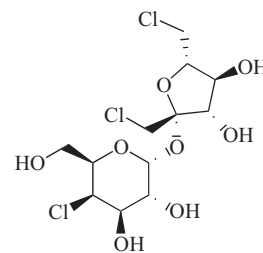
points of diversity, the greater the MW; therefore, there is always a balance between increasing diversity and MW.

Another approach to bias screening collections in favor of molecules likely to show biological activity is to consider *privileged structures*.<sup>[77]</sup> Evans and coworkers at Merck first introduced this term for certain molecular scaffolds that appear to be capable of binding to multiple receptor targets, and, consequently, with appropriate structure modifications, could exhibit multiple pharmacological activities.<sup>[78]</sup> This phenomenon was earlier mentioned by Ariëns and coworkers without referring to them as privileged structures.<sup>[79]</sup> The Merck group used benzodiazepines as a primary example of this phenomenon, because the benzodiazepine scaffold is found not only in antianxiety and anticonvulsant drugs that act through the  $\gamma$ -aminobutyric acid-activated ion channel but also in compounds that interact with opioid and cholecystokinin receptors. The latter two receptors are members of another major class of drug targets, the G-protein-coupled receptors (GPCRs; see Chapter 3, Section 3.1), which are quite distinct in their macromolecular structure from ion channels. Note that library synthesis around the benzodiazepine scaffold was the focus of [Scheme 2.2](#); the privileged structure concept formed the basis for this scaffold. The commonality of molecular features in a variety of drugs is apparent by the revelation that only 32 scaffolds describe half of all known drugs.<sup>[80]</sup> In recognizing a molecule containing a privileged structure, it is important to note that the privileged components frequently consist of two or three rings linked by single bonds or by ring fusion, which constitute a substantial part of the overall size of the compound; otherwise, the contribution of the privileged structure to the activity of the compound would be questionable.<sup>[81]</sup> Additional examples of privileged structures include indoles, purines, dihydropyridines, spiro-piperidines, benzimidazoles, benzofurans, and benzopyrans. Examples of indoles, dihydropyridines, and benzimidazoles that interact with diverse biological targets are shown in [Figure 2.5](#).

Analogous to the small number of scaffolds found in a large number of drugs, there are a small number of moieties that account for a large majority of the side chains found in drugs.<sup>[82]</sup> The average number of side chains per molecule is four. If the carbonyl side chain is ignored, then 73% of the side chains in drugs are from the top 20 most common side chains. Accordingly, efforts to incorporate privileged scaffolds and privileged side chains are common considerations when identifying compounds to add to a screening collection.

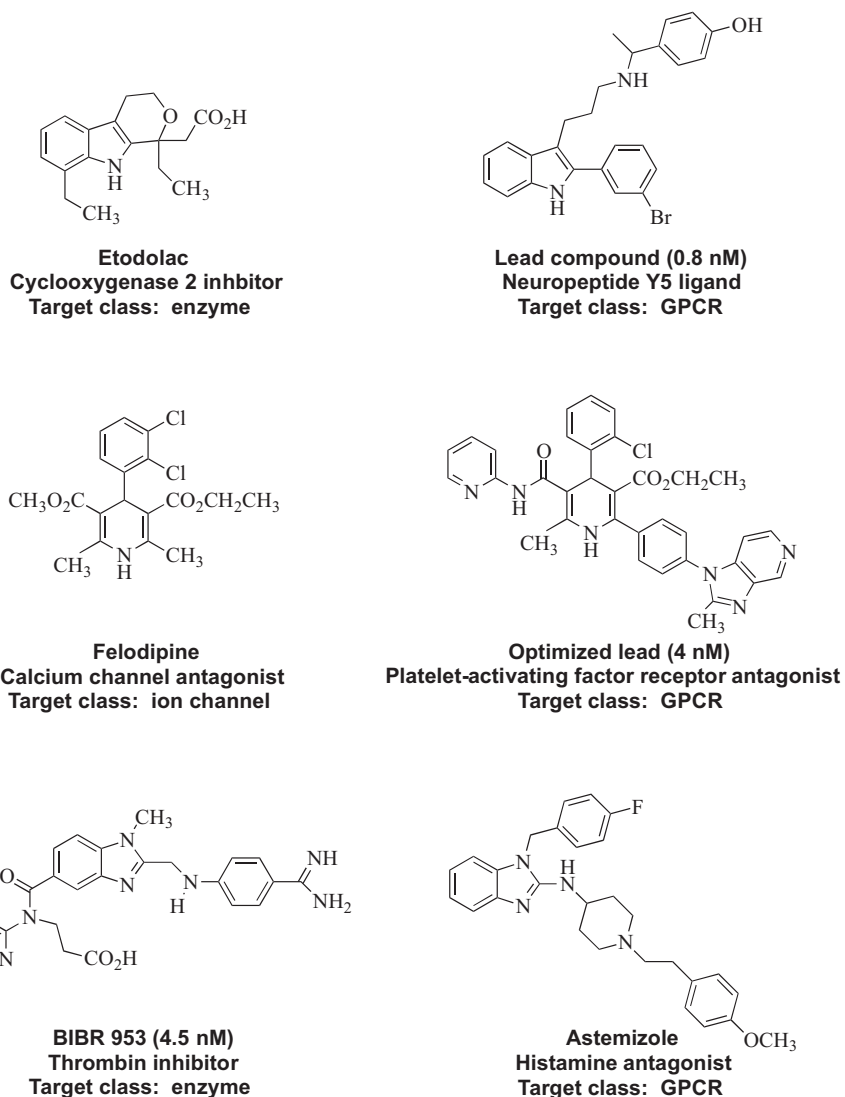
An additional filter for many screening collections is to remove (or at least flag) compounds containing functional groups viewed as undesirable in a drug, usually because these groups have been found, or can be hypothesized, to have undesirable effects *in vivo*. These so-called *toxicophoric* groups can generally be classified into one of two different types: (1) those functional groups that may

have undesirable effects by their own right and (2) those functional groups that can be converted by metabolic processes to moieties that may have undesirable effects<sup>[83]</sup>; representative examples of each type of toxicophore are shown in [Tables 2.4 and 2.5](#), respectively. One approach to identify toxicophoric groups is illustrated in a study by Kazius et al.<sup>[84]</sup> The investigators took a *chemoinformatics* approach by computationally comparing the structures of over 4000 compounds, about half of which were mutagenic and half of which were nonmutagenic, and ascertaining which substructures were prominent in the mutagenic set. It should be noted, however, that the presence of a so-called toxicophoric group in a molecule does not imply that the substance is necessarily unsafe for human consumption. For example, an alkyl halide is frequently considered to be a toxicophoric group (because it is an electrophile), yet this has not prevented the FDA-approved human consumption of the popular artificial sweetener sucralose ([2.34](#), Splenda).






**Sucralose**  
**2.34**

A further approach to optimize a screening collection is to minimize the number of compounds that will ultimately prove to be *false positives* across many different high-throughput screens. False positives are compounds that appear to be *hits* (compounds that have a level of activity that the researcher believes is sufficient to pursue further), but upon additional investigation are found to be inactive against the target. Shoichet and coworkers<sup>[85]</sup> and others<sup>[86]</sup> have shown that a frequent source of false positives is the formation of colloidal aggregates of compounds in the screening mixture. Such aggregates frequently interact with targets in a nonspecific manner, and hence the component compounds have been characterized as *promiscuous binders*. The activity observed for such compounds may be counteracted by the addition of a detergent in the screening solution, which provides the basis for a straightforward method to identify this source of false positives at an early stage. Aggregate formation in the screening medium can be detected using an NMR assay.<sup>[87]</sup> Of course, another source of false positives is impurities in the samples, supporting the necessity to screen pure samples whenever practical.



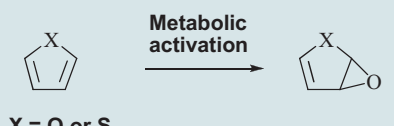
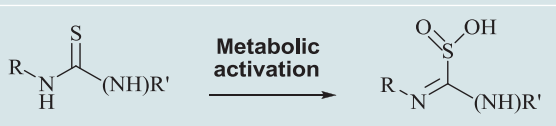
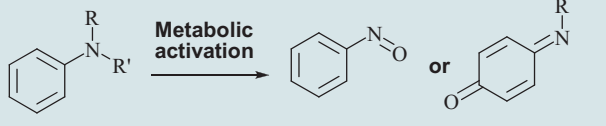
**FIGURE 2.5** Pairs of compounds containing a privileged structure (indole, dihydropyridine, or benzimidazole) and binding to diverse target classes

**TABLE 2.4** Representative Groups Viewed as Toxicophoric Because of the Reactivity

Toxicophoric Group	Rationale
 EWG = electron withdrawing group, e.g., carbonyl, cyano, etc.	Michael acceptor; electrophilic group that can alkylate biological nucleophiles, for example, cysteine -SH
	Epoxide; electrophilic group that can alkylate biological nucleophiles
	Imidazole; can chelate metals, for example, iron in heme proteins such as cytochrome P450 enzymes



**TABLE 2.5** Representative Groups Viewed as Toxicophoric Because They May be Metabolized to Undesirable Moieties

Toxicophoric Group	Rationale
 <p>Metabolic activation of a furan or thiophene ring (X = O or S) leads to an epoxide intermediate.</p> <p><b>X = O or S</b></p>	Furans and thiophenes; tend to be metabolized to electrophilic epoxides
 <p>Metabolic activation of a thioamide or thiourea (R-NH-C(=S)-NH-R') leads to an imine intermediate (R-N=C(SOH)-NH-R').</p>	Thioamides and thioureas; tend to be metabolized to electrophilic imines
 <p>Metabolic activation of an aniline derivative (R-NH-Ph) leads to either a nitroso derivative (R-N=O) or a quinone derivative (R-N=O-C6H4=O).</p>	Anilines; tend to be metabolized to electrophilic nitroso or quinone derivatives

### 2.1.2.3.3. Random Screening

Given a high-throughput assay and access to an appropriate collection of compounds, how do you select which compounds to screen? In the absence of known drugs and other compounds with desired activity or structural information about the target, a random screen is the most common approach. *Random screening* in its simplest form involves no intellectualization; compounds are tested in the bioassay without regard to their structures. However, as discussed above, it is desirable to maximize lead-like and drug-like molecules in your random screening library.

Prior to 1935 (the discovery of sulfa drugs), this was essentially the only approach; today this method is still a very important approach to discover leads, particularly because it is now possible to screen such large numbers of compounds rapidly with (ultra) high-throughput screens.

An example of a random screen of synthetic and natural compounds is the “war on cancer” declared by Congress and the National Cancer Institute (NCI) in the early 1970s. Any new compound submitted was screened in a mouse tumor bioassay. Few new anticancer drugs resulted from that screen, but many known anticancer drugs also did not show activity in the screen used, so a new set of screens was devised, which gave more consistent results. In the 1940s and 1950s, a random screen of soil samples by various pharmaceutical companies in search of new antibiotics was undertaken. However, in this case, not only were numerous leads uncovered, but two important antibiotics, streptomycin and the tetracyclines, were found. Screening of microbial broths, particular strains of *Streptomyces*, was a common random screen methodology prior to 1980; it is now regaining importance in the search for new leads.

In recent years, attempts have been made to increase the efficiency of random screening by using computational methods to select a representative subset of compounds from

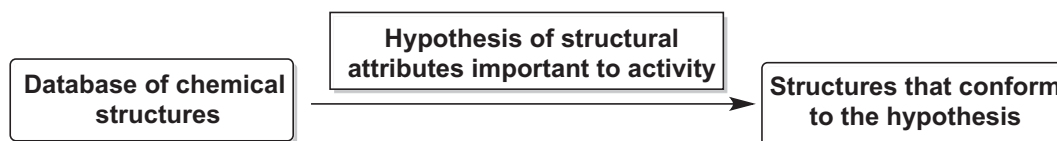
a compound collection. This usually entails grouping (*clustering*) compounds that are structurally similar, and then choosing a few members from each cluster for screening. Methods for quantifying the similarity between molecules are discussed in the next section. If hits are identified in the initial screen, then further screening of other compounds that are structurally similar to the initial hits, a technique known as *hit-directed nearest neighbor screening*, is often productive for identification of additional hits.<sup>[88]</sup> This subsequent round of screening is a special case of targeted (or focused) screening, which is also discussed in the next section.

Another technique proposed to increase efficiency has been to screen mixtures of compounds, generated either as a result of the synthetic method<sup>[89]</sup> or by intentionally mixing pure compounds. However, as noted above (Section 2.1.2.3.2), many screening collections contain a significant number of compounds that tend to aggregate, leading to false positives. When mixtures of compounds are used, these aggregates can also mask the identification of compounds that are active when screened alone.<sup>[90]</sup> Therefore, the likelihood of a high rate of *false negatives* (an active compound that does not show activity) is also considerable when screening mixtures.

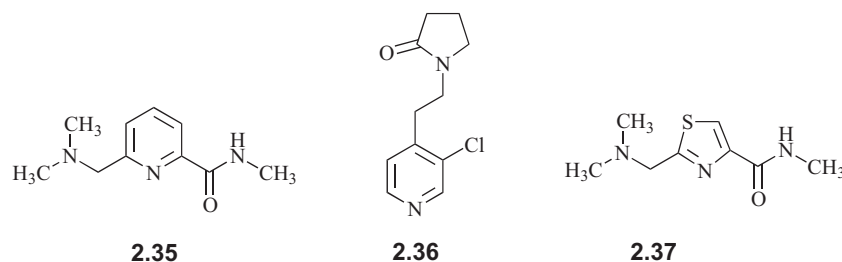
### 2.1.2.3.4. Targeted (or Focused) Screening, Virtual Screening, and Computational Methods in Lead Discovery

Information about one or more known ligands for a target or about the structure of the target itself may be used to narrow a large screening collection to a smaller set of compounds that may be more likely to hit the target, thereby saving screening resources. The screen is then regarded as *targeted or focused*, in contrast to the random approach discussed in the previous section. The most common computational method for selection of the compounds is called *virtual*





**SCHEME 2.6** The process of virtual screening to identify compounds that conform to a hypothesis specifying properties (that are discernible from a compound's structure) that are required for activity



**FIGURE 2.6** Hypothetical example illustrating that substructure search (e.g., using pyridine as the search query) may not retrieve the most structurally similar compounds in a compound collection

*screening*, which involves the rapid in silico (by computer) assessment of large libraries of chemical structures to identify those structures that most likely bind to a drug target, such as a protein receptor or enzyme.<sup>[91]</sup> The goal is to identify new scaffolds, especially ones that may be in the existing collection. In its most general form, virtual screening can be described by the process shown in [Scheme 2.6](#).

Two components are needed: (1) a database of structures in a form that can be computationally analyzed for structural attributes and (2) a hypothesis or model of the structural attributes that are important for activity, for example, the hypothesis that structural similarities to a known active ligand should yield similarly active compounds or a hypothesis of the shape and charge density of a binding pocket that defines what features a complementary ligand structure should have (see discussions below).

**2.1.2.3.4.1. Virtual Screening Database** A key criterion for the structures that will be virtually screened is that physical samples of the compounds will be available if they are identified as compounds of interest by the virtual screen. This criterion would generally argue for the inclusion of compounds in an organization's corporate collection as well as compounds that are offered for sale commercially. Saario et al.<sup>[92]</sup> used two databases of structures in a virtual screen for fatty acid amide hydrolase inhibitors, one representing compounds in the LeadQuest collection offered commercially by Tripos (St Louis, MO, USA) and another screening collection offered commercially by Maybridge (Cornwall, England, UK). Databases that compile compounds from the catalogs of many vendors include the commercial Accelrys ACD with almost 4 million chemicals or the free "ZINC" database with almost 19 million commercially available compounds, 4 million lead-like compounds, and over 13 million drug-like compounds.<sup>[93]</sup> The virtual screening

database might also contain other compounds that could be considered reasonably accessible. For example, compounds believed to be easily synthesizable might be included in a virtual screening database. Such compounds may range from those that have been previously synthesized in the organization, and for which detailed procedures are available, to members of combinatorial libraries for which general synthetic procedures have been reported in the literature.<sup>[94]</sup> Toward the latter set of compounds, the reviews by Dolle et al.<sup>[95]</sup> provide detailed lists of published combinatorial library syntheses and a rich source for generation of such virtual compounds. Among published library syntheses are many that target privileged structures, which should be of particular interest.

**2.1.2.3.4.2. Virtual Screening Hypothesis** Many methods have been developed to describe properties against which a compound might be assessed to estimate the likelihood that it will interact with a given target. For example, if a known ligand for the target exists, then searching a database of compounds for structures that are similar to the known ligand is a reasonable approach. Although this, in principle, could be accomplished by a seasoned medicinal chemist by visual inspection, to do so for many thousands of compounds is clearly impractical; moreover, computers can sometimes discern similarity features that the naked eye would miss. One simple and easily understood method is searching for other molecules that contain a *substructure* (part of the total structure) in common with the active molecule. To understand the shortcomings of this approach, assume that the structures in [Figure 2.6](#) are three among thousands of compounds in a screening collection. A substructure search of the corresponding database using the structure of pyridine as the query would retrieve compounds **2.35** and **2.36**, but not **2.37**. Yet inspection of the structures might reasonably suggest that

2.35 and 2.37 would be more likely to share similar biological activity than 2.35 and 2.36. Therefore, computational methods more sophisticated than the substructure approach have been developed. These methods can be generally categorized according to the following models:

1. 2D similarity models
2. 3D-QSAR models
3. Structure-based pharmacophore models and computational docking

Each of the above methods is discussed in this section. Companies such as Tripos (St Louis, MO), Accelrys (San Diego, CA), and Chemical Computing Group (Montreal, Quebec, Canada) have specialized in developing sophisticated computational chemistry software to assist in the use of such models.

*Two-dimensional similarity models* (2D because they mirror the similarity between flat structures, as drawn on paper) for assessing similarities between two molecules typically rely on defining a set of so-called 2D *descriptors* and then assessing how a given molecule conforms to each descriptor. Many types of descriptors have been developed and applied,<sup>[96]</sup> but simple examples include properties such as “contains an NC(O)O fragment”, “contains a sulfur-containing heterocycle”, or “contains a group IIIA element”, as part of a set containing, say, 80–150 descriptors. Frequently, the descriptors are formulated in such a way that, for a given molecule, the assessment results are an answer of either “yes” or “no” or, in computer language, “1” or “0”. Then, for a set of descriptors listed in a given order, a corresponding sequence of 1’s and 0’s can be generated that defines a *fingerprint* for that molecule. The concept is illustrated in Table 2.6, where the fingerprints are shown for two compounds (1 and 2) as defined by a set of 18 descriptors A through R (again, in most real-life cases, the number of descriptors used is considerably larger). The extent to which the two compounds share the same property is noted according to how often both molecules have a value of 1 for a given descriptor (gray areas). The *Tanimoto coefficient*  $T$  is a frequently used index to quantify similarity<sup>[97]</sup> and is defined as:

$$T = \frac{N_{11}}{n - N_{00}}$$

where  $N_{11}$  is the number of descriptors for which both values are 1,  $N_{00}$  is the number of descriptors for which both values are 0, and  $n$  is the total number of descriptors used. For the example in Table 2.6,  $T = 7/(18-4) = 0.50$  (50% structurally similar). A computer can quickly determine 2D fingerprints for each structure in a database, and from these, quickly determine the level of similarity to a query molecule, for example, a known active ligand, to help select a set of compounds to be assayed in a real screen. This is a widely used similarity search method in the early stages of lead discovery when there are limited SAR and target structure data available. It allows the identification of a few actives that can be used in more sophisticated 3D virtual screening approaches, such as pharmacophore mapping and docking.<sup>[98]</sup> *Extended-connectivity fingerprints*, topological fingerprints designed to capture molecular features relevant to drug activity, were developed for substructure and similarity searching and are available in the commercial software called Pipeline Pilot (Accelrys, San Diego, CA, USA).<sup>[99]</sup>

It bears repeating that the assumption that compounds with similar structures are likely to have similar biological activity must be exercised with some caution. It has been shown that only 30% of compounds considered to be at least 85% structurally similar ( $T \geq 0.85$ ) to an active compound will themselves have the same activity.<sup>[100]</sup> Adding just one methylene group to a 4-hydroxypiperidine analog changed it from a poor binder of the chemokine receptor CCR1 into a potent binder.<sup>[101]</sup> Nevertheless, given an active compound, the use of these methods to select additional active compounds from a data set is still far superior to random selection.

*Three-dimensional quantitative structure–activity relationships* (3D-QSARs) quantitative structure–activity relationship (QSAR) analysis is a method that permits correlations between different series of molecular structures and their biological function at a particular target. Various QSAR methods, which have served as valuable predictive tools for the design of drug candidates, have been developed over more than a 100 years. Classical 2D QSAR methods considered only 2D structures and are discussed later in this chapter (Section 2.2.4.2) as part of a historical overview of computational methods in lead modification.

**TABLE 2.6** Illustration of Data Used to Calculate Tanimoto Similarity

Descriptor	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Compound 1	0	1	1	0	0	1	1	0	0	0	1	1	0	1	0	1	1	0
Compound 2	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	1	0	1

A set of descriptors is assigned a value of either 1 or 0, depending on whether that descriptor applies or does not apply, respectively, to the molecule. The string of 0s and 1s found for each molecule defines its descriptor-based fingerprint.

*Three-dimensional QSAR* was a natural extension of 2D-QSAR and was first proposed in the 1980s. The general approach of 3D-QSAR is to select a group of molecules, each of which has been assayed for a particular activity; align the 3D conformations of the molecules according to some predetermined orientation rules; calculate a set of spatially dependent parameters for each molecule determined in the receptor space surrounding the aligned series; derive a function that relates each molecule's spatial parameters to its respective biological property; and establish self-consistency and predictability of the derived function. There are a variety of computer-based methods that have been used to correlate molecular structure with receptor binding, and, therefore, activity. Some are mentioned here; others are cited in the General References at the end of the chapter.

Crippen and coworkers<sup>[102,103]</sup> devised a linear free energy model, termed the *distance geometry* approach, for calculating QSAR from receptor binding data. The distances between various atoms in the molecule, compiled into a table called the distance matrix, define the conformation of the molecule. Rotations about single bonds change the molecular conformation and, therefore, these distances; consequently, an upper and lower distance limit is set on each distance. Experimentally determined free energies of binding of a series of compounds to the receptor are used with the distance matrix of each molecule in a computerized method to deduce possible binding sites in terms of geometry and chemical character of the site, thereby defining a 3D pharmacophore. This approach requires considerably more computational effort and adjustable parameters, but it is thought to give good results on more difficult data sets.

The distance geometry approach was extended by Sheridan et al.<sup>[104]</sup> to treat two or more molecules as a single ensemble. The ensemble approach to distance geometry can be used to find a common pharmacophore for a receptor with unknown structure from a small set of biologically active molecules. A virtual screen of this type of model was used to identify inhibitors of human immunodeficiency virus type 1 integrase (HIV-1 IN) as potential anti-AIDS drugs.<sup>[105]</sup> HIV-1 IN mediates the integration of HIV-1 DNA into host chromosomal targets and is essential for effective viral replication. From a known inhibitor of HIV-1 IN, a pharmacophore hypothesis was proposed. On the basis of this hypothesis, a 3D search of the NCI database of compounds was performed, which produced 267 structures that matched the pharmacophore; 60 of these were tested against HIV-1 IN, and 19 were found to be active. The relevance of the proposed pharmacophore was tested using a small 3D validation database of known HIV-1 IN inhibitors, which had no overlap with the group of compounds found in the initial search. This new 3D search supported the existence of the postulated pharmacophore and also suggested a possible second pharmacophore. Using the second pharmacophore

in another 3D search of the NCI database, 10 novel, structurally diverse, HIV-1 IN inhibitors were found.

Hopfinger<sup>[106]</sup> developed a set of computational procedures termed molecular shape analysis for the determination of the active conformations and, thereby, molecular shapes during receptor binding. Common pairwise overlap steric volumes calculated from low-energy conformations of molecules are used to obtain 3D molecular shape descriptors, which can be treated quantitatively and used with other physicochemical parameter descriptors.

Two other descriptors for substructure representation, the atom pair<sup>[107]</sup> and the topological torsion,<sup>[108]</sup> have been described by Venkataraghavan and coworkers. These descriptors characterize molecules in fundamental ways that are useful for the selection of potentially active compounds from hundreds of thousands of structures in a database. The atom pair method can select compounds from diverse structural classes that have atoms within the entire molecule similar to those of a particular active structure. The topological torsion descriptor is complementary to the atom pair descriptor, and focuses on a local environment of a molecule for comparison with active structures.

One of the most widely used computer-based 3D-QSAR methodologies, developed by Cramer and coworkers,<sup>[109]</sup> is termed *Comparative Molecular Field Analysis* (CoMFA).<sup>[110]</sup> In this method, the molecule–receptor interaction is represented by the steric and electrostatic fields exerted by each molecule. A series of active compounds are identified, and 3D structural models are constructed. These structures are superimposed on one another and placed within a regular 3D grid. A probe atom, with its own energetic values, is placed at lattice points on the grid, where it is used to calculate the steric and electrostatic potentials between itself and each of the superimposed structures. At each lattice point, one steric value and one electrostatic value are saved for each inhibitor in the series. The results are represented as a 3D contour map in which contours of various colors represent locations on the structure where lower or higher steric or electrostatic interactions would increase binding. However, because simple steric and electrostatic fields are unlikely to represent a complete description of a drug–receptor interaction, alternative and modified forms have been proposed.<sup>[111]</sup> Because it is assumed that the molecules bind with similar orientations in the receptor, which may not necessarily be the case, correct alignments are almost impossible, particularly for compounds with a large number of rotatable bonds, which limits the applicability of CoMFA. *Comparative Molecular Similarity Indices Analysis* (CoMSIA) is similar to CoMFA in the aspect of atom probing.<sup>[112]</sup> However, CoMSIA uses a different potential function; therefore, not only steric and electrostatic, but also hydrophobic, fields can be calculated. Different from CoMFA and CoMSIA, which are ligand-based approaches, *Comparative Binding Energy Analysis*

(COMBINE) takes advantage of structural data of ligand–receptor complexes and applies them to a 3D-QSAR paradigm.<sup>[113]</sup> This technique is based on the hypothesis that the free energy of binding can be correlated with a subset of energy components calculated from the structures of receptors and ligands in bound and unbound forms.

CoMFA, CoMSIA, and COMBINE require molecular alignment prior to the calculation of descriptors. If the structures of the macromolecules are known, the alignment can be guided by the binding conformations of receptor–ligand complexes (COMBINE is only useful when the protein structure is known). Otherwise, when CoMFA and CoMSIA are employed for 3D-QSAR analyses, purely computational alignment has to be postulated to superimpose all ligand structures in space. The 3D descriptors and their corresponding 3D-QSAR models, therefore, are related to molecular rotation and translation. In the past two decades, much effort has been made to develop 3D-QSAR models that are independent of subjective alignment rules. Several methods have been proposed, including *Comparative Molecular Moment Analysis* (CoMMA),<sup>[114]</sup> EVA,<sup>[115]</sup> Weighted Holistic Invariant Molecular (WHIM) descriptors,<sup>[116]</sup> and Grid-independent descriptors (GRIND).<sup>[117]</sup> CoMMA, EVA, and WHIM do not give an intuitively 3D display of the resulting models. In contrast to CoMMA, EVA, and WHIM, GRIND was devised to overcome the problem of interpretability that is common to alignment-independent descriptors.

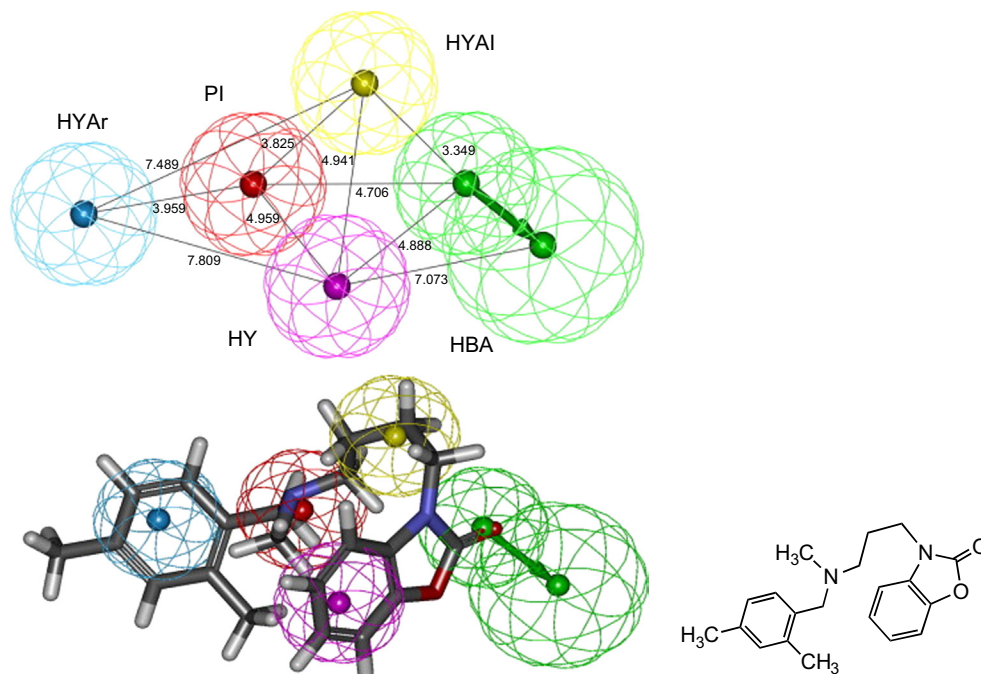
Another popular 3D-QSAR method is an approach known as *topomer similarity searching*.<sup>[118]</sup> A *topomer* is a *molecular descriptor* (any property, measured or calculated, of a molecule, such as melting point or PSA) that focuses on the shape of a molecule, as represented by a combination of the shapes of different fragments of the molecule. This is a method to search 3D molecular structures in conventional structural databases and compare them as sets of fragments (or topomers) by superimposition of their fragmentation bonds, which allows comparison of the molecules by their pharmacophoric features. This method is an improvement over the 2D-QSAR similarity metric, Tanimoto coefficient<sup>[119]</sup> (see Section 2.1.2.3.4.2). CoMFA and topomer similarity technologies were merged by Cramer<sup>[120]</sup> into a 3D-QSAR methodology called *Topomer CoMFA*. In this approach, structures in a series are each broken into two or more fragments at central acyclic single bonds while removing core fragments that are common to the series. The method requires a common scaffold among the molecules in the series, but the commonality can be as simple as a key sp<sup>3</sup> carbon. Topomer 3D models are constructed for each fragment, and a set of steric and electrostatic fields is generated for each topomer set. The Topomer CoMFA results can be used to query virtual libraries already composed of topomer structures to identify fragment structures having increased potency. The advantages of this method are that

it minimizes the preparation needed for 3D-QSAR analysis, automates the creation of models for predicting biological activity, which are created much quicker than traditional CoMFA, and is more user friendly than traditional CoMFA analysis. Other popular 3D methods that focus on shape or volume similarity between molecules include *Surflex-Sim* and *Flex-S*.<sup>[121]</sup> *Hologram QSAR* uses molecular holograms and partial least squares (PLS) analysis to generate a fragment-based SAR but does not require the alignment of molecules, which allows for automated analysis of very large data sets.<sup>[122]</sup>

A *pharmacophore* model is a 3D representation of the regions of ligands that are believed to be responsible for interactions with the biological target. An example<sup>[123]</sup> is shown in Figure 2.7. When such a model is derived from known ligands for the target, it is called a *ligand-based pharmacophore model* (in contrast to a *structure-based pharmacophore model*, which is based on knowledge of the receptor structure, see below). Computer software, such as Catalyst (Accelrys, Inc., San Diego, CA, USA), DISCO (Tripos, Inc., St Louis, MO, USA), LigandScout (Inteligand, Wien, Austria), Phase (Schrodinger, Portland, OR, USA), or MOE (Chemical Computing Group, Montreal, Canada), is used to generate one or more models, given the structures of a collection of known ligands, often including ligands with diverse structures. This technique is called *receptor mapping*.<sup>[124]</sup> It is founded on the premise that receptor topography is complementary to that of drugs, but in this case the structure of the lock is deduced from the shape of the keys that fit it. A variety of receptor mapping techniques have been described. An approach termed *steric mapping*<sup>[125]</sup> uses molecular graphics to combine the volumes of compounds known to bind to the target receptor. This composite volume generates a receptor-excluded volume map, which defines that region of the binding site available for binding by drug analogs and, therefore, not occupied by the receptor itself. The same procedure is, then, carried out for similar molecules that are inactive. The composite volume is inspected for regions of volume overlap common to all the inactive analogs. These are the receptor-essential regions, sites required by the receptor itself and unavailable for occupancy by ligands. Any other molecule that overlaps with these regions should be inactive. This approach has been termed an Active Analog Approach.<sup>[126]</sup> A pharmacophore-based virtual screen, then, would involve the identification of compounds possessing the appropriate pharmacophore that filled the receptor-excluded regions and that avoided the receptor-essential regions.

Some types of targets, specifically soluble enzymes such as kinases and proteases, are amenable to crystallization and hence to structure determination by X-ray crystallography. The structure of a protein that is similar to one for which the crystal (or NMR) structure has been determined can sometimes be deduced with a reasonable degree of





**FIGURE 2.7** Example of a computer-generated pharmacophore model. From Laurini et al *Bioorg. Med. Chem. Lett.* 2010 (Ref. 111)

accuracy using a process known as *homology modeling*.<sup>[127]</sup> This technique involves the alignment of the amino acid sequence of the protein of unknown structure onto the corresponding positions in the experimentally determined structure (the template structure), followed by energy minimization. Naturally occurring homologous proteins have similar protein structure, and 3D protein structures are evolutionarily more conserved than expected because of sequence conservation.<sup>[128]</sup> The sequence alignment onto the template structure can be used to produce a structural model of the target. Membrane-associated proteins, such as GPCRs and ion channels, are much less amenable to crystallization, but steady progress has been made to experimentally determine the structures of these targets as well.<sup>[129]</sup> When the structure of a biological target, or preferably of the target complexed with a ligand, is available, then the information can be used to develop models for use in virtual screening. A model that has been derived based on an experimentally determined *target structure* is referred to as a *structure-based* model (see below).

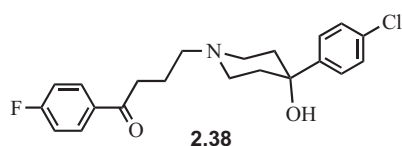
When the X-ray crystallographic structure or the NMR solution structure of a target receptor is known, an analysis of the active site can be performed to facilitate drug design. Two popular approaches, GRID<sup>[130]</sup> and multiple copy simultaneous search (MCSS),<sup>[131]</sup> have been employed to identify the energetically favored sites in the active site for ligand binding. Goodford's program *GRID* uses a grid force field that includes a very good description of hydrogen bonding.<sup>[132]</sup> Because the energetics and shape complementarity of a drug-receptor complex are vital to its stability,

this method simultaneously displays the energy contour surfaces and the macromolecular structure on the computer graphics system. This allows both the energy and shape to be considered together when considering the design of molecules that have an optimal fit in the receptor, and it determines probable interaction sites between various functional groups on the ligand and the enzyme surface. MCSS uses numerous small chemical group copies simultaneously, each transparent to the others (i.e., noninteracting) but each subjected to full force minimization in the receptor. This approach provides exhaustive information of the possible binding sites and orientations for small chemical groups in a known protein structure.

After an analysis of the active site, sophisticated computational programs such as DOCK,<sup>[133]</sup> AutoDock,<sup>[134]</sup> FlexX,<sup>[135]</sup> Glide,<sup>[136]</sup> GOLD,<sup>[137]</sup> Surflex,<sup>[138]</sup> and MolDock,<sup>[139]</sup> are capable of *docking* (inserting, on the computer, an unbound ligand into the binding site of the target) ligands into the biological target.<sup>[140]</sup> The ability for the software to independently dock a ligand in a way that corresponds closely to an experimentally determined structure for the same complex serves as validation of the docking method used. It is important to recognize, however, that the lowest energy structure of the ligand does not have to be the one that binds to the receptor; that is, the *bioactive conformation* can be a higher energy conformation of the molecule.<sup>[141]</sup> Currently available software programs are capable of carrying out virtual docking experiments across large numbers of compounds, including multiple conformations of each molecule, in a short period of time.



The algorithm DOCK, which was originally restricted to rigid ligands and receptors, was modified<sup>[142]</sup> for flexible ligands by representing the ligand as a small set of rigid fragments. This approach focuses on molecular shapes, and like most docking methods, DOCK ranks molecules based on polar, steric, hydrophobic, and solvation terms. Starting with a high-resolution structure (X-ray crystal structure or NMR spectral structure) of the receptor *with a bound ligand*, the ligand is removed from the binding site on the graphic display; then DOCK fills the binding site with sets of overlapping spheres, where a set of sphere centers serve as the negative image of the binding site. When a crystal structure of a receptor is available, but without a ligand bound, DOCK characterizes the entire surface of the receptor with regard to grooves that could potentially form target-binding sites, which are filled with the overlapping spheres. Next, DOCK matches structures of putative ligands to the image of the receptor on the basis of a comparison of internal distances and searches 3D databases of small molecules and ranks each candidate on the basis of the best orientations that can be found for a particular molecular conformation.<sup>[143]</sup> The drawbacks of this approach are the assumptions that binding is determined primarily by shape complementarity and that only small changes in the shape of the receptor occur upon ligand binding. An important advantage, though, is that this method is not limited to docking of known ligands. A library of molecular shapes can be scanned to determine which shapes best fit a particular receptor-binding site. In fact, DOCK was used to identify the antipsychotic drug haloperidol (**2.38**, Haldol)<sup>[144]</sup> and fullerenes<sup>[145]</sup> as potential inhibitors of HIV-1 protease.



An example of the application of DOCK to the identification of new leads for the ubiquitous GPCRs, which have been an important focus of the pharmaceutical industry for many years, came from Shoichet and coworkers.<sup>[146]</sup> Because few crystal structures of GPCRs are available, lead discovery efforts have largely been ligand based. Crystal structures of the  $\beta_2$ -adrenergic receptor with two partial inverse agonists (see Chapter 3, Section 3.2.3) bound<sup>[147]</sup> allowed a structure-based approach. About 1 million commercially available lead-like molecules were docked into this structure; the 25 top hits were tested, and 25% of them were active inverse agonists of this receptor. Impressively, one of them had a  $K_i$  of 9 nM, the most efficacious inverse agonist for the  $\beta_2$ -adrenergic receptor to that date. A crystal structure of this high-potency molecule bound to the  $\beta_2$ -adrenergic receptor<sup>[148]</sup> revealed the same overall fold

observed for the previous crystal structures and exhibited the same binding conformation predicted by DOCK.

Given the wide variety of models and methods that are available for virtual screening, it is of both theoretical and practical interest to understand which ones are most effective. Somewhat surprisingly, systematic comparisons have frequently led to the conclusions that 2D similarity methods have similar effectiveness to 3D similarity methods, and that ligand-based pharmacophore models are frequently as effective as structure-based models.<sup>[149]</sup>

A comparison of hits obtained by HTS and by virtual screening of the same compound library against the protein cruzain, a target for Chagas disease, revealed the strengths and weaknesses of the two approaches and demonstrated the power of integrating the two.<sup>[150]</sup> Experiments by both approaches with a 198,000-member library led to 146 well-behaved hits, representing five different chemotypes. Two of the chemotypes were discovered through HTS alone, two came from the virtual screen, and one resulted from a combination of the two methods. Testing of these compounds gave potencies ranging from 65 nM to 6  $\mu$ M. Integration of these two approaches can be very beneficial to identify and prioritize hits.

Another dramatically different computational approach for lead identification is *structure-based de novo design*. This approach is used to design, from scratch (i.e., de novo), a bioactive compound that does not exist in your known compound libraries. It is often applied when the 3D structure of the target protein or a specific set of pharmacophores is known. It therefore provides an opportunity to explore and utilize other areas of chemical space that have not been explored by your known compound libraries. *De novo* design approaches were first proposed in 1980s and can primarily be divided into structure-based approaches and ligand-based approaches. In the former case, the 3D structure of the receptor is known or can be modeled by homology modeling (vide supra), and the de novo design is based on the structural information of the target. In the latter case, the structure of the target is unknown, and the pharmacophore information of ligands is used to guide the design of new structures. Five different approaches have been developed for receptor-based de novo design depending on the method of structure sampling<sup>[151]</sup>: (1) planar structure fitting, (2) atom or fragment growing, (3) fragment linking, (4) target protein lattice-based sampling, (5) and molecular dynamics simulation-based sampling. Sophisticated computational programs for this approach include LUDI, LEGEND, and BOMB (Biochemical and Organic Model Builder).

The program *LUDI*<sup>[152]</sup> uses statistical analyses of nonbonded contacts in crystal packings of organic molecules to establish a set of rules that define the possible nonbonded contacts between proteins and ligands. Using these rules it also can search databases to find structures that fit a particular binding site in a protein based not on shape, as in DOCK,

but on physicochemical properties, such as hydrogen bonding, ionic interactions, and hydrophobic interactions.

Some software programs grow molecules from atoms added into receptor structures. LEGEND<sup>[153]</sup> grows molecules by adding atoms one by one up to the specified molecular size using random numbers and force field energy calculations. BOMB,<sup>[154]</sup> another de novo ligand-growing lead discovery program, grows molecules by adding substituents to a core that is isolated or that has been placed in a binding site. BOMB has a library of about 700 possible substituents, including the most common heterocycles and substituted phenyl groups. The core may be as simple as ammonia or benzene or it may represent a polycyclic framework of a lead series. The user specifies a template, which includes the core, the topology, and the substituents, and all molecules corresponding to the template are grown. The template is generally selected because it conforms to the geometry of the target-binding site and because of synthetic ease. A thorough conformational search is performed for each molecule that is grown, and the dihedral angles for the conformers are optimized along with their position and orientation in the binding site. The resultant lowest energy conformer is evaluated with a docking-like scoring function to predict activity.

New developments in the field of de novo design have led to the generation of scaffold hopping (see Section 2.2.6.3) and fragment hopping. Different from structure-based de novo design, which aims to generate entire ligands, scaffold hopping is an attempt to replace only the core motif of a known ligand, while conserving key substituents.<sup>[155]</sup> This approach can lead to the identification of compounds that have similar biological activities, but totally different scaffolds. A pharmacophore-driven de novo design strategy for fragment-based drug discovery (see Section 2.1.2.3.6) is fragment hopping.<sup>[156]</sup> The core of this approach is the derivation of the minimal pharmacophoric elements for each pharmacophore. The minimal pharmacophoric element can be an atom, a cluster of atoms, a virtual graph, or vectors. The new fragments that match the requirements of the minimal pharmacophoric elements are generated and hopped onto the corresponding position in the active site. After linking the fragments, new inhibitors with novel scaffolds can be generated. Key features for both ligand-binding affinity and isozyme selectivity (when there are multiple isozymes of the target protein) can be included in the definition of minimal pharmacophoric elements, which leads to the generation of new inhibitors with diverse scaffolds and greater isozyme selectivity.

Although the interaction of a drug with multiple protein targets generally leads to side effects, many diseases, such as CNS diseases, infectious diseases, and cancer, involve multiple proteins. In these cases, it would be desirable to have a drug that can interfere with more than a single target. A computational method for the design of small molecules

that bind to multiple desirable targets, in favor of proteins that could cause side effects, was developed; 800 ligand–target predictions were tested experimentally of which 75% were confirmed.<sup>[157]</sup>

Structure-based drug design has broader applications than just virtual screening for lead discovery and is discussed further in the context of lead modification in Section 2.2.6.

### 2.1.2.3.5. Hit-To-Lead Process

The *hit-to-lead* phase of the drug discovery process is the follow-up to HTS, where a *hit* is any compound that exhibits a level of activity that the researcher believes is worth pursuing further. Large-scale HTS campaigns generate enormous amounts of data that must be processed, analyzed, and ultimately acted upon if the program is to move forward. Because of this, certain activities need to be carried out to help avoid potential pitfalls and improve the chances that downstream efforts will ultimately result in a successful drug candidate, ideally within a reasonable time frame.<sup>[158]</sup> The main focus of such hit-to-lead efforts is not to identify the best compound, which normally takes place during the later lead optimization stage, but rather to provide data from the hits and related compounds that will support a decision to advance one or more series into the lead optimization stage. A central tenet of the hit-to-lead process is that identification of liabilities that are significant enough to disqualify a series of compounds for further work is of greatest value prior to lead optimization efforts. The precise activities undertaken during a hit-to-lead process may vary according to the organization carrying out the work. The following activities of a hit-to-lead phase are typical:

- *Confirmation of the structure, purity, and activity of the compound (hit confirmation).* Does the screening sample still contain the expected compound? Are there other compounds in the sample that might be responsible for the observed activity? It is useful to repeat the assay with a range of doses using freshly prepared solutions made from pure material that has been stored as a powder and for which purity and identity can be verified by NMR spectroscopy, MS, and HPLC. If dry pure sample is not available, it is often prudent to resynthesize or reisolate the compound.
- *Computational assessments.* Computational support can be applied to several aspects of hit-to-lead evaluations.<sup>[159]</sup> It is common to organize the hits into groups of similar compounds (*clusters*) in order to organize the information around structure classes. The finding that a number of structurally similar compounds possess similar biological activity lends credence to the data arising from any given hit within the cluster. By contrast, data for *singletons* (a compound that has no other similar structures among the hits) have no such substantiation, which increases the

need for gathering independent verifications of structure, purity, and activity at an early stage. Computation is also applied to calculating properties such as CLog *P* and PSA that are believed to correlate with drug-like properties and oral bioavailability (see Section 2.1.2.3.2). More sophisticated calculations that can be used to assess hits include predictions of solubility<sup>[160]</sup> and membrane permeability. Predicting these properties computationally can save time compared to determining them experimentally. Poor aqueous solubility affects results in biological assays and in absorption and distribution; the two most important descriptors to predict aqueous solubility are the aromatic proportion<sup>[161]</sup> of the molecule and the MW.<sup>[162]</sup>

- **Early ADME-tox assessments.** Measurement of the stability of hit compounds and close analogs by incubation with liver microsome preparations gives an early indication of the degree of metabolic stability in vivo. *In vitro* systems for measuring membrane permeability (e.g., with human epithelial (Caco-2) cells<sup>[163]</sup>) are also available and can provide an early indication of the likelihood that compounds will be absorbed from the gastrointestinal (GI) tract. Assessment of hits for inhibition of cytochrome P450 enzymes, important enzymes that metabolize drugs,<sup>[164]</sup> gives an early indication of potential drug–drug interactions.<sup>[165]</sup> Ways to reduce cytochrome P450 inhibition include lowering the lipophilicity of the molecule, adding steric hindrance, and adding an electron-withdrawing substituent (e.g., a halogen) to reduce the  $pK_a$ .<sup>[166]</sup> Assessment of hits for interactions with hERG<sup>[167]</sup> potassium ion channels gives an early indication of potential adverse cardiac toxicity; inhibition of the hERG channel is a major cause for compound attrition and withdrawal from the market.<sup>[168]</sup> While improving ADME-tox properties is frequently a major objective of the lead optimization phase, such early assessments can help in the prioritization of different series and further give an early indication of what parameters should be of concern during lead optimization.
- **Intellectual property assessments.** Patent searches are time consuming and thus difficult to conduct thoroughly on a large number of hits. Nevertheless, an early evaluation of whether the chemical space around a hit is very crowded or less so can be obtained by carrying out substructure searches across the Chemical Abstracts Registry File, noting how many of the retrieved publications are patent documents.
- **Early structure–activity relationship (SAR) assessments, synthetic accessibility, and ligand efficiency (LE).** Once the activity and identity of a hit have been verified, it is common practice to synthesize a number of close analogs of the hit for biological assessment. Such a set of analogs is frequently termed a *focused library* around the hit; the most efficient and desirable way to accomplish

this is by parallel synthesis (see Sections 2.1.2.3.1.3.1 and 2.1.2.3.1.3.2). It is helpful to observe a range of biological activities among the analogs, which lends confidence to the prospect of eventually increasing potency through structure modifications. This process also helps to prioritize a series on the basis of synthetic accessibility since, other factors being equal, those series that are more easily synthesized can generally proceed more rapidly through the lead optimization process.

While the natural inclination is to place the highest value on the most potent compounds, the concept of LE offers an interesting alternative perspective, one that takes into account not only potency but also MW, which we found (Section 2.1.2.3.2) might be related to oral bioavailability.<sup>[169]</sup> LE is defined (Eqn (2.1)) as the binding energy per ligand atom:

$$\text{Ligand efficiency} = \Delta G/N \quad (2.1)$$

where  $\Delta G = -RT(\ln K_d)$ , and  $N$  = the number of nonhydrogen atoms in the ligand. Thus, a small ligand with moderate potency could have a higher LE than a more potent, but significantly larger, molecule. Accordingly, LE is a way of normalizing potency at a target (pharmacodynamics) and molecular size (a contributor to pharmacokinetics), and is therefore useful for comparing compounds with a range of potencies and MWs. In refinements of the concept, the term  $\Delta G$  in the above equation can be substituted by the  $pK_i$  or  $pIC_{50}$  (where  $IC_{50}$  is the concentration that gives 50% inhibition) and  $N$  may be replaced by terms for CLog *P*, MW, or PSA to normalize potency against these other parameters that are critical to drug-likeness.<sup>[170]</sup> The LE should remain relatively constant during optimization if the scaffold is preserved and optimal substitutions are incorporated into the lead. In comparing the properties of a set of drugs with the leads from which they were derived, in general,  $pK_i(\text{drug}) \gg pK_i(\text{lead})$ , but the CLog *P* (drug) = CLog *P* (lead), resulting in LLE (drug)  $\gg$  LLE (lead),<sup>[171]</sup> where the LLE<sup>[172]</sup> is the *ligand lipophilicity efficiency* =  $pK_i - \text{CLog } P$ . One of the keys to success in a lead optimization program is the maintenance of low levels of lipophilicity as the MW inevitably increases. The LLE links the potency and lipophilicity to estimate drug-likeness of compounds. However, the LLE does not include the LE term; therefore, an alternative term can be used that stresses the importance of lipophilicity and LE, called LELP (*ligand efficiency and log P*), which is  $\log P/LE$ .<sup>[173]</sup> The higher the LELP, the less drug-like is the lead. The accepted lower limit of LE for a lead is 0.3, and lead-like compounds have  $-3 < \log P < 3$ ; therefore  $-10 < \text{LELP} < 10$  is an acceptable LELP range for leads. In general, the closer the LELP is to zero in the positive range, the better. If a good hit or lead has an  $LE > 0.4$  and  $0 < \log P < 3$ , then an LELP between 0 and 7.5 is an

excellent range. With regard to their impact on ADME, safety properties, and binding thermodynamics, both LLE and LELP are helpful in identifying higher quality compounds; however, LLE is not as useful as LELP with fragment-based hits (see section below).<sup>[174]</sup>

### 2.1.2.3.6. Fragment-based Lead Discovery

Despite several successes,<sup>[175]</sup> HTS has not yet completely fulfilled the original expectations of bringing medicines to the market rapidly,<sup>[176]</sup> because HTS has some inherent fundamental limitations. First, a typical HTS campaign utilizes approximately  $10^5$ – $10^6$  compounds, which is much less than the potential chemical diversity space, estimated to be about  $10^{60}$  molecules containing  $\leq 30$  nonhydrogen atoms.<sup>[177]</sup> Second, corporate libraries are filled with compounds that have drug-like rather than lead-like properties, i.e., having relatively high MWs (on average, 400 Da) and high lipophilicity,<sup>[178]</sup> which limit lead optimization efforts. Finally, for many targets, suitable lead molecules will be absent from the compound collections or the HTS hit rate will be very low.

The awareness of concepts such as lead-likeness<sup>[179]</sup> and drug-likeness<sup>[180]</sup> and their importance in the construction of compound collections should yield improved success rates in HTS-based lead discovery efforts.<sup>[181]</sup> Hann et al.<sup>[182]</sup> showed that poor ligand–receptor interactions increase exponentially with the size and complexity of the ligand, suggesting that the probability of small, simple molecules binding to the receptor, although with low affinity, is much higher than HTS-sized compounds. Indeed, LE (see Section 2.1.2.3.5) calculations<sup>[183]</sup> of HTS hit compounds show that the average contribution to binding per atom can be rather modest, suggesting that small molecules might have greater potential as starting points for lead optimization.

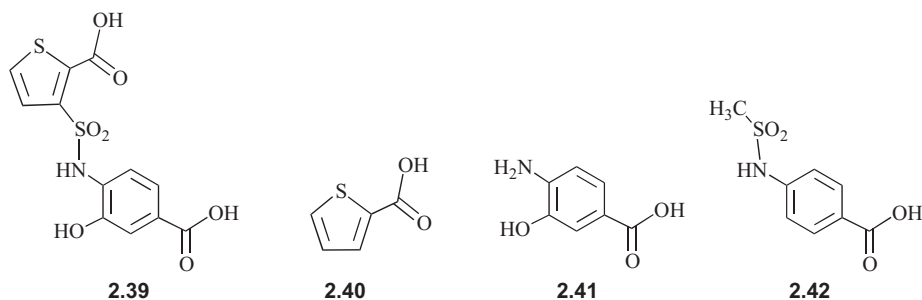
*Fragment-based lead discovery*<sup>[184]</sup> involves the screening of low-MW building blocks (*fragments*), followed by the application of various methods to increase potency. Typically, the focus is on ligand efficiencies of fragments, rather than potency, when prioritizing hits for follow-up. The interactions with the individual fragments are often rather weak since the small molecular structure usually offers only a small number of points of contact with the target. A significant rationale for the fragment-based discovery approach is that interactions with a biological target might be identified in an isolated fragment, whereas such interactions might have been obscured if the same fragment were part of a larger molecule containing structural elements that interfere with binding to the target. The molecular mass of these fragments is typically in the range of 150–300, having less functionality; fragments are expected to be much less potent (millimolar to 30  $\mu$ M potency range) than hits from HTS campaigns (30  $\mu$ M to nanomolar potency range). Because of the poor binding

affinities of fragments, standard assay methods generally cannot be used, as they are not sufficiently sensitive. Attempts made to utilize standard screening approaches with the fragments at high concentrations (millimolar rather than the typical micromolar concentrations commonly used for normal HTS) are usually unsuccessful. Therefore, one of the serious limitations of fragment-based methods is the requirement to implement sensitive biophysical techniques, such as NMR spectroscopy,<sup>[185]</sup> X-ray crystallography,<sup>[186]</sup> MS,<sup>[187]</sup> and surface plasmon resonance<sup>[188]</sup> to screen fragments because of their weak binding to the target.

Nonetheless, fragment-based screening offers a number of attractive features compared to HTS. First, the larger compounds typically found in HTS libraries are less able to adapt to a variety of binding sites; however, a high proportion of the atoms of a fragment directly interact with the receptor, which allows for optimal positioning in the binding site. Therefore, a hit fragment generally has a higher LE.<sup>[189]</sup> Second, the number of potential fragments with  $\leq 12$  nonhydrogen atoms (<160 Da) has been estimated to be about 14 million.<sup>[190]</sup> Therefore, the number of fragments that need to be screened is only in the range of hundreds to a few thousands, which still explores a much larger percentage of fragment chemical space (14 million) relative to the percentage of drug-like space ( $10^{60}$  compounds) that a million compounds screened in an HTS campaign explores. This also leads to much higher hit rates for fragment screens than HTS screens (in one report, 10–1000 times higher hit rates).<sup>[191]</sup> Furthermore, developing and maintaining a small set of fragments is easier than maintaining a large HTS library. Third, the subsequent structural optimization of a hit fragment has many more options and can result in a higher success rate for generating novel chemical structures. Finally, starting with a low molecular mass and low lipophilic fragment is likely to produce leads with small, simple structures, allowing for the typical molecular mass and, if necessary, lipophilicity (CLogP) increases during the lead optimization process.<sup>[192]</sup>

Intelligent construction of fragment screening collections is beneficial for more rapid lead discovery. One approach is to focus on fragments containing moieties that are frequently found in known drugs or other compounds that interact with proteins, since they have already passed toxicity and ADME studies.<sup>[193]</sup> In analogy to the Rule of 5 for drug-like molecules (Section 2.1.2.3.2), a *Rule of 3* (MW < 300 Da, CLogP  $\leq 3$ , number of hydrogen bond donors and acceptors each  $\leq 3$ , number of rotatable bonds  $\leq 3$ , and PSA  $\leq 60 \text{ \AA}^2$ ) has been proposed as a guideline for the selection of fragments.<sup>[194]</sup> Such constraints should, in principle, enhance the probability that drug-like molecules will result after the fragments are linked. Virtual screening methods discussed earlier (Section





2.1.2.3.4) may be productively applied to computationally predicting fragments that are likely to interact with the target; indeed, because small fragments are likely to be less conformationally mobile than larger molecules, virtual screening has at least one less confounding factor when applied to fragments as opposed to larger, more flexible molecules.

A retrospective analysis of 18 different drug leads confirmed that fragments should not be larger than 20 nonhydrogen atoms or about 300 Da (for some targets, the upper limit was set to 250 Da). However, a lower limit to the MW also should be taken into account in a fragment library<sup>[195]</sup> because smaller, less complex fragments that only contain single rings with small substituents have a greater likelihood of binding in multiple orientations; therefore elaborated fragments may have different binding geometries than unelaborated fragments.<sup>[196]</sup> For example, the crystal structure of **2.39** bound to AmpC  $\beta$ -lactamase ( $K_i$  1  $\mu$ M) was compared to the crystal structures of fragments (**2.40** ( $K_i$  40 mM), **2.41** ( $K_i$  19 mM), and **2.42** ( $K_i$  10 mM)) derived from **2.39**. None of the fragments bound in the corresponding positions when they were part of **2.39**. In fact, they were in different orientations, and the fragments bound in two entirely different binding sites. It is normally assumed that the geometries of the parts of larger, more potent molecules from elaboration of fragments are the same as the fragments from which they were derived. However, by this converse experiment, deconstruction of molecules into fragments, it is apparent that small fragments can bind differently than those fragments bind when part of a more complex molecule, implying that there will be some potentially good inhibitors missed in molecules constructed from different fragments in a fragment-based approach. Because of the potential for different orientations of small fragments, a lower limit for fragment sizes of approximately 150 Da minimizes the chance that a fragment might bind in a different orientation in the target upon elaboration.<sup>[197]</sup> On the other hand, similar larger molecules seem to have a high degree of structural conservation to a binding site. A survey of the *Protein Data Bank* (PDB), which stores experimentally determined protein structures, showed that

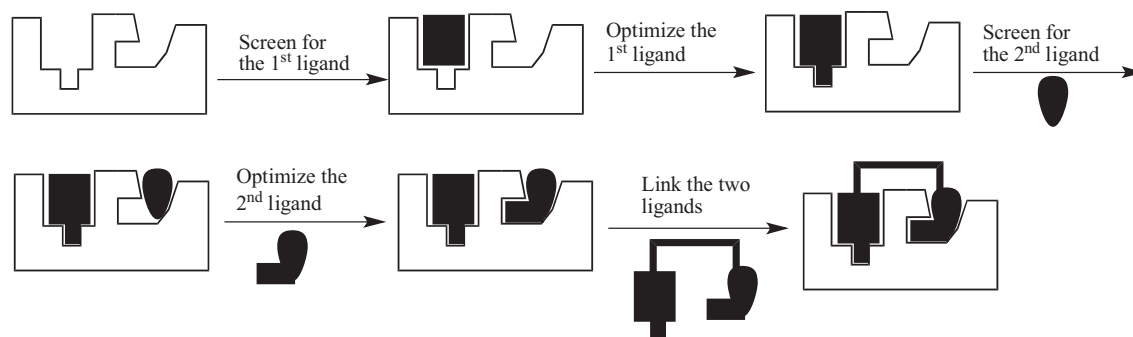
the binding orientation of a majority of structurally similar ligands in a protein is conserved, especially when the MWs are greater than 370 Da; however, binding site side-chain movements occur in half of the ligand pairs.<sup>[198]</sup> This supports the tenet in drug design that making small modifications to lead molecules will retain activity. For simple fragments, effective molecular recognition elements are important. Hydrophobic and electrostatic interactions are two important forces between ligands and proteins (discussed further in Chapter 3). Most structures in a generic fragment library, therefore, should include a hydrophobic group<sup>[199]</sup> and a strong hydrogen bonding or charged group.<sup>[200]</sup>

A comparison of fragment-based drug design (FBDD) approaches and HTS is given in Table 2.7.<sup>[201]</sup> The theory upon which FBDD is based can be tracked to Jencks in 1981,<sup>[202]</sup> who showed that binding efficiencies can be thought of as a combination of two or more moieties of the molecule; experimental evidence was provided by Nakamura and Abeles when they rationalized the potency of the first statin, mevastatin, as a combination of two “fragments” binding into separate, but adjacent, binding pockets.<sup>[203]</sup>

The actual exploitation of the method came in 1996 with a report by Fesik and coworkers at Abbott Laboratories of a new technique called *SAR by NMR*, an approach for screening fragments and elaborating them into a potent lead using NMR spectrometry.<sup>[204]</sup> The first step of the process (Figure 2.8) involves screening a library of small compounds, 10 at a time, by observing a <sup>15</sup>N-chemical shift in the heteronuclear single quantum coherence NMR spectrum for a specific amide nitrogen of the protein. Once a fragment is identified that causes a notable change in this chemical shift, a library of similar analogs is screened to identify compounds with optimal binding at that site. Then, with a saturating (excess) concentration of the first optimized ligand, a second library of compounds is screened to find a compound that binds at a nearby site, and then the second compound is optimized by screening a library of related compounds. On the basis of the NMR spectrum of the ternary complex of the protein and the two bound ligands, the location and orientation of each ligand is determined, and compounds are synthesized

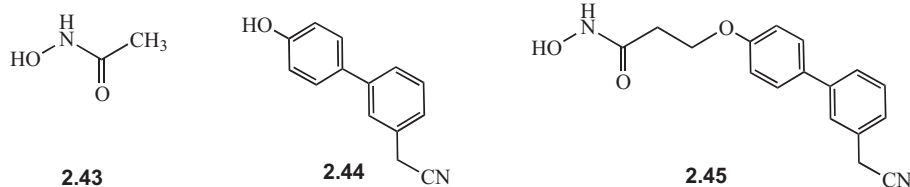
**TABLE 2.7** Comparison of Fragment-Based Approaches and High-throughput Screening

Fragment-Based Approaches	HTS
Emphasis on efficiency	Emphasis on potency
Screen a few hundred to a few 1000 compounds	Screen hundreds of thousands of compounds
MW range 120–250	MW range 250–600
Hit activity millimolar–30 $\mu$ M	Hit activity 30 $\mu$ M–nanomolar
High proportion of atoms in pharmacophore, i.e., high ligand efficiency	Hits contain groups that contribute poorly to binding or act as scaffold; low ligand efficiency
Biophysical screening techniques (NMR, X-ray, surface plasmon resonance) required because of weak binding	In vitro screening; often generates false positives and high attrition during validation
Protein structure-based information key to validation and prioritization of hits	Chemical (re)synthesis required for validation and prioritization of hits
Hit to lead usually requires synthesis of only a few compounds	Usually requires several iterations of high-throughput chemistry; protein structure can lower this
Design intensive	Resource intensive
Requires expertise and knowledge in protein-structure and protein–ligand interactions	HTS requires extensive infrastructure for storing and handling compound collections, screening, automation, data processing, and chemistry

**FIGURE 2.8** SAR by NMR methodology

in which the two ligands are covalently attached. When two low-affinity fragments are linked into a single molecule that effectively delivers each fragment to its respective site of interaction with the target, then a compound with *much* higher affinity results. This is because the free energy of binding becomes the sum of three free energies: those of the two ligands plus a free energy to reflect the effect of linking (note that the *sum* of the free energies of the two ligands translates to the *product* of their binding affinities!). The free energy from linking likely has numerous components that might individually result in either a positive or negative effect, but a positive entropic effect (reduced “randomness” of the individual fragments) is likely a major contributor. An example of this is the identification of the first potent inhibitor of the enzyme stromelysin, a *matrix metalloprotease* (a family of zinc-containing hydrolytic enzymes responsible

for degradation of extracellular matrix components, such as collagen and proteoglycans, in normal tissue remodeling and in many disease states such as arthritis, osteoporosis, and cancer),<sup>[205]</sup> as a potential antitumor agent.<sup>[206]</sup> Matrix metalloproteases are generally inhibited by compounds that contain a hydroxamate moiety to bind to the zinc ion. A library of hydroxamates was screened, and acetohydroxamic acid (**2.43**) was identified with a  $K_d$  of 17 mM (generally regarded as exceedingly weak binding affinity). A focused screen of hydrophobic compounds was carried out in the presence of saturating amounts of acetohydroxamic acid, and biphenyl analogs were identified; optimization led to **2.44** with a  $K_d$  of 20  $\mu$ M. From the NMR spectrum, the best site for a linker was expected to be between the methyl of acetohydroxamic acid and the hydroxyl group of **2.44**. Consequently, alkyl linkers of varying chain length were



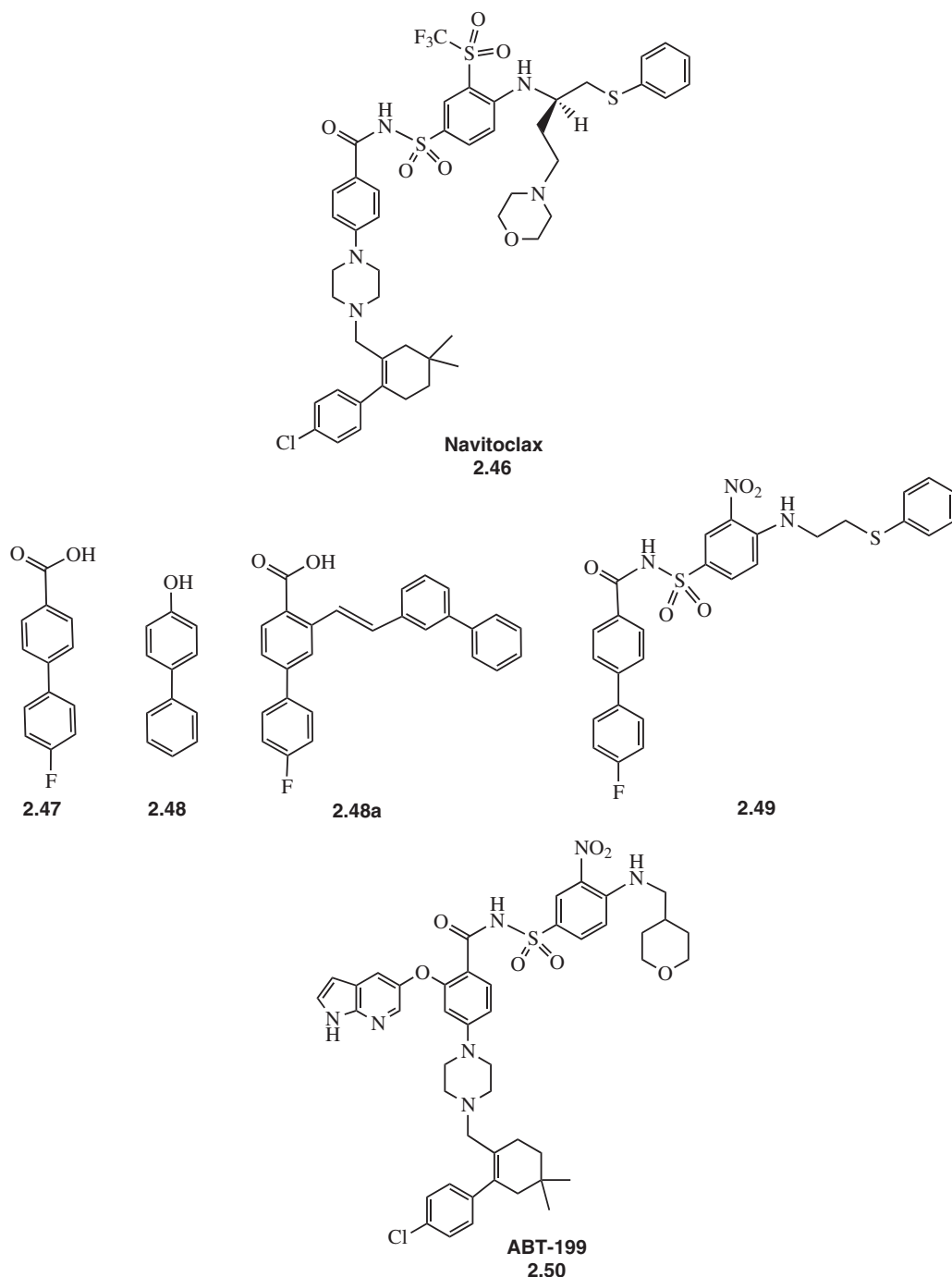
tried, and the best was a one-carbon linker, giving **2.45** having a  $K_d$  of 15 nM! The  $\Delta G$  for **2.43** is  $-2.4$  kcal/mol, for **2.44** is  $-4.8$  kcal/mol, and for the linker is  $-2.6$  kcal/mol; the total, therefore, is  $-9.8$  kcal/mol. It took about six months to identify this inhibitor; prior to this study, 115,000 compounds had been screened with no leads.

The first compound in clinical trials derived from the SAR by NMR method is navitoclax (**2.46**), an anticancer drug that inhibits the protein Bcl-x<sub>L</sub>, an antiapoptotic B-cell lymphoma protein.<sup>[207]</sup> Normal cellular homeostasis is regulated by expression of antiapoptotic proteins, such as Bcl-x<sub>L</sub>, Bcl-2, and Bcl-w and proapoptotic proteins, such as Bak, Bax, and Bad.<sup>[208]</sup> For some cancers, *apoptosis* (programmed cell death) is circumvented by overexpression of the antiapoptotic proteins Bcl-2 or Bcl-x<sub>L</sub>, which makes them targets for the development of new anticancer drugs.<sup>[209]</sup> There is a hydrophobic groove on the surface of these proteins to which the proapoptotic proteins bind, and the 3D structure of this binding region was determined by NMR spectrometry.<sup>[210]</sup> SAR by NMR was used to identify inhibitors that bind in the hydrophobic groove of Bcl-x<sub>L</sub>.<sup>[211]</sup> A 10,000-compound fragment library was screened, and **2.47**, with a  $K_i$  of 300  $\mu$ M, was identified as a ligand for Bcl-x<sub>L</sub>. Comparison of the structure of this ligand complex to that of the Bcl-x<sub>L</sub>/Bak peptide complex suggested a proximal second site. A second screen was run in the presence of an excess of **2.47** using a 3500-fragment compound library, which identified **2.48**,  $K_i$  6 mM. A variety of possible linkers were assessed, and a *trans*-olefin was deemed best, giving **2.48a** with a  $K_i$  of 1.4  $\mu$ M. Further synthetic manipulation resulted in **2.49**,  $K_i$  36 nM. It was found that **2.49** was too hydrophobic, making it poorly aqueous soluble and tightly bound to serum albumin. Consequently, the polarity of **2.49** was increased for improved pharmacokinetics, leading to the antilymphoma drug **2.46**. Note that, despite the relatively good pharmacokinetic properties of **2.46**, its molecular mass (974 Da) far exceeds the Rule of 5 maximum of 500 Da for good oral bioavailability. In Phase II clinical trials it was found that **2.46** caused thrombocytopenia (low platelet count), and it was terminated. The cause for the platelet loss was found to be inhibition of Bcl-x<sub>L</sub>; this led the Abbott group to modify **2.46** in search of a selective Bcl-2 inhibitor. With the aid of cocrystal structures of small molecules in Bcl-2, the first-in-class Bcl-2-selective inhibitor, ABT-199 (**2.50**) was developed, which showed potent antitumor activity (chronic lymphocytic leukemia)

without platelet loss.<sup>[212]</sup> Several other drugs discovered from fragment-based approaches, rather than high-throughput screens, are reaching clinical trials.<sup>[213]</sup>

Sounds simple, doesn't it? But let's think about what is involved in carrying out SAR by NMR. The method requires screening compounds and observing a specific <sup>15</sup>N-amide chemical shift for binding. Where did the <sup>15</sup>N come from? This had to be incorporated into the protein because natural abundance <sup>15</sup>N is not sufficiently high to detect. To incorporate <sup>15</sup>N, it is necessary to be able to express the protein in a microorganism, and then grow the microorganism on <sup>15</sup>NH<sub>4</sub>Cl as its sole nitrogen source. This gives the protein with all <sup>15</sup>N-containing amino acids. To perform the NMR experiments, large amounts of soluble (>100  $\mu$ M) protein (>200 mg per spectrum) are needed; therefore, an efficient overexpression system for the protein is needed. Then the protein has to be purified, and its complete structure determined by 3D and 4D NMR techniques, so that the position of every amino acid residue in the protein is known (which is needed to determine when the two ligands are bound in nearby sites). This means that the protein target should have a mass less than about 40 kDa (the current limit for rapid protein NMR spectra, although spectra of larger proteins is possible<sup>[214]</sup>). Although it appears that this is a highly specialized technique, it is used widely because molecular biology and protein chemistry techniques have been well developed, making overexpression of proteins in microorganisms and their purification routine.<sup>[215]</sup> NMR instrumentation and methods also have made structure determination plausible. If the structure can be determined, SAR by NMR provides a technique to screen, by automation, about 1000 compounds a day and identify, relatively rapidly, potent protein binders.<sup>[216]</sup> Integration of a medicinal chemist's input into computational methods can accelerate fragment-based lead discovery.<sup>[217]</sup>

Ellman and coworkers have developed a combinatorial lead optimization approach using the basic principles described above for SAR by NMR, except without the use of NMR spectrometry and without the need for any structural or mechanistic information about the target protein!<sup>[218]</sup> First, a diverse library of compounds is synthesized in which each molecule incorporates a common chemical linkage group (Figure 2.9). Next, the library is screened to identify any member that shows even weak binding to the target. Third, a new library is constructed containing all combinations of any two of the active compounds linked to

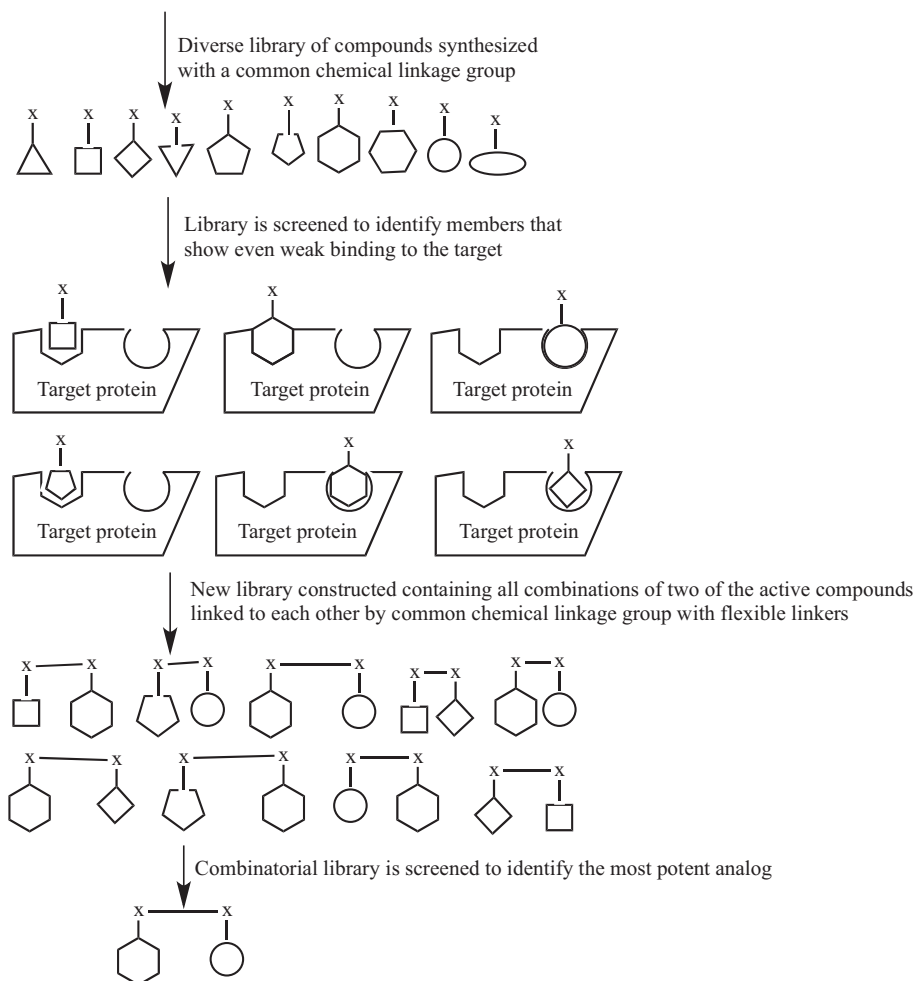


each other by the common chemical linkage group through a set of flexible linkers. Then this combinatorial library is screened to identify the most potent analog. The method depends on two analogs binding in nearby sites (although it is not known which two will bind or where the sites are) and finding the appropriate linker size combinatorially so the linked active compounds take advantage of the additive free energy gain of the three elements, the two compounds and the linker. This approach was used to identify a potent

( $IC_{50}$  64 nM) and selective inhibitor of one type of tyrosine kinase.

A complementary method to SAR by NMR is SAR by MS.<sup>[219]</sup> This is a high-throughput MS-based screen that quantifies the binding affinity, stoichiometry, and specificity over a wide range of ligand-binding energies. A set of diverse compounds is screened by MS to identify those that bind to the receptor. Competition experiments are used to identify the ones that bind to the same site and those that do



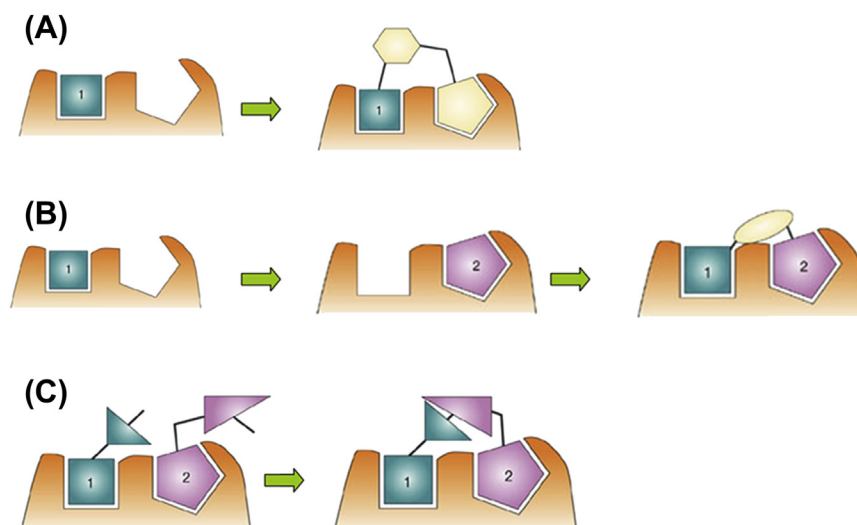
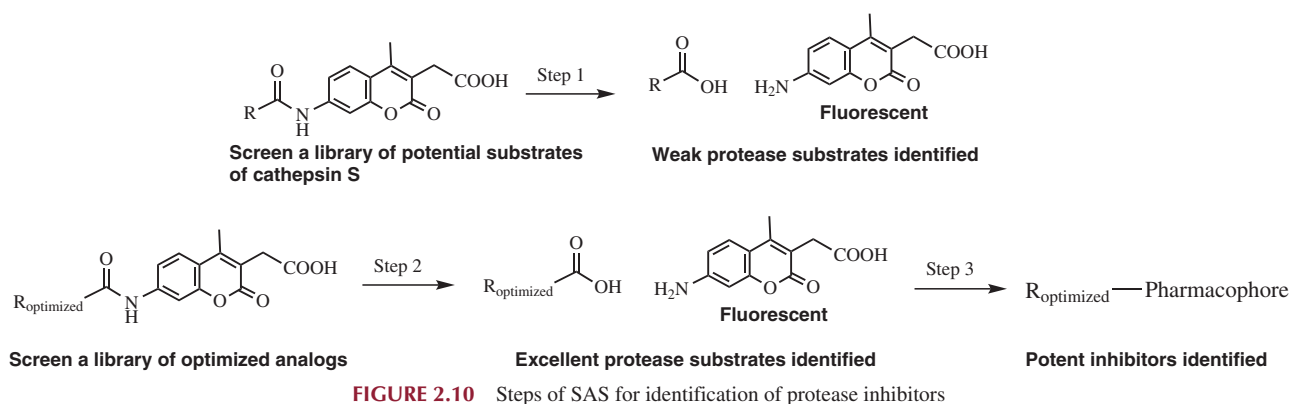


**FIGURE 2.9** Ellman combinatorial methodology for lead generation with an unknown or impure protein

not. If two compounds bind at different binding sites, then a ternary complex of the two molecules plus the receptor is detected in the mass spectrum. If the two compounds bind at the same site, the tighter binding molecule displaces the other from the binding site, and only a binary complex is detected. By varying the substituent size on various classes of compounds and rescreening, it is possible to identify those molecules that bind at nearby sites as the ones that become competitive once a larger substituent is appended to one of the molecules. Once adjacent binding sites are realized, then the same methodology as for SAR by NMR, namely, attaching the two or more molecules to each other with linkers, can be employed. This approach was applied to the development of a new class of small molecules with high affinity for the hepatitis C virus-internal ribosome entry site IIA subdomain, which mediates initiation of viral-ribonucleic acid (RNA) translation.<sup>[220]</sup> MS of the company's compound collection (180,000 compounds) led to the identification of a benzimidazole analog with activity, which was optimized to submicromolar binding affinity for the IIA RNA construct using SAR by MS. The optimized

benzimidazoles reduced viral RNA in a cellular replicon assay at concentrations comparable to the binding constants observed in the MS assay.

Ellman and coworkers developed a substrate-based fragment identification method for protease inhibitors, called *substrate activity screening* (SAS).<sup>[221]</sup> This method addresses two key challenges in fragment-based screening: (1) the efficient identification of weak binding fragments and (2) the rapid optimization of the initial weak binding fragments into high-affinity compounds. SAS has three steps (Figure 2.10): (1) a library of substrates consisting of the substrate-catalytic functionalities, in this case, the amide of the acylaminocoumarin and diverse, low-MW fragments, in this case, the R groups, is screened using a single-step, high-throughput fluorescence-based substrate assay; (2) the activity of the substrate is optimized by rapid analog synthesis and evaluation; (3) the optimized substrates are converted to inhibitors by replacement of the substrate-catalytic functionality with inhibitor pharmacophores, which match the catalytic residues in the active site (in this case, the aminocoumarin



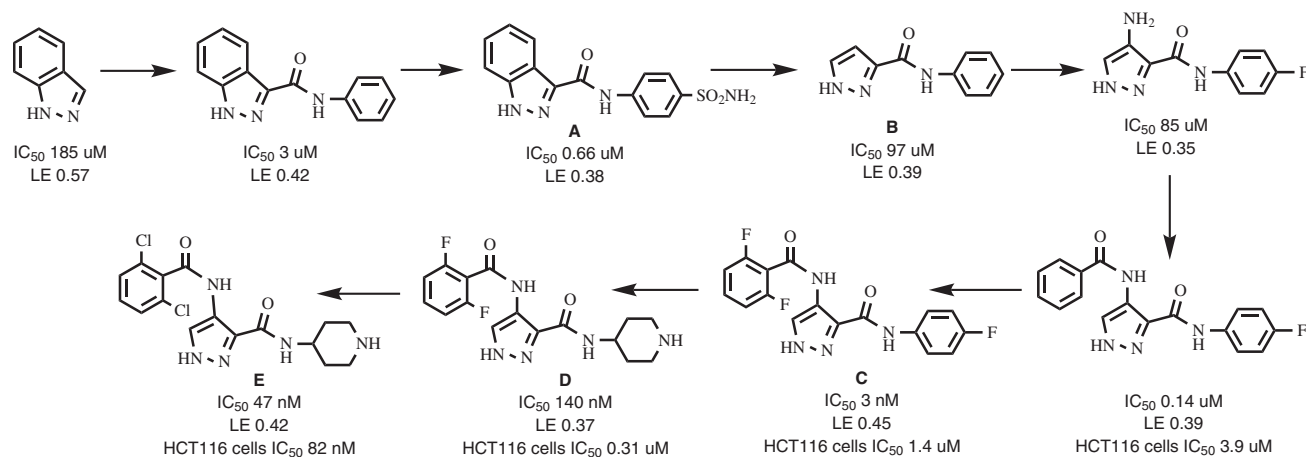
**FIGURE 2.11** Three approaches to linking fragments: (A) fragment evolution, (B) fragment linking, and (C) fragment self-assembly. Reprinted with permission from Macmillan Publishers Ltd: *Nature Reviews Drug Discovery* (Reese, D. C.; Congreve, M.; Murray, C. W.; Carr, R. *Fragment-based lead discovery*. *Nat. Rev. Drug Discov.* **2004**, *3*, 660–672) Copyright 2004.

was replaced by H to give an aldehyde, a known functionality for cathepsin inhibitors). In SAS, both an active enzyme and productive active site binding are required for catalytic function. However, SAS has some prominent advantages, such as being able to detect weak binding fragments because catalytic substrate turnover results in signal amplification (from release of a fluorescent molecule), and therefore, even very weak substrates can be identified at concentrations where only minimal binding to the enzyme occurs. Also, it is a high-throughput and straightforward technique to perform. Using this method, a 9 nM inhibitor of cathepsin S, which has been implicated in autoimmune diseases such as rheumatoid arthritis and multiple sclerosis,<sup>[222]</sup> was identified.

As illustrated in the foregoing examples, after the fragment hits are identified, the next step is to transform them into a lead structure while maintaining drug-like properties in the generated molecule. There are three general strategies for converting fragments into a drug-like lead compound:

(1) *fragment evolution*, (2) *fragment linking*, and (3) *fragment self-assembly* (Figure 2.11). If the target structure is available, then elaboration of the fragment can be guided by the X-ray crystallographic or NMR spectral data of the fragment bound to the target. The fragment must also be optimized for pharmacokinetic properties.

*Fragment evolution* (Figure 2.11(A)) involves the addition of functionality to the fragment to allow for binding to additional pockets in the target. An example of fragment-based lead discovery incorporating fragment evolution (followed by lead modification to an optimized compound) is shown in Figure 2.12. Note that LE was used to guide the overall process. Ultimately, a balance had to be reached between potency for cyclin-dependent kinase (CDK2) inhibition, pharmacokinetics, and tumor cell activity. Changes in structure **A** in Figure 2.12 did not lead to large increases in potency, so a different strategy was taken, i.e., removal of the benzene ring from the benzopyrazole to give **B**, which had much lower potency



**FIGURE 2.12** Example of fragment-based lead discovery incorporating the fragment evolution approach followed by lead modification to an optimized compound. Ligand efficiencies help guide the overall process.

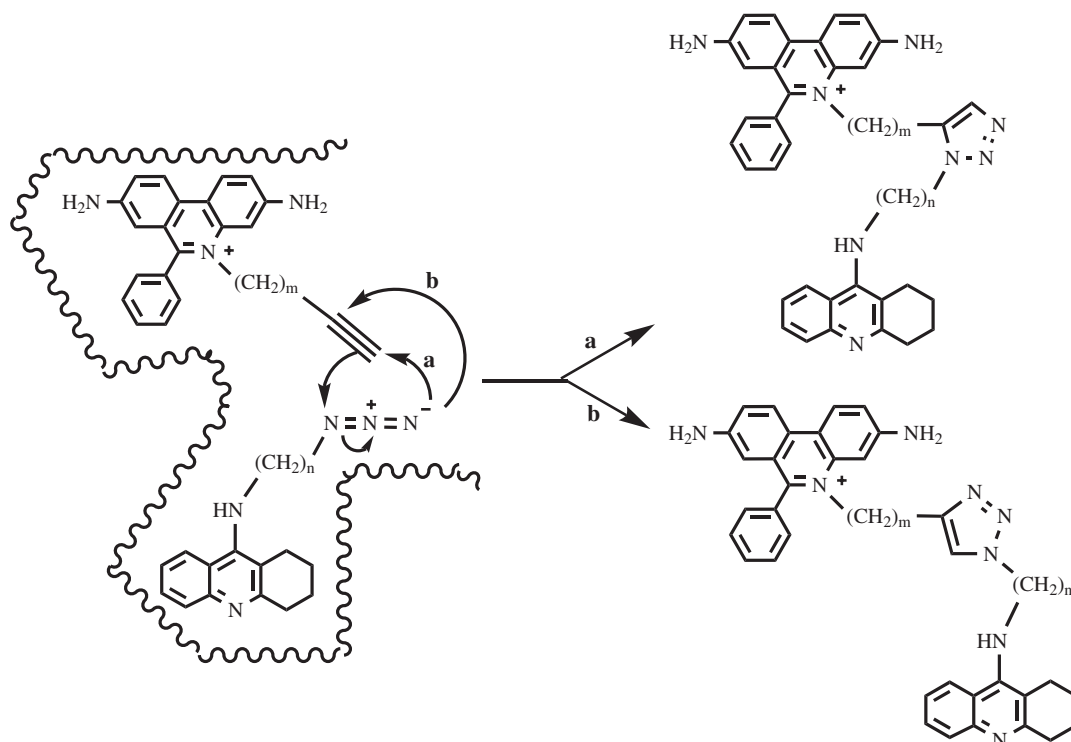
(but similar LE). Growing from the pyrazole ring led to **C**, which was potent and had good pharmacokinetic properties, but activity in tumor cells was only moderate. The measured  $\log P$  was found to be  $>4$ . To increase polarity, the *p*-fluorophenyl substituent was replaced by a 4-piperidiny group (**D**), which lowered the enzyme inhibitory potency, but increased tumor cell activity. Further increases in lipophilicity and size by conversion of the 2,6-difluorophenyl ring of **D** to a 2,6-dichlorophenyl ring in **E** increased enzyme inhibitory potency as well as antitumor cell activity. This compound showed excellent *in vivo* activity and entered clinical trials.

As the name implies, *fragment linking* (Figure 2.11(B)) involves the linkage of two or more fragments that bind in proximal pockets, leading to higher affinity. The streptomycin example given in the discussion of SAR by NMR above was fragment linking: fragments in two adjacent binding pockets were linked to produce a  $10^6$  increase in potency relative to the hydroxamate fragment.

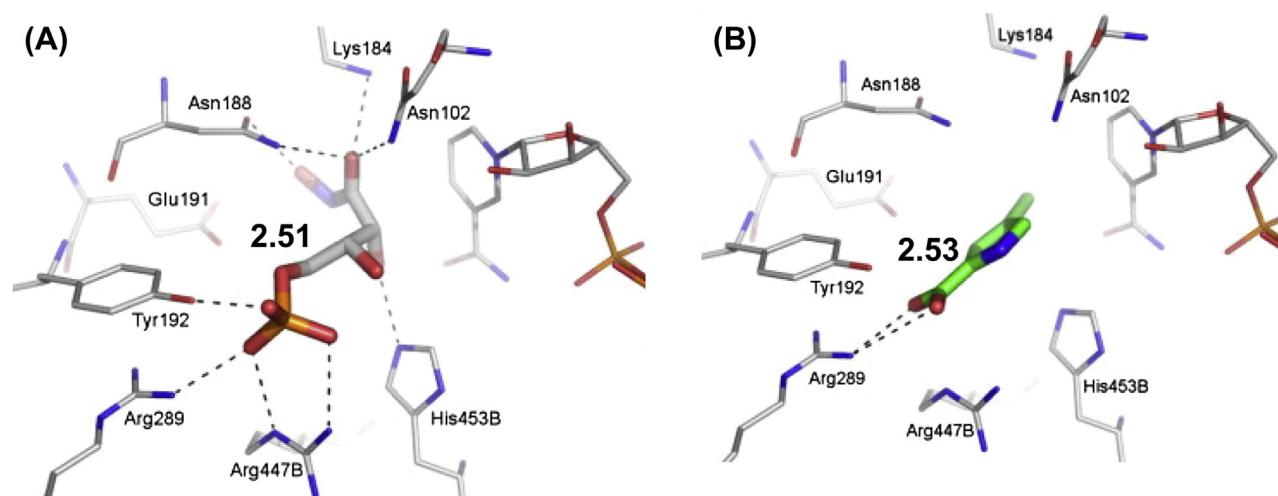
*Fragment self-assembly* (Figure 2.11(C)) is when fragments with complementary functional groups are allowed to react within the binding sites of the target. An example of this is the origins of “click chemistry”, where a series of alkyne analogs and azide analogs were incubated with acetylcholinesterase; the alkyne analog and azide analog that bound in adjacent binding pockets were held in the optimal position to react and give the corresponding triazoles, one of which had femtomolar inhibitory potency (Figure 2.13).<sup>[223]</sup>

A lead discovery example that illustrates the fragment-based approach as well as several other concepts discussed throughout this chapter follows. *Trypanosoma brucei* is the causative parasite of African sleeping sickness, one of the most widespread and lethal diseases in Africa. Ruda, et al.<sup>[224]</sup> set out to discover a new inhibitor

of 6-phosphogluconate dehydrogenase (6PGDH), a key enzyme for the function and survival of *T. brucei*. An X-ray crystal structure of 6PGDH in complex with a known inhibitor (**2.51**) was available. It is noteworthy that although **2.51** is a potent inhibitor of the enzyme, it does not possess trypanocidal (trypanosome-killing) activity. This deficit is attributed to the inability of the inhibitor to pass through membranes of the organism, thereby preventing it from reaching the enzyme target (pharmacokinetics). The poor membrane permeability is attributed to the double-negatively charged phosphate moiety (the basis for such reasoning will be discussed further in Sections 2.2.5.3 and 2.2.5.4). The crystal structure revealed that the enzyme contains a cluster of positively charged moieties that interact with the negatively charged phosphate. Therefore, the objective was to identify a new class of inhibitors that still contained a negatively charged moiety (to retain binding properties), but that was less likely to preclude permeability through membranes. The researchers started with an electronic database of commercially available chemicals and filtered it to retain only molecules that had  $MW < 320$  and one of the following negatively charged groups: phosphonate, sulfonate, sulfonamide, carboxylic acid, or tetrazole.<sup>[225]</sup> This operation resulted in a set of 64,000 compounds. The compounds were computationally docked into the active site of the enzyme, with the requirement that the negatively charged group docked into the same region as the phosphate group of **2.51**. To validate the docking method, it was demonstrated that when **2.51** was docked computationally, a model resulted that closely resembled the crystallographically determined enzyme–inhibitor complex. About 6000 compounds gave reasonable docking poses with the enzyme. Using a computationally determined similarity approach, the 6000 compounds were divided into similar



**FIGURE 2.13** Example of fragment-based lead discovery incorporating the fragment self-assembly approach: click chemistry



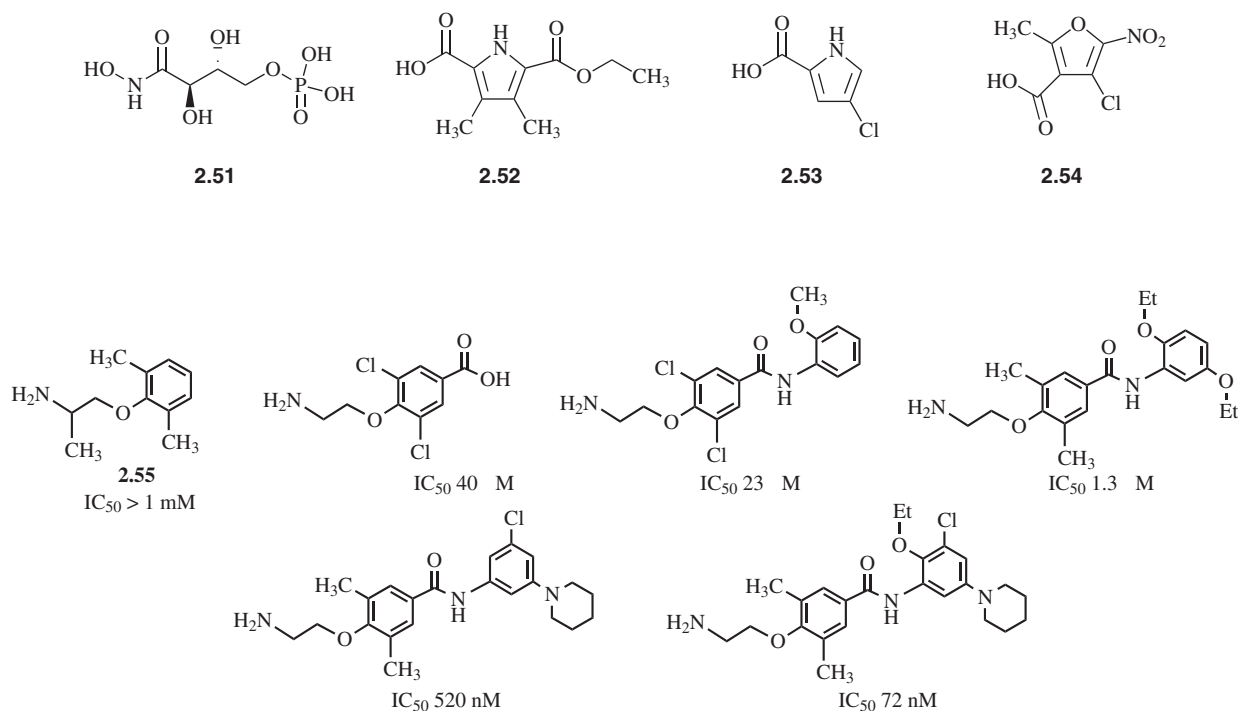
**FIGURE 2.14** (A) Structure of **2.51** complexed with 6-phosphogluconate dehydrogenase (6PGDH) determined by X-ray crystallography. (B) Structure of **2.53** complexed with 6PGDH predicted by computational docking. *From Ruda, et al. Bioorg. Med. Chem.* **2010**, *18*, 5056–5062.

groups (clusters), and 71 molecules were selected for purchase. The 71 compounds were tested as enzyme inhibitors, first at a very high concentration (200  $\mu\text{M}$ ), and then promising compounds were tested at a range of doses to determine binding potency. In this way, compounds **2.52**, **2.53**, and **2.54** were identified as fragments with moderate affinities but high ligand efficiencies, and thus as reasonable starting points for further modification. A

comparison of the computationally derived docking pose of **2.53** with the complex of the protein and **2.51** (Figure 2.14, B vs A)) suggests where potential substituents could be added to **2.53** for additional productive interactions with the enzyme.

Another fragment-based HTS approach, rather than starting with a diverse random fragment library, is to start





**FIGURE 2.15** Progression from mexiletin (**2.55**, identified by fragment-based screening) to a potent orally bioavailable uPA

with a library of small known drugs, which have already been shown to have drug-like pharmacokinetic and safety characteristics (because they already are drugs) and see if they have other activities, a method Wermuth has called *selective optimization of side activities* (SOSA).<sup>[226]</sup> Because essentially all drugs can bind to more than one target, this approach searches for the minor off-target hits, and then optimizes the side activity into the main activity, diminishing (or eliminating) the original target activity. A library of small drug molecules can be purchased from Prestwick Chemical (Washington, DC, USA).<sup>[227]</sup> For example, the antiarrhythmic drug mexiletin (**2.55**, Fig. 2.15, Mexitil, MW 179) was found to be a weak inhibitor ( $IC_{50} > 1$  mM) of urokinase-type plasminogen activator (uPA),<sup>[228]</sup> a serine protease that, when bound to its receptor, catalyzes the conversion of plasminogen to plasmin, which is responsible for a variety of proteolytic processes in the extracellular matrix.<sup>[229]</sup> Therefore, uPA is implicated in the progression of disease states associated with abnormal tissue destruction, such as multiple sclerosis<sup>[230]</sup> and cancer.<sup>[231]</sup> Figure 2.15 shows the structural progression from mexiletin to a potent orally bioavailable uPA inhibitor, using X-ray crystallography to guide the optimization.

The earlier discussion on the hit-to-lead process (Section 2.1.2.3.5) also applies in part to fragment-based lead discovery. For example, confirmation of activity, many computational methods, and early SAR assessments are

already an inherent part of fragment-based approaches. On the other hand, taking the opportunity to perform basic calculations of physical properties and to conduct early ADME-tox and intellectual property assessments on leads discovered by fragment-based methods is still a good idea before proceeding into full-scale lead modification.

The SOSA example is a segue to the next section (Section 2.2), where we take a detailed look at how leads described in Section 2.1 are modified, resulting in a drug candidate ready for advanced preclinical studies. During this next phase of the drug discovery process, there is enhanced concern with pharmacokinetic (ADME) and toxicological properties as the potency at the intended target (pharmacodynamics) is being increased.

## 2.2. LEAD MODIFICATION

Once your lead compound is in hand, how do you know what to modify in order to improve the desired pharmacological, toxicological, and pharmacokinetic properties? The lead modification process, often referred to as *lead optimization*, can be context dependent, that is, the approach may vary depending on what property or properties most require improvement. In the discussion below, as well as in subsequent chapters, general principles and case examples are presented that provide a flavor for how specific challenges might be approached by the medicinal chemist.