Contents

Abbreviations

ADME: absorption, distribution, metabolism, and excretion
BLA: Biologics License Application
CADD: computer-aided drug discovery
DEL: DNA-encoded compound library
DHHS: U.S. Department of Health and Human Services
DMPK: drug metabolism and pharmacokinetics
FBDD: fragment-based drug discovery
FDA: U.S. Food and Drug Administration
GPU: graphics processing unit
HCV: hepatitis C virus
HDL: high-density lipoprotein
HMG-CoA: 3-hydroxy-3-methylglutaryl coenzyme A
HTS: high-throughput screening
IND: Investigational New Drug
LDL: low-density lipoprotein
mRNA: messenger RNA
NDA: New Drug Application
NIH: National Institutes of Health
NMEs: new molecular entities
PDUFA: Prescription Drug User Fee Act
SBDD: structure-based drug design
siRNA: small interfering RNA

Drug Discovery or Drug Invention?

The conventional phrase *drug discovery* makes sense for therapeutic compounds obtained from plants and other organisms. Today, however, only a fraction of the new drugs introduced each year are discovered in nature. Instead, most drugs are not discovered, but are totally new compounds, painstakingly optimized against many criteria through an interplay of design and experimentation. In that sense, today's new drugs are more invented than discovered.

The current paradigm for drug development grew out of synthetic organic chemistry, which arose as the dye industry in the late 19th century and has continued to flourish. Dyes are colored compounds with selective affinity across various biological tissues. Study of these interactions stimulated Paul Ehrlich to postulate the existence of chemical receptors in tissues that interacted with and "fixed" the dyes. Similarly, Ehrlich thought that unique receptors on microorganisms or parasites might react specifically with certain dyes and that such selectivity could spare normal tissue. Ehrlich's work culminated in the invention of *arsphenamine* in 1907, which was patented as "salvarsan," suggestive of the hope that the chemical would be the salvation of humankind. This and other organic arsenicals were used to treat syphilis until the discovery of *penicillin*. Gerhard Domagk demonstrated that another dye, *prontosil* (the first clinically useful sulfonamide), was dramatically effective in treating streptococcal infections, thereby launching the era of antimicrobial chemotherapy. The collaboration of pharmacology with chemistry on the one hand and clinical medicine on the other has been a major contributor to the effective treatment of disease, especially since the middle of the 20th century.

Early on, new compounds could be tested for their activities only in whole organisms. This is how the nonsteroidal anti-inflammatory drug *indomethacin* was discovered, for example (Brune and Hinz, 2004). In the past 70 years, researchers have begun to understand in considerable detail the cellular and molecular mechanisms of disease. As a result of this basic biomedical research, it is possible to do initial testing of compounds *in vitro* ("in glass"), using cellular and molecular assays. For example, one could look for the cellular responses due to inhibition of a protein involved in a disease process. In this scenario, by testing enough appropriately chosen compounds, one could develop at least a partial understanding of which types of compounds are most likely to be active and then use this information to steer the program of chemical synthesis and testing toward increasingly potent compounds.

In the 1980s, it became practical to determine high-resolution three-dimensional structures of complex organic molecules and even larger molecules such as proteins, using and refining the techniques of X-ray crystallography pioneered by Hodgkin, Kendrew, and Perutz in the mid-20th century. It was already known that many drugs worked by binding tightly to a disease-related protein and thereby modulating (e.g., inhibiting or activating) its biological function, but the atomic details of these interactions had remained mysterious. As a consequence, the only way to advance a drug discovery project had been by synthesizing and testing one compound after another. Now, with the protein's three-dimensional structure in hand, one could finally hope to design a compound that would bind with high affinity by fitting snugly into a pocket in the protein, such as an enzyme's active site. Thus, protein crystallography enabled *structure-based drug design* (SBDD), where the three-dimensional structure of the drug target is used to guide creation of tight-binding compounds, often called *ligands*.

Around the same time, computer technology began to advance rapidly. This accelerated the data processing needed to go from X-ray diffraction patterns to protein structures (i.e., three-dimensional atomic coordinates) and enabled interactive visualization of complex protein structures comprising thousands of atoms. It also opened new vistas in *computer-aided drug discovery* (CADD), including the use of molecular simulations to model the physical interactions of compounds and proteins, and the development of tools to encode, archive, share, and analyze chemical and pharmacological data. In parallel, automation and miniaturization have dramatically increased experimental throughput, notably through robotic *high-throughput screening* (HTS), in which hundreds of thousands of compounds can be tested rapidly and at relatively low cost in cellular or molecular activity assays. Today, excitement about the power of artificial intelligence motivates wide-ranging efforts to apply these technologies to drug discovery.

The following section goes into more detail regarding the process of drug discovery, focusing on so-called *small-molecule* drugs, organic compounds with molecular weights typically less than 500 Da, which have traditionally been the most common type of drug. Subsequent sections introduce biological drugs, such as antibodies and other engineered biomolecules.

Target Identification

Today, most small-molecule drug discovery projects grow out of basic research that implicates a specific macromolecule, usually a protein, as a key player in a disease and, further, suggests that a small molecule which binds this macromolecule could be used to treat the disease. The macromolecule thus becomes a candidate *drug target*. Many small-molecule drugs are inhibitors (antagonists), which work by reducing the activity of their macromolecular target. Examples include the statins, which reduce cholesterol synthesis by binding and inhibiting the enzyme 3-hydroxy-3-methylglutaryl (HMG) coenzyme A (CoA) reductase, and β-lactam antibiotics, which kill bacteria by inhibiting enzymes involved in the synthesis of bacterial cell walls. However, some small molecules are activators (agonists) rather than inhibitors. Activators frequently target proteins whose normal role involves cell signaling, such as hormone receptors. For example, the asthma medication *albuterol* dilates bronchi by binding and activating β adrenergic receptors on bronchial smooth muscle, thereby mimicking the effect of adrenaline (epinephrine; see Chapter 10).

Candidate drug targets have been identified in many ways (Hughes et al., 2011). For example, the enzymes targeted by the β-lactam antibiotics were unknown in advance and were discovered precisely because they are bound by these naturally occurring antibiotics. In contrast, the target of the statins, HMG-CoA reductase, was identified by elucidation of the pathways of cholesterol synthesis (Tobert, 2003), and this information was used to help discover the first statins. Similarly, as researchers have determined the regulatory functions of human protein kinases—enzymes that change the activities of other proteins by covalently attaching phosphate groups to their hydroxyl-containing side

chains—specific kinases have been targeted for small-molecule drug discovery (Cohen et al., 2021). Many kinase inhibitors are anticancer agents that work by inhibiting protein kinases that accelerate cell proliferation. Some of these targeted kinases carry abnormal, cancer-associated mutations that make them hyperactive, so inhibiting them returns their regulatory activities toward normal. The pioneering example of this scenario is the drug *imatinib*, which inhibits a cancer-associated mutant protein kinase, the Bcr-Abl tyrosine kinase, and is used to treat chronic myelogenous leukemia (Buchdunger et al., 2002).

In recent years, technological advances enabling genome-wide experimentation (*omics*) have opened new approaches to identifying candidate targets (Lindsay, 2003; Paananen and Fortino, 2020). Fast, inexpensive genome sequencing facilitates genome-wide association studies, in which variations in the susceptibility to a disease across many people are correlated with variations in specific genes, leading to suggestions for gene products (i.e., proteins), that may be suitable drug targets. The growing availability of patient genomic data in the context of patients' electronic medical records will likely open new opportunities for data mining in support of target discovery in the coming years. It has also become routine to measure the quantities of messenger RNA (mRNA) transcribed from thousands of genes simultaneously (the *transcriptome*) and to quantify thousands of translated proteins (*proteomics*). By comparing such data between, for example, cancer cells and normal cells, one can identify proteins transcribed or present at elevated or depressed levels in the disease state. Mining data about these proteins from sources such as biomedical databases, scientific articles, and patents, and integrating it with the omics data, may suggest certain proteins as candidate drug targets.

A totally different approach starts with the use of high-throughput instrumentation and robotics to test a large collection of small molecules (a *chemical library*) for biological activity in a *phenotypic screen* (Swinney and Lee, 2020), which might use automated microscopy and image analysis to determine which compounds produce desired biological effects, such as the activation of a desired gene in cultured human cells or the death of a parasitic microorganism in culture. Various methods may then be used for *target deconvolution* (i.e., to determine how the active small molecules work). For example, candidate targets of compounds found to kill the malarial parasite *Plasmodium falciparum* were identified by cultivating these organisms in gradually increasing concentrations of the compound to select for resistant protozoa and then using omics methods to determine which genes had changed. The proteins encoded by these genes may then become candidate drug targets (Flannery et al., 2013).

Target Validation

After a candidate drug target has been identified, additional research is usually warranted to *validate* it by seeking stronger evidence that a small molecule that binds and modulates it will actually treat the disease (Jones, 2016; Lansdowne, 2018; see Box 1–1). For example, the fact that a protein is more abundant in cancer cells than normal cells by no means proves that it is a suitable drug target. Instead, this might be a correlate rather than a cause, so further research is needed to assess its role. Accordingly, *target validation* aims to "de-risk" a project by lowering the probability that a compound carefully developed to hit the targeted protein will fail in clinical trials, whether because hitting the target does not influence the disease as expected or because the compound generates unanticipated toxicity, termed *on-target* or *mechanism-based* toxicity.

There are no absolute criteria for target validation, nor is there a single method. One approach is to use a *chemical probe*, a small molecule that binds the target, and study its biological effects (Quinlan and Brennan, 2021). This approach requires that such a probe be available, and the fields of *chemical genetics* (Stockwell, 2000) and *chemogenomics* (Bredel and Jacoby, 2004) aim to create selective chemical probes for as many proteins in the human genome as possible. Alternatively, one may use gene silencing via small interfering RNA (siRNA) to block production of the target protein, thereby mimicking the effect of an inhibitor of the protein's activity. Additional insight into the biological role of a candidate drug target may sometimes be obtained by studying genetically modified mice, including *knockout mice,* in which the gene coding for the target has been disabled entirely, and *transgenic mice*, in which expression of the target's gene is placed under the control of a promoter that can be turned on by feeding the animals a specific compound, such as *tetracycline* (Lindsay, 2003).

BOX 1–1 ■ Target Validation: The Lesson of Leptin

Biological systems frequently contain redundant elements or can alter expression of drug-regulated elements to compensate for the effect of the drug. *In general, the more important the function, the greater the complexity of the system.* For example, many mechanisms control feeding and appetite, and drugs to control obesity have been notoriously difficult to find. The discovery of the hormone leptin, which suppresses appetite, was based on mutations in mice that cause loss of either leptin or its receptor; either kind of mutation results in enormous obesity in both mice and people. Leptin thus appeared to be a marvelous opportunity to treat obesity. However, on investigation, it was discovered that obese individuals have high circulating concentrations of leptin and appear insensitive to its action.

Target Druggability

It is important to know whether the candidate target is *druggable*, that is, whether it can, in principle, bind a small molecule with sufficient affinity. If the protein has been the target of a prior drug discovery effort, there may be informative small-molecule binding data in a public database, such as BindingDB (Gilson et al., 2016), PubChem (Kim et al., 2021), or ChEMBL (Gaulton et al., 2012), or in an article or patent not yet curated by one of these databases. One may also check the Protein Data Bank (Berman et al., 2000; Berman and Gierasch, 2021) for a crystal structure of the target, which may assist in locating a suitable binding pocket for the small molecule to be developed as a drug. This is frequently true for metabolic enzymes and receptors that have evolved to bind small substrate and transmitter molecules. Many proteins belong to families, such as the protein kinases, whose members have similar properties (e.g., an ATP binding pocket), so that if one member of a family is druggable, then the others probably are also. In contrast, receptors for proteins often have large, relatively flat binding surfaces, rather than small binding pockets suitable for a small-molecule drug, and are thus less likely to be druggable and influenced by small molecules. Efforts are under way to systematically search for all druggable targets encoded by the human genome (Nguyen et al., 2017; Finan et al., 2017; Hopkins and Groom, 2002) and to gain traction against targets hitherto considered *undruggable* (Zheng et al., 2017).

The ultimate validation of a candidate target is the successful development of a novel drug that works by binding to it. Such a novel drug is termed *first-in-class*. A first-in-class drug is a true innovation and may represent a medical breakthrough, so one might expect first-in-class to be the goal of every drug discovery project. In fact, however, pharmaceutical companies often engage in less innovative, more predictable projects by developing *me-too drugs* against old targets that are already fully validated by a first-in-class drug. Such projects aim to improve on the first-in-class drug through, for example, greater potency, reduced side effects, or more convenient dosing (e.g., oral instead of intravenous), and ideally to produce a new drug considered *best-in-class*. For example, Merck's *lovastatin* broke ground as the first statin, the first in a class of drugs that lower cholesterol by inhibiting the enzyme HMG-CoA reductase (see Chapter 37); but other statins, such as *atorvastatin*, have also achieved enormous commercial success.

Beyond Single-Protein Drug Targets

A number of drugs, whether by accident or by design, hit multiple protein targets, a phenomenon termed *polypharmacology* (Peters, 2013). This phenomenon is particularly common when the target is a member of a family of proteins with similar binding sites. For example, the

specific chemical group of the protein target, often a relatively reactive amino acid side chain within an enzyme's catalytic site. In principle, covalent drugs should require smaller, less frequent dosing, because a covalently bound drug will not dissociate from the protein as the concentration of free drug dwindles over time following a dose (but note that some boron-containing compounds form *reversible* covalent bonds [Diaz and Yudin, 2017]). Drug developers have tended to avoid covalent drugs because they necessarily possess chemically reactive groups that risk reacting not only with the desired target but also with other proteins and biomolecules, with the potential for causing undesired biological effects. However, selectivity can be achieved by specific noncovalent interactions between the drug and the protein that pull the compound into a location and conformation where it is poised to form the desired covalent bond.

Covalent binding has been used to successfully target and inhibit a member of the RAS GTPase family, KRAS G12C, which had been viewed as virtually undruggable. As a result of such targeted positioning, the cancer drug *sotorasib* gains both potency and specificity by forming a covalent bond with a cysteine side chain present in an oncogenic mutant form of KRAS but not in normal KRAS (Lanman et al., 2020).

Experimental Approaches to Drug Discovery

Given a validated target, the next major milestone in a drug discovery project is arrival at a *clinical candidate*, a small molecule that binds the target with high affinity and specificity, has the desired effect on it, and meets a range of other criteria for a safe, efficacious drug (Hefti, 2008). Some of these criteria relate to *pharmacokinetics*: How well will the compound be absorbed if given orally? How well does it distribute to the targeted organs and tissues? How rapidly and by what mechanisms is it eliminated? Is it metabolized to an active metabolite? These properties are often lumped together as absorption, distribution, metabolism, and excretion (*ADME*) or drug metabolism and pharmacokinetics (*DMPK*).

It is also essential to confirm that the compound does not show evidence of toxicity. Both pharmacokinetics and toxicity can be initially studied *in vitro*. For example, there are *in vitro* methods that examine the ease with which the compound enters cells (see Chapter 4) and the likelihood that liver enzymes (see Chapter 5) will chemically modify the compound. Compounds also can be evaluated *in vitro* for evidence of toxicity and mutagenicity. However, *in vitro* studies cannot fully model the complexities of a living organism; animal studies are still required to minimize the chances that a compound will be problematic when first given to human subjects. For example, toxicity is usually assessed by long-term monitoring of the health of two species of animals, generally one rodent (usually mouse) and one nonrodent (often rabbit), when dosed with the compound. A good clinical candidate should also meet some nonbiological criteria. In particular, it must be amenable to large-scale synthesis and high-grade purification at acceptable cost, and it should be possible to create a formulation (e.g., a tablet or injection) that is sufficiently water soluble and stable.

Sophisticated technologies have been developed to speed the process of generating a clinical candidate. These mainly focus on the discovery or design of compounds that will bind the protein target with high affinity (*potent ligands*). Less progress has been made toward designing in safety and favorable pharmacokinetics. These properties pose more complex challenges, because they go far beyond how a small molecule and a protein interact with each other and instead involve the interactions of the small molecule with thousands of different biomolecules in a living system. The technologies for ligand discovery are both experimental and computational, and different methods are applicable in different settings. The following subsections touch on broad approaches but are not comprehensive. Note, too, that various approaches can be used in combination, so the distinctions made here are ultimately somewhat artificial.

Medicinal Chemistry

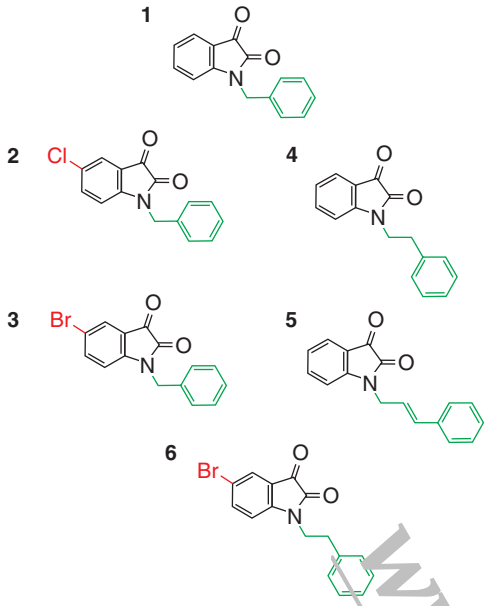
Synthetic organic chemistry remains at the heart of small molecule drug discovery, where it is specialized and known as medicinal chemistry. Medicinal chemists typically are part of a project team that includes, among others, biologists, assay specialists, and computational chemists; their role is to reduce chemical concepts to practice by synthesizing and purifying compounds that may ultimately lead to a new drug. In addition to providing the expertise needed to synthesize compounds of interest, they also help guide the design and selection of the compounds to be made. A key consideration is the complexity of a compound's synthesis, or "synthetic accessibility", which must be balanced against the level of interest in the compound. For example, it can be difficult to generate pure stereoisomers of compounds with multiple chiral carbon atoms, and certain chemical structures can by synthesized only via demanding, multi-step syntheses. A compound that is too difficult to make or purify will not only slow down the research effort but may also lead to a drug that is too costly to manufacture.

Medicinal chemists also inform the drug design process by providing insights into the properties of various chemical groups that might be incorporated into a drug, such as the attractive or repulsive interactions they may form with the targeted protein, their susceptibility to metabolic changes following administration, their potential to spontaneously form undesired covalent bonds with biomolecules, and their influence on the compound's ability to cross the blood-brain barrier (which may be desirable or undesirable, depending on the goal of the project). This expertise comes into play, for example, when a compound binds the target well but is rapidly metabolized by the liver into an inactive product. In this setting, the medicinal chemist may try substituting the part of the compound that is metabolized with a "bioisostere", a different chemical group with a similar shape and ability to interact with the protein but with reduced susceptibility to metabolic modification. More broadly, decades of experience have led to a number of rules of thumb for what makes a compound "drug-like", such as the "rule of five" (Lipinski, et al., 2001). These may be useful guides during drug discovery projects, but there are also many exceptions to the rules (Zhang et al., 2007).

High-Throughput Screening

If nothing is known about the structure of the target protein and what small molecules can bind it, it is common to turn to HTS, in which thousands to millions of compounds are tested using automation and robotics (Waldo et al., 2017). Tiny samples of each compound are drawn from a stored *chemical library* and deposited into multiwell plates for testing. Substantial effort often must be invested to devise an assay that works reliably in miniature and without user intervention. Most provide an optical readout, such as a change in luminescence, fluorescence, or color, as these can be efficiently measured with an optical plate reader. The compounds screened can range from part of the vast, in-house compound collection that a major pharmaceutical company has assembled over the years to a smaller set purchased from a commercial vendor. A screening library is often designed for the particular application. For example, one can purchase libraries tuned for activity against protein kinases, libraries with reactive groups that can form covalent bonds to the protein, and libraries designed to sample a wide range of compounds through high *chemical diversity*. A compound chosen at random from a screening library has a very low probability, typically 0.1% or less, of being active against a given target (Shun et al., 2011), and HTS measurements are subject to experimental error. Therefore, many of the compounds that appear active on an initial screen (*hit compounds*) are false positives, so careful data analysis and confirmatory testing are essential.

Even the confirmed hits from a high-throughput screen are far from being drugs. Their affinity for the target usually is orders of magnitude too weak, they may lack the desired specificity, and they do not meet DMPK or safety criteria. However, they offer an initial toehold on the challenge of finding a potent drug candidate. The next step is to purchase (*analogue by catalog*) and/or synthesize (*medicinal chemistry*) similar compounds that ultimately give a picture of how various changes in



Compound	ALDH1A1	ALDH2	ALDH3A1
1	0.02	82	7.7
2	0.06	2.1	16
3	0.58	2.1	69
4	0.07	3.5	0.45
5	0.07	>100	0.31
6	2.0	0.05	18

Figure 1–1 *Structure-activity relationship: scaffolds and substituents.* Five inhibitors of the aldhyde dehydrogenase family of enzymes have a common chemical scaffold (black) while having different chemical substituents at two positions (red, green). The table lists the IC_{50} (μM) of each compound for three members of the aldehyde dehydrogenase family of enzymes: ALDH1A1, ALDH2, and ALDH3A1; i.e., the concentration of compound needed to provide 50% inhibition of each enzyme. The lower the IC_{50}, the more potently the compound inhibits the enzyme. Focusing first on compounds **1**, **2**, and **3**, one can see that adding an increasingly bulky halogen atom (Cl, Br) on the six-membered ring tends to reduce the compound's potency against ALDH1A1 and ALDH3A1 but to increase it against ALDH2. Focusing next on compounds **1**, **4**, and **5**, one can see that adding increasingly bulky, nonpolar, aromatic substituents at the nitrogen modestly reduces the potency against ALDH1A1, initially improves but then destroys potency against ALDH2, and consistently improves potency against ALD3A1. Such patterns can guide the design of new compounds with desired potency and selectivity. For example, the substituents in compounds **3** and **4** each reduce potency against ALDH1A1 while increasing potency against ALDH2, so it is not surprising that compound **6**, which combines both substituents, has particularly low potency against ALDH1A1 and high potency against ALDH2. Note, however, that this kind of reasoning can only offer guidelines; its predictions are not always borne out by experiment. Data drawn from Kimble-Hill et al., 2014.

the chemical structure influence activity against the target (*structure-activity relationships, or SAR*) and other properties (Figure 1–1). This information is used to guide the synthesis of often hundreds of compounds with gradually improving properties. The most promising early molecules (*lead compounds*) serve as starting points for further improvement (*lead optimization*), ultimately generating, hopefully, a clinical candidate, potentially accompanied by several *backup compounds* in case the leading candidate fails.

Fragment-Based Drug Discovery

Even a large-scale screen can fail to provide useful hits (Keserü and Makara, 2009). This result becomes understandable when one recognizes that the number of stable, drug-sized, organic compounds is on the order of 10^{60} (Reymond et al., 2010), so a screen of even 10^6 compounds scarcely touches the vastness of *chemical space*. This vastness results from the combinatorial explosion of ways of connecting various chemical substructures, such as benzene rings, hydroxyl groups, and cycloalkanes. To be a good binder, a compound has to get multiple substructures positioned so they all form favorable interactions with complementary groups in the targeted binding pocket. If it has two chemical components suitable for binding the target but a third that is inappropriate or in the wrong place on the compound, it may fail to bind the target. This perspective motivates another method of discovering binders, *fragment-based drug discovery* (FBDD) (Erlanson, 2012; Lamoree and Hubbard, 2017). In FBDD, one conceptually breaks down drug-sized compounds into their substructures (*fragments*) and tests simple substructures against the target. Although such fragment-like molecules can bind only very weakly, such studies can, nonetheless, identify a small set of chemical substructures that are suitable for the target, and one can then buy or synthesize larger compounds assembled from these components. When either X-ray crystallography (Patel et al., 2014) or nuclear magnetic resonance spectroscopy (Shuker et al., 1996) is used to detect or analyze fragment binding, specific information is usually available about where each fragment binds the protein. This information can be used to stitch together designed compounds that place the appropriate fragments at the right places in the protein's binding pocket (*fragment linking*) or to optimize and expand one selected fragment (*fragment growing*). In this way, FBDD avoids the combinatorial explosion of possible compounds made from various chemical components and allows researchers to focus quickly on compounds made from only a productive subset of chemical components. The drug *vemurafenib*, which targets an oncogenic mutation of B-Raf kinase and was developed with a fragment-growing strategy, is usually referenced as the first FBDD success story (Bollag et al., 2012).

Emerging Experimental Technologies

The difficulty and cost of drug discovery, coupled with the market and human need for new medications, have driven ongoing innovation in drug discovery technologies. For example, DNA-encoded compound libraries (DELs) dramatically expand the number of compounds that can be tested, relative to conventional HTS (Halford, 2017). Unlike a traditional HTS compound library, where each compound is kept in its own separate container or well, a DEL is a mixture of compounds in a single container and can include far more compounds—into the billions and even trillions. Each unique compound in the mixture is covalently bound to a corresponding unique short DNA molecule, which serves as an identification tag. Such libraries can be synthesized and tagged with the methods of *combinatorial chemistry*, where a mixture of compounds is split into multiple portions, each portion is modified with a different chemical step and its DNA tags modified accordingly, and the portions are mixed again. This process is iterated until the synthesis is complete. To screen the DEL for active compounds, one may immobilize the target of interest on a solid surface, expose the surface to the DEL mixture, and then wash the surface to remove all the DEL compounds that have not bound tightly to the target. The binders are then removed from the target by more aggressive washing, and the active compounds in the wash are identified by sequencing the DNA tags they carry.