4
Pharmaceutical Drug Discovery: Designing the Blockbuster Drug

David Jesse Cummins

Twenty years ago, drug discovery was a somewhat plodding and scholastic endeavor; those days are gone. The intellectual challenges are greater than ever but the pace has changed. Although there are greater opportunities for therapeutic targets than ever before, the costs and risks are great and the increasingly competitive environment makes the pace of pharmaceutical drug hunting range from exciting to overwhelming. These changes are catalyzed by major changes to drug discovery processes through application of rapid parallel synthesis of large chemical libraries and high-throughput screening. These techniques result in huge volumes of data for use in decision making. Besides the size and complex nature of biological and chemical data sets and the many sources of data "noise", the needs of business produce many, often conflicting, decision criteria and constraints such as time, cost, and patent caveats. The drive is still to find potent and selective molecules but, in recent years, key aspects of drug discovery are being shifted to earlier in the process. Discovery scientists are now concerned with building molecules that have good stability but also reasonable properties of absorption into the bloodstream, distribution and binding to tissues, metabolism and excretion, low toxicity, and reasonable cost of production. These requirements result in a high-dimensional decision problem with conflicting criteria and limited resources. An overview of the broad range of issues and activities involved in pharmaceutical screening is given along with references for further reading.

1 Introduction

The pharmaceutical industry is rapidly approaching a crisis situation. Epidemics, such as AIDS, increasing rates of cancer, the threat of biological warfare agents, and an increasing elderly population, mean that the demand for useful therapeutic drugs is greater than ever. At the same time, pressure is increasing to reduce costs in the face of the daunting challenge of discovering and developing therapeutic agents. Approximately 50% of drugs in development fail due to safety issues and 25% fail due to efficacy issues. Most researchers estimate that the process of developing a new drug from early screening to drug store shelves costs \$600 to \$900 million and takes 8 to 15 years.

In this chapter, a *screen* refers to a biochemical test (or *assay*) to see if small molecules bind to a target. The usual sense of this term suggests an experiment performed on some (physical) experimental unit. There is a hierarchy of results: human clinical trials are the ultimate answer, which are approximated with animal testing, animal testing is approximated with in vitro testing (cell cultures, enzyme studies), and any of the above can be approximated *in silico* by the use of predictive models (a *virtual* screen).

Although the greatest expenses in drug discovery and development are incurred in the clinical trials phases, this chapter focuses on the early screening stage, before the first human dose. Well-planned studies at this stage have great potential to reduce expenses at later stages of the process. If it were possible to weed out the toxic molecules prior to the clinical trials phase, fully 40% of the expenses incurred in clinical trials would be eliminated! Even a small dent in this expensive process would result in enormous savings. If this could be done through virtual screens using predictive models, then additional savings would be achieved through less animal toxicity testing and this would also reduce the overall drug development time.

Drug discovery is a multidisciplinary endeavor with critical work at the interface of biology, chemistry, computer science, and informatics. In biology, a major activity is to make the linkages between what can be assayed and a disease response, but activities also include design and validation of animal models, cell cultures, biochemical screen design, and assay variability studies. Another important area in biology, the pace of which has especially intensified in the last decade, is the assessment in vivo of the extent of absorption into the blood stream, distribution and binding to tissues, and the rates of metabolism and excretion. This is denoted by *ADME* and is discussed in Section 11. In chemistry, the major responsibility is to provide the creative spark to navigate effectively the large space of possible compounds (another word for molecules) towards the blockbuster drug. Other activities include synthesis of new molecules, analytical characterization of existing molecules (including purity of batches, pKa, logP, melting point, and solubility) and construction of libraries. Important issues in computer science include data storage and extraction, implementation and scale-up of algorithms, management of biological and chemical databases, and software support. Activities in informatics or chemoinformatics (Leach, 2003) include design of experiments, development of new chemical descriptors, simulation, statistical analysis, mathematical modeling, molecular modeling, and the development of machine learning algorithms.

The successful development of a new drug depends on a number of criteria. Most importantly, the drug should show a substantial beneficial effect in the treatment of a particular disease (*efficacy*). This implies that, in addition to intrinsic activity, the drug is able to reach its *target* (a biological gateway that is linked to a disease state—a large organic molecule that may be a receptor, a protein, or an enzyme) and does not produce overwhelming toxic effects. Many active drugs fail in later phases of the development process because they do not reach their intended target.

The main challenges in drug discovery fall into four categories:

1. Potency: the drug must have the desired effect, in the desired time frame, at a low dosage.

2. Selectivity: the drug should produce only the desired activity and not cause side effects. There are so many possible targets in the body that achieving high selectivity is difficult. Side effects may be caused by metabolites of the drug, by-products produced when the body uses enzymes to break down the drug in the elimination process (Section 11.1).
3. ADME and pharmacokinetics (or *PK*): the drug must reach the site of action. If taken orally, it must be water soluble, survive the stomach, be absorbed through the intestine, survive attack by many enzymes, and be transported into the target cells across the cell membrane. It must not be metabolised too quickly, but also must not be so stable or protein bound that it accumulates in the body. Another important factor to control is whether a compound crosses the blood–brain barrier (or BBB).
4. Toxicity: there are many mechanisms by which a compound can be toxic. Toxicity issues may arise from PK or ADME or selectivity issues, depending on the mechanism. Alternatively a compound may simply react harmfully with tissues or organs in a direct manner.

One may think of an iterative model for the preclinical discovery screening cycle. A large number of compounds are to be mined for compounds that are *active*; for example, that bind to a particular target. The compounds may come from different sources such as vendor catalogues, corporate collections, or combinatorial chemistry projects. In fact, the compounds need only to exist in a virtual sense, because in silico predictions in the form of a model can be made in a *virtual screen* (Section 8) which can then be used to decide which compounds should be physically made and tested. A mapping from the structure space of compounds to the *descriptor space* or *property space* provides covariates or explanatory variables that can be used to build predictive models. These models can help in the selection process, where a subset of available molecules is chosen for the biological screen. The experimental results of the biological screen (actives and inactives, or numeric potency values) are then used to learn more about the *structure–activity relationship* (*SAR*) which leads to new models and a new selection of compounds as the cycle renews.

The relationship between the biological responses and the changes in chemical structural motifs is called SAR or QSAR (*quantitative structure–activity relationship*). Small changes to the chemical structure can often produce dramatic changes in the biological response; when this happens, chemists and biologists will often describe the SAR as *nonlinear*, by which they mean that the SAR has a "sensitive" or "rough" or "unstable" response surface. Often the chemical compounds are considered to be the experimental unit even though, in actual experiments, an animal or cell culture is the unit and the compound is a treatment. This is because the important asset to the pharmaceutical company is the compound. The vast size of the set of potential experimental units (potential compounds), coupled with the high dimensionality of the response being optimized (potency, selectivity, toxicity, and ADME) and the "roughness" of the response landscape make drug discovery a challenging arena. The level of noise in biological data can be extremely high as well.

This chapter covers a selection of problems and case study examples. Perspectives specific to Eli Lilly and Company (or Lilly) are distinguished from broader perspectives believed to be shared by most companies across the industry. The chapter covers both design and analysis issues, and touches on topics such as simulation, computer experiments, and pooling. Section 2 gives an overview of drug design. In Section 3 the issue of false negatives and false positives in drug screening is addressed. Molecular diversity is discussed in Section 4, and machine learning is the topic of Section 6. Section 7 describes a lower-throughput iterative approach to screening and virtual screening in drug discovery projects in an iterative *Active Learning* strategy. A brief mention of pooling strategies is made in Section 9 and Section 10 discusses expectations for rare events. Section 11 describes aspects of a molecule that determine its ability to be safely transported to the area of the body where it can be of therapeutic benefit. Finally, in Section 12 the problem of multicriteria decision making in drug discovery is addressed.

2 Overview of Drug Design

2.1 Process Overview

The entire process of drug discovery and development can be depicted as a rocketship with stages, an image that portrays the “funneling” effect as fewer compounds are under consideration at successive stages of the process. The focus of this chapter is screening issues in lead generation (the first stage of the rocket) which begins with the chemical entity and the biological target. The chemical entity (compound) may be a small molecule, a peptide, or a large protein. Typically, chemical entities are purchased from external providers or synthesized within a company. The compound can be viewed as binding or docking to the biological receptor or target in order to competitively inhibit, or else to induce, some biological signal such as the production of a protein or hormone, resulting in a specific response. The number of molecules tested is dependent on reagent costs and other practical factors. This chapter adopts the following paradigm for drug discovery and development.

1. A target is validated to establish a direct link (such as a gene or a process in the body, or a virus or a parasite) to the disease state of interest and the feasibility of controlling the target to obtain the desired therapeutic benefit. This stage involves scientific study that can be catalyzed by genomic and proteomic technologies. Target validation requires careful scientific experiments designed to explore how a target influences a biological response or disease state.
2. A high-throughput screen (HTS) is designed, optimized, calibrated, validated, and run to obtain biological response data at a single concentration for

200,000 compounds (Section 4). This may involve whole cells, enzymes, or other in vitro targets. Reducing variability is crucial. Sittampalam et al. (1997) introduced the concept of *signal window*, a method whereby two controls are used to diagnose the ability of the assay to distinguish between actives and inactives in the presence of background noise.

3. The most promising compounds, or *actives*, typically numbering from 1000 to 5000, are then tested in a *secondary* screen which involves testing each compound at 5 to 10 different concentrations. These results are modeled with a nonlinear dose–response curve and for each molecule a summary measure is computed such as a 50% inhibitory concentration (IC50) or a 50% efficacious concentration (EC50).
4. The secondary assay reduces the set of actives to those for which potency reaches at least 50% of the maximum potency of a reference compound, at some concentration. Typically there are 500 to 1000 of these compounds, and they are called *hits*. Many hits may be nonspecific or for other reasons may offer no prospect for future development. (In subsequent sections the distinction between *active* and *hit* is blurred.)
5. The hits are examined in a series of careful studies in an effort often called *hit to lead*. Chemists look at the hits and classify them into four to eight broad series and, within each series, they try to find a structure–activity relationship. The chemists characterize these SARs by testing (in the secondary assay) a few hundred or a few thousand more molecules, thus expanding each SAR. Out of these SARs, the chemists and biologists choose a few hundred compounds to be tested in cell-based or enzyme in vitro screens. These screens require careful design and validation. From the molecules run through the in vitro testing, 100 or so may go through in vivo single-dose tests using a rodent or some other animal model. Some 10 to 40 of these molecules are finally tested for in vivo efficacy in a full dose–response experiment performed on the animal of choice.
6. The lead compounds undergo careful studies in an effort known as *lead optimization*. At this point any remaining issues with metabolism, absorption, toxicity, and so on, are addressed through molecular modification.

Some research groups contend that the HTS step should be eliminated and replaced with a number of rounds of iterative medium-throughput screening (Section 7). It is an issue of quantity versus quality. The lower-throughput screens tend to have lower variability ("noise") and less dependence on the single concentration test as an initial triage. The iterative approach is closely akin to a strategy in machine learning (Section 6) known as *Active Learning*.

In a successful project, the steps outlined above will lead to a *First Human Dose* (FHD) clinical trial. How well those prior steps are done will, in part, determine the success or failure of the human clinical trials. The adoption of high-throughput screening and combinatorial chemistry methods in the early 1990s offered promise that a shotgun approach to drug discovery would be possible. It was soon learned that simply increasing the volume of screening results cannot be the answer. The number of potential chemical entities is staggering, being estimated to be between

10^{20} and 10^{60}. The efficient exploration of a landscape of this magnitude requires prudent use of machinery, human expertise, informatics, and, even then, an element of fortuity. On average, for every 5000 compounds that enter a hit-to-lead phase, only five will continue on to clinical trials in humans, and only one will be approved for marketing. It is analogous to searching for a small needle in a whole field of haystacks. In these terms, the future of drug design lies in no longer searching for the needle but, instead, constructing the needle using available clues.

3 False Negatives and False Positives

In primary screening, compounds are tested at a single concentration; those whose response exceeds a prespecified threshold are labeled as "active" and the rest as "inactive". Typically, 200,000 compounds are screened, giving numeric potency results for each, then, based on exceeding a threshold, about 2000 are labeled as active and 198,000 as inactive. The actives are studied further at multiple concentrations and the inactives are henceforth ignored. A *false positive* error occurs when a compound labeled as active is, in fact, inactive when studied in the more careful multiple concentration assay. The false positive rate can be lowered by raising the decision threshold, or "hit limit", but at the cost of increasing the false negative error rate. In most HTS screens, of those compounds flagged as active in a primary screen, roughly 30% to 50% are found to be inactive in the multiple concentration–response follow-up study.

A *false negative* error occurs when a compound that is actually active is not labeled as active. Biological noise, for example, and the choice of hit threshold can affect the false negative error, as well as mechanical failures such as a leaking well. Mechanical failure errors are unrelated to the true potency of the molecule. The false negative rate is unknowable because the vast majority of compounds are not studied further, but it can be estimated from small studies. From past HTS screens at Lilly, we have estimated that a mechanical failure false negative occurs in roughly 7% to 12% of compounds in an HTS screen, with a total false negative error rate ranging from 20% to 30%. Aside from the mechanistic type of false negative, the false negative rate can be viewed as a function of the activity level—the greater the activity of the molecule, the lower the chance of a false negative error.

Experimental results from an HTS assay are not the "truth" but merely an estimate of the true potency of a molecule. Because molecules are only measured one time in the HTS setting, there is a high degree of uncertainty. One thing that can be done is to look for highly similar molecules and treat them as pseudo replicates of the same "parent" molecule. See Goldberg (1978) for further discussion on estimating the error rate.

One effective way of dealing with experimental errors is to build a predictive model and score the screening results through the model, then to look at discrepancies between the experimental screening result and the model prediction. Often the highly potent but mechanical failure type of false negatives or false positives

TABLE 1. Breakdown of hit rates from the screening follow-up results.

Compound source	Hit rate
150,000 original compounds	3%
2050 new compounds	34%
250 potential false negatives	55%

can be identified. In practice, the false positives will be tested in the secondary screen and found to be inactive when that screening result is observed. Some researchers favor raising the threshold to reduce the number of compounds labeled as active. The false negatives can then be identified by a statistical or predictive model and rescreened. In one recent project at Lilly we followed up a 150,000 compounds HTS with a small library (that is, a small collection of molecules) of 2300 compounds. A predictive model was trained (or fitted) using the 150,000 primary results and used to select 2050 molecules that had not yet been tested. The same model was used to identify 250 molecules that were screened in the 150,000 and found to be inactive, yet scored by the model as highly active. These 250 were the potential false negatives that were to be retested. Fully 55% of these false negatives were active upon retesting. The breakdown of hit rates is given in Table 1.

The 3% hit rate in the primary screen was a concern, as such a high number suggests a problem in the assay. It was found that there was a carryover problem in which sticky compounds were not being completely washed from the robotic tips. Such a trend can be found easily by analysis of hit rate as a function of well location. This problem was resolved before the secondary runs (rows 2 to 3 of the table) were made.

A computer experiment was done to confirm and further explore the above findings. An iterative medium-throughput screening operation was simulated with different levels of false negative rates, reflecting historical error rates seen across past screens. For each level of false negative rate, the appropriate proportion of true actives was randomly chosen (from a nonuniform distribution that is a function of the "true" activity level of the molecule, based on historical data) and relabeled (incorrectly) as inactive. The model was trained on this "polluted" data set and used to select the set of compounds for the next round of testing. Computer experiments of this type can be run many times to explore the behavior of the predictive models under realistically stressed circumstances. For this particular experiment, the model was able to find 25 times more false negatives than a random (hypergeometric) search would produce, up to a false negative rate of 30% at which point the enrichment over random decreases from 25-fold to about 15-fold higher than random.

In both screening and predictive modeling, the relative cost of false negatives versus false positives is an important component of decision making. From a business standpoint, false negatives represent lost opportunities and false positives represent wasted efforts chasing down "red herrings." The resources wasted chasing false positives is particularly troubling. The current trend is to reduce the false positives and to tolerate the increased number of false negatives that results.

4 Molecular Diversity Analysis in Drug Discovery

In the last two decades, three technologies have been co-developed that enable a significant shift in the process of lead generation:

Combinatorial Chemistry ⇒ Large libraries of molecules
High-Throughput Screening ⇒ Many biological data points
Cheminformatics ⇒ Many molecular descriptors

The adoption of high-throughput screening and combinatorial chemistry methods in the early 1990s led to an immense interest in molecular diversity. It was widely expected that making diverse libraries would provide an increase in the number of hits in biological assays. It took a while to realize that this was the wrong expectation. Molecular diversity designs do offer great benefits, but more in the enhancement of the quality, rather than quantity, of information from a screen. It became clear that other properties of molecules, beyond mere structural novelty, need to be considered in screening. This led to extensive work on "druglikeness" and an attempt to achieve a balance between diversity and medicinal reasonableness of molecules.

4.1 Molecular Diversity in Screening

Molecular diversity analysis is useful in several contexts:

- Compound acquisition: this avoids purchasing a compound very similar to one that is already owned.
- General screening for lead identification: screening a diverse library is a sensible approach when little or nothing is known about the target or possible lead compounds.
- Driving an SAR effort away from prior patent claims.

The second context, general screening, involves selecting subsets of molecules for lead generation. Experimental designs are considered because it is not feasible to screen all molecules available. Even with the application of high-throughput screening, the demand for screening outpaces the capacity. This is due to the growth of in-house chemical databases, the number of molecules synthesized using combinatorial chemistry, and the increasing number of biological targets fueling discovery projects. In addition, novel screens that are not amenable to HTS automation may be attractive from a competitive standpoint but the gap between screening capacity and screening opportunities in this case is particularly daunting. Given this imbalance, methods for selecting finite subsets of molecules from potentially large chemical databases must be considered. Possible selection

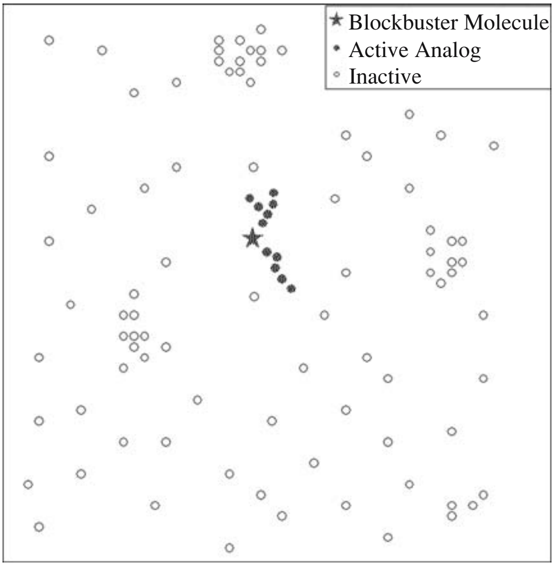

FIGURE 1. SAR paradigm. Fictitious two-dimensional projection of the property space of feasible druglike molecules.

strategies include: random, diverse, and representative selection, each of which may be performed as a biased or directed analysis if information such as a drug-likeness score is available to weight the analysis in favor of certain classes of molecules. A requirement of every selection method considered here is its computational feasibility for the databases of hundreds of thousands to millions of compounds that are now common with the application of combinatorial synthesis. For example, many distance-based selection strategies involve computation and storage of all pairwise distances for molecules in a database. If the number of molecules n is 300,000, then there are approximately 45×10^9 (calculated from $\binom{n}{2}$) distances to compute and/or to store. This is a formidable task, necessitating creative computational solutions.

Figure 1 illustrates a modern paradigm of drug hunting processes. In this fictitious two-dimensional projection of the space of feasible druglike molecules, open circles represent compounds that are not active relative to a specific target, solid circles are active compounds, shown in one contiguous series of related molecules, and the star is the blockbuster drug that is still undiscovered. In a primary screen, finding a single solid circle is all that is needed. The medicinal chemistry teams can follow up by making systematic changes to any one of the active compounds to explore the whole series and find (or invent) the blockbuster drug. An important point is that the blockbuster may not exist in the corporate collection. A typical lead generation or lead optimization project involves not only testing molecules in current libraries, but also synthesis of new molecules. Molecular modification and subsequent testing is the way the trail gets blazed, through characterizing the SAR.

In recent years there has been strong dogma contending that, in filling a fixed screening capacity, it is important to screen “backups”, that is, molecules that are closely related. This argument is motivated by the high rate of false negatives in primary screening. Thus screening two or more compounds from the same related series effectively gives pseudo replicates. If one compound turns out to be a false negative, it is likely that another from the same series will screen as positive and thus, the active series will not be missed. This rationale is popular in the industry. However, at Lilly we have demonstrated, through both simulations and retrospective analysis, that it is better to tolerate the false negatives in favor of sampling a larger number of different series. The motivating principle for this position is that testing two closely related compounds (or *analogues*) is often equivalent to testing the same hypothesis twice, which comes at the expense of testing a different hypothesis; see Wikel and Higgs (1997).

Optimizing molecular diversity has the potential to maximize the information gained about an SAR in the early stages of screening. Suppose a random screening gives the same number of hits as a diverse screening. Then one would favor the diverse set of hits, because this increases the chance of at least one structural lead with a favorable ADME and toxicity profile. In fact, for primary screening, it is often better to have 10 novel hits than 200 hits that are analogues of each other. The proper focus, in our view, is quality of information gleaned from the screen.

Most pharmaceutical companies have clusters of compounds (for example, Lilly has many SSRIs, cephalosporins, and so on). There are many analogues clustered tightly in local subregions of chemical space, reflecting historical SARs investigated around related targets. A random sample will reflect the densities of the compound classes in the collection; thus testing a random sample of molecules for a certain biological activity will be equivalent to testing the same hypothesis many times over.

4.2 Descriptors

Computationally, a *structure space* (represented as a set of two-dimensional graphs of molecule structures) is mapped to *property (or chemical) space* ($\mathcal{R}^p$) (for example, Figure 2), where each point is a vector of values of each of p variables,

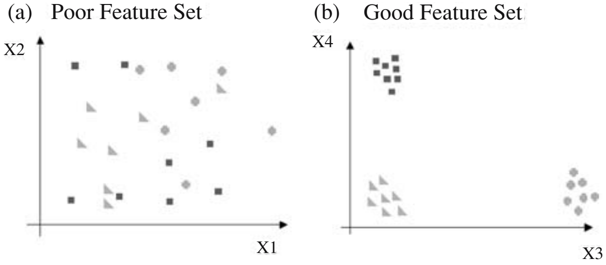

FIGURE 2. Descriptor validation example: (a) poor feature set; (b) good feature set.

called *descriptors*, or sometimes *properties*. The descriptors can be binary, integer counts, or continuous variables. A descriptor may be as simple as a count of the number of oxygen atoms in a molecule, or as sophisticated as an estimate of the three-dimensional polar surface area around the molecule. Molecules are assigned positions in this high-dimensional descriptor space through their properties. The relationships defining their molecular diversity are, therefore, represented through their coordinates or positions in this space. The distance metrics most often used are Euclidean and Mahalanobis for properties, and Tanimoto (Jaccard) for binary bit strings; see Section 4.5.

Prior to selecting a set of molecules from a database, it is often necessary to preprocess the molecular descriptors to replace missing descriptor values and to scale the descriptors. Although it is possible to develop distance metrics that are tolerant to missing values, at Lilly we have focused on imputing (replacing) missing values and using distance metrics that assume all descriptor values are present.

A set of molecules is commonly described with anywhere from 4 to 10,000 descriptors. It is also possible to represent molecules with sparse descriptors numbering up to 2 million. Variable selection, or descriptor subset selection, or descriptor validation, is important, whether the context is supervised or unsupervised learning (Section 6).

4.3 Molecule Selection

Discussions about molecular diversity involve the concepts of "similarity" and "dissimilarity" and may be confusing as their meanings are content related. Similarity is in the eye of the beholder. Chemists may find similarity hard to define, but they generally are quick to identify it when they see it and at times are willing to debate the similarity of one structure to another. Similarity is not absolute, but relative to the space in which it is defined. In chemistry, this means the definitions must always be held in context to the property space used to define the structures.

If characteristics are known about the biological target then this target-specific information may be used to select a subset of molecules for biological testing. Various database searching methods, molecular similarity methods, and molecular modeling methods could be used to identify a favored (or biased) subset of molecules for biological testing. This corresponds to the second row of Table 2. One example of this situation is the case of neuroscience targets. If very little

TABLE 2. Subset selection strategies for primary screening at Lilly.

Situation	Strategy
No target information	Diversity Selection
Expert judgment or literature information about target	Directed Diversity Selection
Experimental results related to target, or structure of target known	QSAR, Predictive Modeling

is known except that the receptor of interest is in the brain, a biased diversity selection would be more useful than an unbiased one. For example, one might construct a weight function based on the number of positively charged nitrogen atoms in a molecule, because this is often observed to be present in desirable neuroscience drugs. If there are no positively charged nitrogen atoms, or if there are more than two, the weight function is very low and otherwise very high. Other factors related to toxicity, solubility, and other aspects of medicinal viability of molecules could be included in the weight function. Then a weighted diversity selection could be performed to construct a reasonable starting set for initial screening.

4.4 Descriptor Validation and Variable Selection

The concept of molecular similarity is strongly linked with the "SAR Hypothesis" of Alexander Crum-Brown (Crum-Brown and Fraser, 1869) which states that compounds that are similar in their structure will, on average, tend to display similar biological activity. A modest extension holds that one can build mathematical models from the numerical descriptors to describe a relationship between the chemical structure and the biological activity. When chemists discuss similarity of two molecules, they often make arguments about the biological effects or binding potential of the compounds. There is a concept of *bioisostere* which says that some chemical fragments function in the same way as other chemical fragments (for example, a sulfur may behave like a methyl group). An ideal set of molecular descriptors would be one that contains properties characterizing all aspects important to potency, selectivity, ADME, and toxicity. Because our understanding of any one of these processes is limited, expert judgment is needed. Descriptors considered generally important include those describing lipophilicity, molecular shape, surface area and size, electronic properties, and pharmacophoric elements such as hydrogen bond donors and acceptors.

Just as with variable subset selection in linear regression, there are risks akin to over-fitting a training set. (A *training set* is the subset of data used to fit the model.) The topic of how best to do descriptor validation has been hotly debated, and numerous ideas have been proposed, but the general goal is to select that subset of descriptors that best achieves some sense of separation of the classes of compounds, as illustrated in Figure 2. This figure illustrates three structural series in two hypothetical two-dimensional configurations. Descriptors x_3 and x_4 are more useful because they separate the different structural classes.

There are dimensionality issues. Later we propose Mahalanobis distance (Section 4.5) as a good metric for diversity analysis. With p descriptors in the data set, this metric effectively, if not explicitly, computes a covariance matrix with $\binom{p}{2}$ parameters. In order to obtain accurate estimates of the elements of the covariance matrix, one rule of thumb is that at least five observations per parameter should be made. This suggests that a data set with n observations can only investigate approximately $\sqrt{2n/5}$ descriptors for the Mahalanobis distance computation. Thus, some method for subset selection of descriptors is needed.

In the case of molecular diversity, there is no response to guide the variable subset selection (unsupervised learning) and hence creative ways to do subset selection are needed. Ideally, chemical similarity is defined by the target or receptor. If one has information about the target/receptor then it is more useful to do QSAR (last row of Table 2). If such information is lacking, then one must impose an arbitrary definition on chemical similarity in order to avoid testing duplicate, or very similar, hypotheses in screening. Thus, at Lilly, our molecular diversity tools are generic and depend on a generic notion of similarity that is relatively independent of biology (rows 2 and 3 of Table 2).

4.5 Distance Metrics

Distance-based methods require a definition of molecular similarity (or distance) in order to be able to select subsets of molecules that are maximally diverse with respect to each other or to select a subset that is representative of a larger chemical database. Ideally, to select a diverse subset of size *k*, all possible subsets of size *k* would be examined and a diversity measure of a subset (for example, average near neighbor similarity) could be used to select the most diverse subset. Unfortunately, this approach suffers from a combinatoric explosion in the number of subsets that must be examined and more computationally feasible approximations must be considered, a few of which are presented below.

Given two molecules *a* and *b*, let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote their vectors of descriptors. The Mahalanobis distance between *a* and *b* is defined as:

$$d(a, b) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{V}^{-1} (\boldsymbol{x} - \boldsymbol{y})},$$

where $\boldsymbol{V}^{-1}$ denotes the inverse of the covariance matrix, $\boldsymbol{V}$, of the vectors of the descriptor values of all the molecules. If $\boldsymbol{V} = \boldsymbol{I}$ the result is Euclidean distance:

$$d(a, b) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2},$$

where x_i and y_i are the *i*th elements of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

The effect of the $\boldsymbol{V}^{-1}$ is to divide each descriptor by its standard deviation, so that some descriptors do not dominate others due to mere differences of scale. Many cheminformaticians compute the standard deviations explicitly, but this alone is not sufficient. The off-diagonal elements of the inverse covariance matrix adjust for overweighting (due to high correlations between descriptors) of latent aspects of a molecule, such as size.

A common practice is to scale each descriptor to have standard deviation of 1. Another is to compute principal components and confine the analysis to the first *h* components, where *h* may range from 1 to 20. This is an ad hoc form of dimension reduction that does not remove irrelevant information from the analysis. At Lilly, we prefer a careful descriptor validation to avoid including many irrelevant descriptors into the analysis, combined with a dimension reduction criterion using the $\sqrt{2n/5}$ rule of thumb, followed by a Mahalanobis distance computation using all the descriptors that remain.

For presence or absence of features in the molecules, represented by binary bit strings $\boldsymbol{x}$ and $\boldsymbol{y}$ as descriptors, the Tanimoto coefficient is a popular metric for similarity:

$$sim(a, b) = \frac{(\text{bits on in both } \boldsymbol{x} \text{ and } \boldsymbol{y})}{(\text{bits on in } \boldsymbol{x}) + (\text{bits on in } \boldsymbol{y}) - (\text{bits on in both } \boldsymbol{x} \text{ and } \boldsymbol{y})}.$$

Then

$$d(a, b) = 1 - sim(a, b).$$

Now consider $d(a, b)$ to be a generic distance metric of which Tanimoto, Euclidean, and Mahalanobis are three cases. Then, the distance between molecule a and the set of molecules B is defined as follows,

$$d(a, B) = \overset{\min}{_{b \in B}}\, d(a, b),$$

and the overall dissimilarity of a set of molecules M is defined as

$$dis(M) = \frac{1}{n} \sum_{a \in M} d(a, M \backslash a), \tag{1}$$

where $M \backslash a$ denotes the set M with the molecule a removed.

These metrics are used by design algorithms for selecting dissimilar molecules for chemical analysis (see Section 5.2).

5 Subset Selection Strategies

A requirement for any subset selection method is the ability to accommodate a set of previously selected molecules, where augmentation of the pre-existing set is desired. For example, when purchasing compounds, the goal is to augment what is already owned so that the current corporate collection would be used in the analysis as the pre-existing set of molecules. The goal then is to select a subset of the candidate molecules that optimizes a specified criterion with reference to the molecules in both the candidate set and the previously selected set.

In the case of iterative medium-throughput screening, at any given point in the process, the set of molecules that have been screened thus far is the previously selected set for the next round of screening. In choosing molecules for the next iteration, one may have a selection criterion such as predictive model scores but a diversity criterion may also be applied: it is not desirable to screen something identical, or nearly identical, to that which was screened in previous rounds.

There are two main strategies developed to select diverse and representative subsets of molecules, namely, cell-based methods and distance-based methods.

5.1 Cell-Based Methods

Cell-based methods divide the space defined by a set of molecular descriptors into a finite set of "bins" or "buckets". Each molecule is then assigned to one of the bins. Structurally similar molecules will occupy the same or adjacent bins and dissimilar molecules will occupy bins that are spatially well separated. A diverse subset of molecules can be identified by selecting a single molecule from each of the occupied bins. Databases can be compared by examining the occupancy of bins with molecules from different sources. For example, commercial databases such as Comprehensive Medicinal Chemistry (2003), World Drug Index (2002), and Maccs Drug Data Report (2003) contain molecules that can be used to define the historically medicinally active volume (bins) of chemical space. Compounds in another database, or collection, that fall within the bins defined by these databases can then be selected for biological testing.

Cell-based methods have the advantage that they are intuitive and computationally more efficient than many distance-based methods. However, cell-based methods suffer from a problem known as the "curse of dimensionality." Consider a database with each molecule described by 20 molecular descriptors. Subdividing each molecular descriptor into merely 5 segments (or bins) will result in 5^{20}, or approximately 10^{14} bins. Even with large chemical databases, most of the bins will be empty and many bins will contain a single molecule. Outliers wreak havoc. Just one molecule whose molecular descriptors take on extreme values will cause the majority of molecules to be allocated to a small number of bins. In either case, a cell-based method will present problems in selecting a diverse subset of molecules. Thus, cell-based methods require a significant reduction in dimensionality from the many possible molecular descriptors, attention to outliers, and careful consideration of how to subdivide each dimension. An application to drug discovery screening, which addressed the issues of outliers and dimensionality, was applied to large databases by Cummins et al. (1996).

5.2 Distance-Based Methods

Statistics has a long-standing role in design of experiments. There is a long history of the use of information optimal designs (for example, D-optimal designs), which consist of the most informative points and are useful in designed experiments where the "true" model is known. *Space filling* designs are used in numerous contexts including geographical modeling (literal space filling), modeling response surfaces, multivariate interpolation, and chemical library design.

A more in-depth discussion of three selection methods that are computationally feasible with very large chemical databases is now given to highlight the issues that must be considered when applying many of these molecular diversity selection methods. The three design methods described here are edge, spread, and coverage designs. Each design method optimizes a specific objective.

- Edge design objective: obtain minimum variance estimates of parameters in a linear model.
- Coverage design objective: select a subset of molecules that is most representative of the entire library. Heuristically, the distance from the chosen subset to the remaining candidate points should be small. One might imagine a set of umbrellas positioned to cover as many candidate points as possible.
- Spread design objective: select the maximally dissimilar subset of molecules. This requires maximizing the distance of points within the subset from each other. One analogy for this is electron repulsion.

Edge designs are often constructed using D-optimal design algorithms. Molecules selected using D-optimal designs populate the edge of descriptor space by first filling in the corners and then moving around the boundary. Edge designs are appropriate when one intends to fit a linear regression model where the descriptors are the predictors in the model, for example, if biological activity is modeled as a function of the descriptors. This is usually the situation in lead optimization, rather than lead generation.

Spread and coverage designs are space-filling designs. Let C be the *candidate set*, that is, the set of possible design points. Once the criterion (*space filling*) is well defined, selecting the points $M \subset C$ to be space filling is simply an optimization problem.

The objective of a spread design is to identify a subset of molecules in which the molecules are as dissimilar as possible under a given similarity metric. For a given metric to measure the similarity of a subset, all subsets of size k (plus any molecules previously selected) could be evaluated and the subset that produces the lowest similarity measure chosen. In practice, simple non-optimal sequential algorithms are often used to approximate the maximally dissimilar subset: two such algorithms are described below.

1. Maximum Spread Algorithm

The goal: out of all $\binom{n}{k}$ subsets of k molecules from a candidate set C, find the subset M^* where $dis(M^*)$, defined in (1), is largest. The problem is that it is not feasible to enumerate and evaluate all possible subsets. The solution is to use a sequential approximation (greedy algorithm).

a. Select the first compound from the edge of the design space.
b. Select the second compound to be most distant from the first.
c. Select subsequent compounds in order to maximize the minimum distance to all previously selected compounds.

This is the algorithm proposed by Kennard and Stone (1969). At Lilly we have focused on an efficient implementation of this approach applied to large chemical databases and have not implemented design optimization due to the marginal

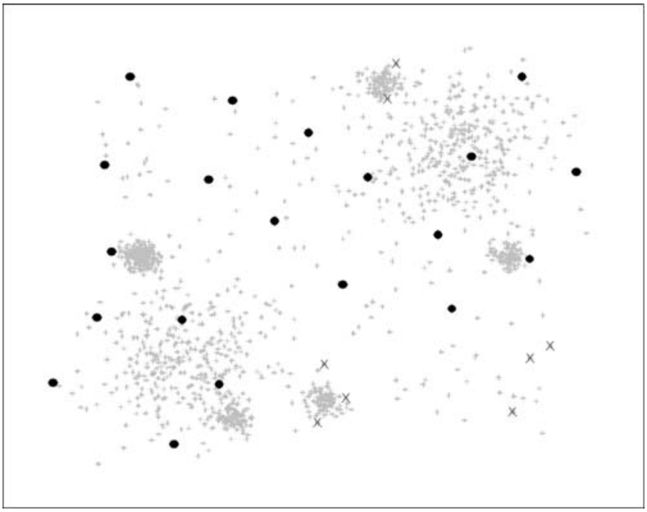

FIGURE 3. Spread design for fictitious example. (Reproduced from Higgs et al., 1997 with permission.)

design improvements and increased computational time. To illustrate, the SAS (2003) OPTEX (*CRITERION* $= S$) procedure was used to select 20 points from the 1400 two-dimensional points shown in Figure 3 using a modified Fedorov optimization algorithm (Cook and Nachtsheim (1982)). The OPTEX procedure seeks to maximize the harmonic mean distance from each design point to all other design points. Eighty different designs were generated using the sequential method of Kennard and Stone and compared with those obtained by the modified Fedorov optimization method. On average, the Fedorov optimization generated a design that was 8.5% better than that obtained from the simple sequential selection method but required eight times more computational time. In larger data sets of 200,000 or more compounds this can mean a choice of eight hours versus three days to find a design.

2. Maximum Coverage Algorithm

Define the coverage of a set M, where $M \subset C$ as:

$$\text{cov}(M) = \frac{1}{n} \sum_{\alpha \in M} d(a, C \backslash M),$$

where $C \backslash M$ is the set C with the set M removed. The goal: out of all $\binom{n}{k}$ subsets of k molecules with descriptor vectors in C, find the subset M^* where $\text{cov}(M^*)$ is smallest. This is often approximated using cluster analysis (see Zemroch, 1986).

In Section 5.3 the different design types are compared.

5.3 Graphical Comparison of Design Types

Figures 3–5 show a fictitious two-dimensional data set reproduced from Higgs et al. (1997) with permission. The data set contains 1400 hypothetical molecules and is constructed to illustrate the differences between edge, spread, and coverage designs. The data set was constructed to have five tightly packed clusters (bivariate normal), two loosely packed clusters (bivariate normal), and molecules uniformly distributed over the two-dimensional design space. For illustrative purposes, eight molecules were randomly chosen and labeled with an "X" as having been selected in a previous design. Future selections should complement these eight molecules. The data were simulated in two dimensions to depict how a pharmaceutical compound collection might appear in some two-dimensional projection. Certain regions are sparse with low density whereas other regions are highly clustered, reflecting the synthetic legacy of the company.

Figure 4 shows 20 molecules selected using an edge (D-optimal) design to augment the previously selected molecules. Two quadratic terms and one linear interaction term were included in the model used to select this design in order to force some interior points into the selection. Figure 5 shows 20 molecules selected using a *k*-means clustering approximation to a coverage design to augment the previously selected molecules. Figure 3 shows 20 molecules selected using the Kennard and Stone approximation to a spread design (see, for example, Johnson et al., 1990) to augment the previously selected molecules.

Although not shown in the figures, a random selection is often considered the baseline method of subset selection. Random sampling typically selects many molecules from the dense clusters, and several molecules near the previously selected molecules. Spread designs select the most diverse subset of molecules

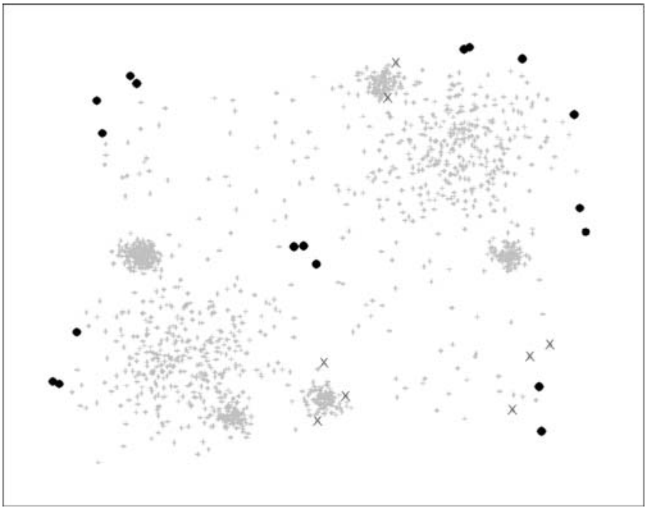

FIGURE 4. Edge design (D-optimal with interactions) for fictitious example. (Reproduced from Higgs et al., 1997 with permission.)

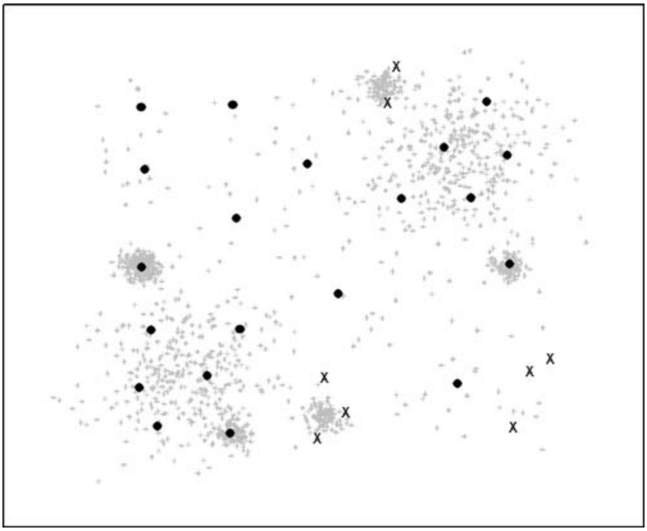

FIGURE 5. Coverage design for fictitious example. (Reproduced from Higgs et al., 1997 with permission.)

(relative to the other methods presented here), including molecules near the edges as well as throughout the design space. Spread designs ignore the density of the candidate points and focus rather on efficient exploration of the *space* populated. Coverage designs select molecules near the center of clusters. Molecules near the edges of the design space are naturally avoided because they are unlikely to be near the center of a cluster.

5.4 Combinatorial Chemistry Example

This example illustrates the usefulness of a tool that assigns a rank ordering to molecules in a set. A combinatorial chemistry collection at Lilly consisted of a number of separate libraries. The question arose as to which of the libraries was the most diverse. To answer this question, a spread design was used to rank the combinatorial molecules. We pooled 22 combinatorial libraries (105,640 molecules) with a set of 32,262 corporate library molecules. We rank ordered the combinatorial molecules relative to the corporate library molecules; that is, the corporate library molecules were marked as pre-selected and the task was for the combinatorial candidates to augment them as well as possible. The spread design chose molecules from the pool irrespective of which library they came from—the only criterion was their diversity. We examined the cumulative number of molecules selected from each combinatorial library as a function of spread design rank, as follows. The first molecule chosen was the one most dissimilar to the corporate collection and received a rank of 1. The next molecule was that which was most dissimilar to both the corporate collection and the first molecule, and received a rank of 2, and so on. Libraries that were drawn from most frequently by the algorithm in

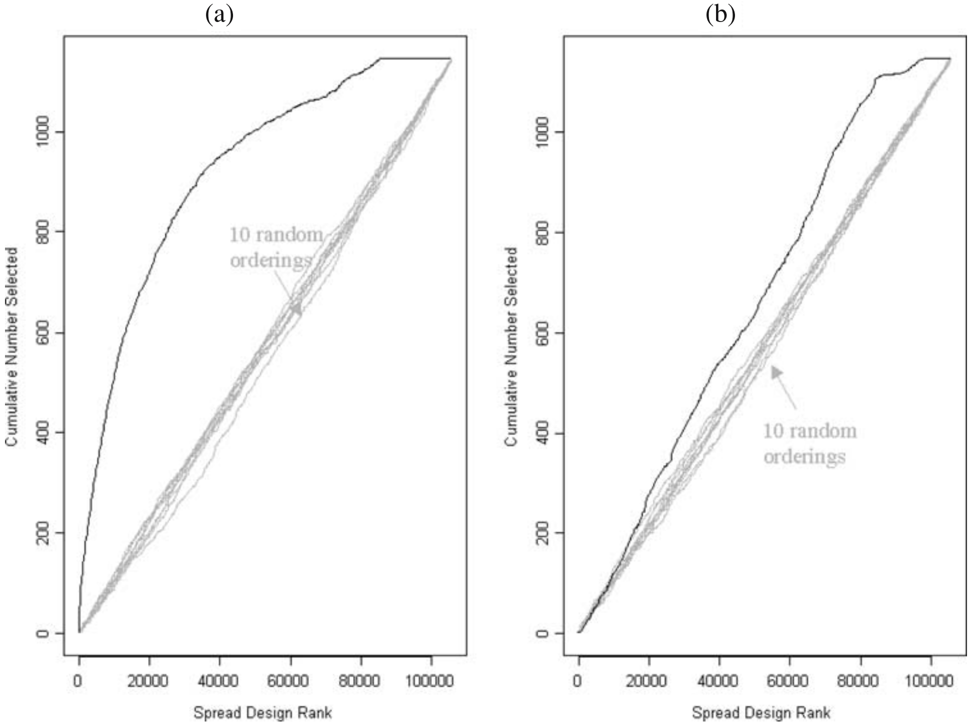

FIGURE 6. Combinatorial libraries comparison: cumulative number of molecules selected versus the rank of the spread design for each of 10 random orderings of molecules within (a) Library A; (b) Library B.

the early stages (early ranks) were taken to be the most diverse libraries. In the case of Figure 6, library A was far better than library B at augmenting the current collection.

This example shows how spread designs can be used to solve practical problems. There is always a descriptor selection problem, as chemists continue to invent new molecular descriptors. Which should be used? Which molecular similarity measure performs best? Controlled experiments are expensive. Simulation can be used as a guide.

All of this effort is invested in the first of a number of iterations in the drug discovery cycle and the later stages are much more rewarding. At Lilly, we move as quickly as possible into the predictive modeling stages.

6 Machine Learning for Predictive Modeling

Machine learning is defined as the use of algorithms to generate a model from data, which is one step in the knowledge discovery process, applied in the context of QSAR (last row of Table 2). The last decade of machine learning advances has seen tremendous increases in prediction accuracy, largely due to model averaging techniques. A good starting point for reading about such ensemble methods is the paper of Breiman (1996) and a valuable discussion about algorithmic versus parametric modeling approaches is provided by Breiman (2001b). Hastie et al. (2001) gave a broad overview of statistical learning (see, especially, the figures on pages 194 and 199). Predictive models can serve as useful tools and have made substantive contributions to many disciplines.

6.1 Overview of Data Handling and Model Building Steps

Figure 7 gives a brief layout of sequential steps for a typical data modeling exercise. The first step, which is by far the most time consuming, starts from a representation of the structures of the molecules and ends with a "training set" of descriptors to be used in the model selection step. Medchem filtering, in step 1, is an application of expert judgment to chemical structural data. Certain fragments of molecules are known to be highly reactive, or carcinogenic, or unstable, or otherwise undesirable, and these molecules can be eliminated at this first step with a simple rule-based algorithm. Data cleaning is, by far, where most of the time is spent.

The data cleaning steps may involve the removal from the data frame of columns (of descriptor values) that are constant or nearly constant, imputing missing values and eliminating columns that are redundant due to a strong relationship with other columns. All these steps are easily automated. Approximate algorithms can easily be developed that are more than 100-fold faster than those available in commercial packages.

The next step of data cleaning is to perform a replicate and pseudo replicate analysis of the experimental values. When replicate data are available, highly discrepant results can point to problems with the experimental data. When replicate results are not available, pseudo replicates are almost always present in the data. Often the same chemical structure exists more than once in the results file, where the different identifiers refer to different batches of the same material. Thus, a

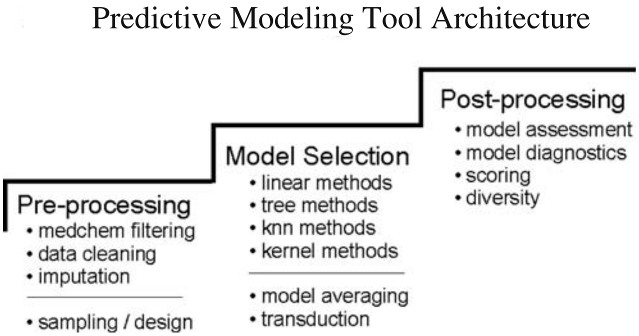

FIGURE 7. Steps involved in predictive modeling.

large discrepancy in the biological response for two such identifiers suggests that a follow-up meeting with the appropriate biologist is needed in order to resolve the experimental discrepancies.

Another aspect of data cleaning arises when data come from different laboratories. Then one is faced with the task of placing the results in a reliable and consistent context (a sort of "metaanalysis"). Another data cleaning task involves the imputation (estimation) of missing values. Often the programs that compute descriptors will fail on unusual molecules and then those molecules are usually removed from further consideration. However, sometimes a failure is not a reflection of the desirability of the molecule and imputation of the missing values is then a reasonable strategy.

The final portion (the sampling step) of the first step of Figure 7 is to create a *training set* which is the set of data to be used for fitting the model in the model selection stage. This may be done in several different ways. The simplest is merely to take random subsamples of the data and to split the data into training and test data. A more rigorous approach involves splitting the data into a training, validation, and test set (Hastie et al., 2001, pages 195–196), where the test set is used only once for assessing the error of the final model (after the model selection studies have finished) and the training and validation sets are used for model selection to compare competing modeling methods.

The "design" question of what proportion of the data to use for training, relative to the test set, is an important one. If the test set is too small, the estimates of error will be unreliable, highly variable, and likely to have a high downward bias. On the other hand, if the training set is too small the estimates of error will have high upward bias.

The second step shown in Figure 7 lists the actual model training steps. These typically involve a model selection exercise in which competing modeling methods are compared and a choice of one or more modeling methods is made. Listed in the figure are four of the many popular classes of modeling approaches. We use all of the methods listed; see Section 6.3.

6.2 Error Rates

Some of the examples and discussion in this chapter draw on the two-class classification problem, which here is "hit" versus "inactive". The word "active" refers to a validated hit, that is, a molecule that truly does exhibit some level of the desired biological response. A key point is that an assay is itself an estimator. With this in mind, definitions and a discussion of error rates are given in the context of predictive models. Borrowing from the terminology of signal detection, the "sensitivity" of a model refers to the fraction of observed hits that are classified as (or predicted to be) hits by the model, and "specificity" refers to the fraction of observed inactives classified as inactives by the model. An observed hit is not necessarily an active molecule, but simply a molecule for which the primary screening result exceeded a decision threshold. Whether such a molecule turns out to be an active is a problem that involves the sensitivity of the assay, but the task at hand is for a model to predict accurately the primary screening outcome and to assess the accuracy of the model for that purpose.

Let $\hat{I}$ denote "predicted by the model to be inactive" and I denote "observed to be inactive in the assay by exceeding the decision threshold", with analogous definitions for $\hat{A}$, "predicted to be a hit", and A, "observed to be a hit". With the null hypothesis that a compound is inactive, we have:

$$\begin{aligned} \text{specificity} &= P(\hat{I} \mid I) = P(\text{model prediction} - \mid \text{observed} -) \\ P(\text{Type I error}) &= P(\hat{A} \mid I) = P(\text{false positive}) = 1 - \text{specificity}. \end{aligned}$$

Similarly, 1 minus the sensitivity gives the probability of Type II error or the false negative rate:

$$\begin{aligned} \text{sensitivity} &= P(\hat{A} \mid A) = P(\text{model prediction} + \mid \text{observed} +) \\ P(\text{Type II error}) &= P(\hat{I} \mid A) = P(\text{false negative}) = 1 - \text{sensitivity}. \end{aligned}$$

The complementary rates are obtained from the opposite conditioning: the fraction of model-predicted hits that are observed hits $(A \mid \hat{A})$ and the fraction of model-predicted inactives $(I \mid \hat{I})$ that are observed inactives. We call these the "positive discovery rate" and "negative discovery rate". It is important to look at these conditional probabilities; a very clear example is in the analysis of gene chip microarray data where the false discovery rate is 1 minus the positive discovery rate as defined above and in Chapter 6; an excellent discussion is given by Benjamini and Hochberg (1995).

An example in the context of blood–brain barrier (BBB) predictions (see Section 6.5) is shown in Figures 8 and 9. Data from different laboratories at Lilly and from various literature sources were pooled together and molecules were assigned binary class labels, BBB+ and BBB−, depending on whether they crossed the blood–brain barrier. A random forest model, defined in Section 6.3, was trained on this data set and molecules that were not part of the training set (called "out-of-bag" in the bagging or random forest terminology) were predicted to be hits BBB+ or inactive BBB− according to a particular score/decision threshold. These predictions were evaluated and three rates were examined: sensitivity, specificity, and positive discovery rates—shown as a function of decision threshold in Figure 8, where the scores are multiplied by 10. If the goal is to obtain equal sensitivity and specificity rates (a common practice), then the optimal threshold is 0.778. Because both sensitivity and specificity are conditioned on the observed class labels, we feel it is important to include a rate that conditions on the predicted score or class label. Thus we include the positive discovery rate in our analysis.

Balancing these three rates equally yields an optimal threshold of 0.846. Both thresholds are indicated by vertical lines in Figure 8. Figure 9 shows the actual predicted scores for the molecules that do cross the blood–brain barrier (BBB+) as well as those that do not (BBB−). The false positive and false negative rates are, of course, direct consequences of which threshold is chosen. The appropriate threshold depends on the goal. For example, if the project is a neuroscience project where BBB+ is the goal, it may be that the team wants to find and reject

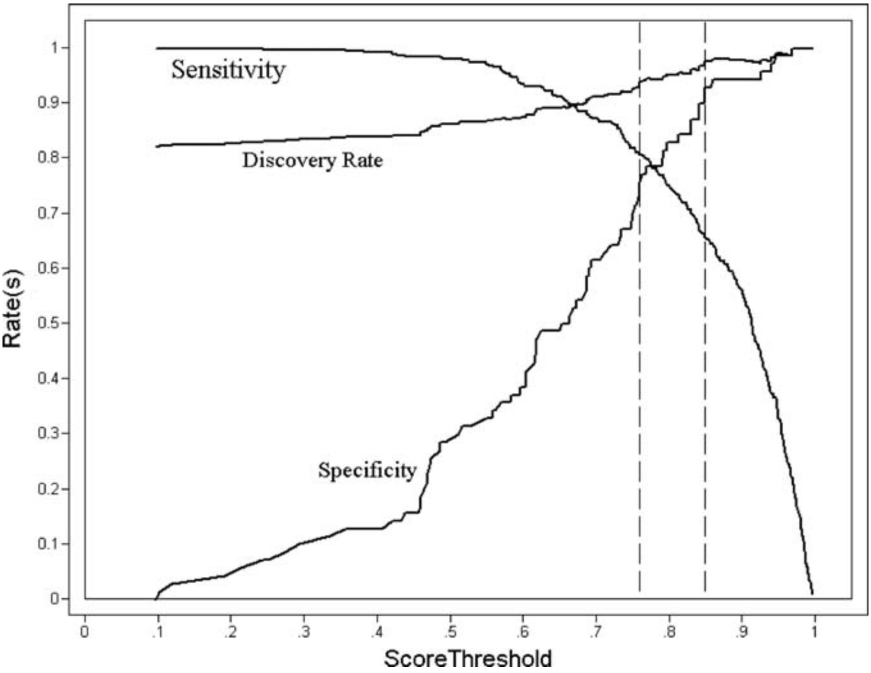

FIGURE 8. Sensitivity, specificity, and positive discovery rate as a function of decision threshold; the two reference lines correspond to two decision thresholds. The rates are estimated from predictions made for molecules not in the training set of the model.

compounds that are BBB− at an early point. Then, the goal would be to maximize sensitivity or to maximize the negative discovery rate (while realizing that going too far with this means losing a number of "good" compounds as well), and an appropriately large weight could be given, say, to specificity in computing the weighted average of the three rates to obtain an optimal threshold for that purpose.

6.3 Machine Learning Methods

Some of the more popular predictive modeling methods used in drug discovery include linear methods, tree-based methods, k-nearest neighbors, and kernel methods. In this section, a brief outline of these methods is given, together with references for reading and further details.

Linear methods include simple linear regression, multiple linear regression, partial least squares, logistic regression, and Fisher's linear discriminant analysis; see Hansch et al. (1962), Frank and Friedman(1993), and Hastie and Tibshirani (1996b). *Tree-based methods* are some of the most widely used methods today; see Breiman et al. (1984) and Rusinko et al. (1999). *Bagging* is a generic strategy

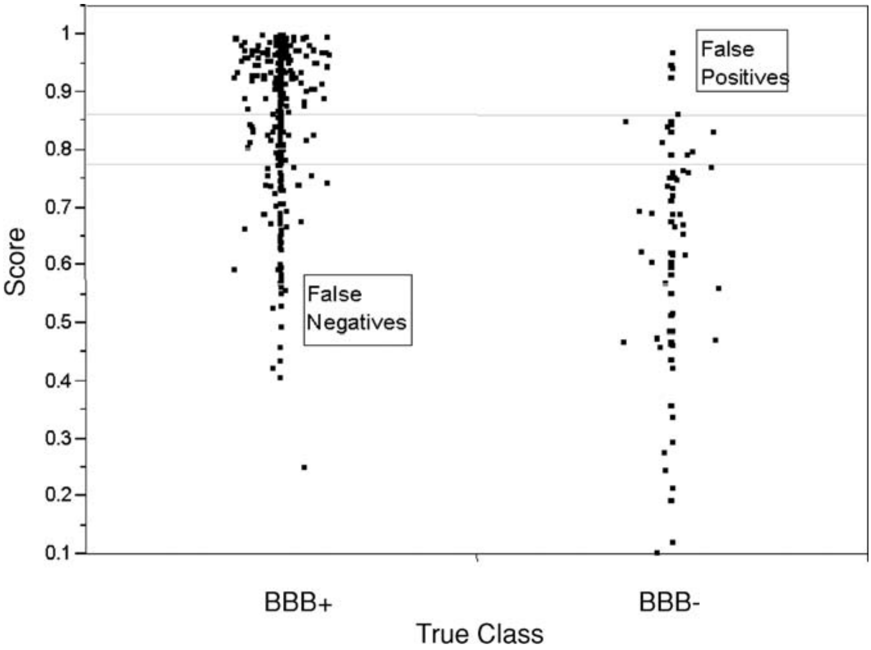

FIGURE 9. Random forest BBB predicted scores for molecules assigned as BBB+ and BBB−; horizontal reference lines correspond to two decision thresholds. All predictions (scores) are for molecules not in the training set.

that is useful in many contexts including tree-based methods. It was introduced by Breiman (1996) who motivated the strategy through the concept of unstable predictors. The bias and variance properties of aggregated predictors were further studied by Breiman (1998). *Random forests* is an improvement to the strategy of tree-based models combined with bagging. Details are given by Breiman (1999, 2001a). This is currently the top-rated algorithm in our project work at Lilly.

A simple, yet useful, and often highly accurate method is that of *K-nearest neighbors* described, for example, by Fix and Hodges (1951) and Dasarathy (1991). A notable recent advance in this method is given by Hastie and Tibshirani (1996a,b). In the case of a single descriptor, *kernel regression* and *smoothing splines* are useful methods of model fitting. However, far more general is the recent development known as *support vector machines*. This method is based on a particular hyperplane in the descriptor or property space that separates the active from the inactive compounds. This plane has the largest possible distance from any of the labeled compounds and is known as the maximum margin separating hyperplane. Support vector machines avoid overfitting by choosing the maximum margin separating the hyperplane from among the many that can separate the positive from negative examples in the feature space. Good starting points for reading about this topic are Burges (1998), Weston et al. (2002), and Vapnik (2000).

Transduction is another generic strategy that is an important recent advance. Standard practice in machine learning is to use *inductive learning*, that is, taking known molecules and, through training or fitting a model, generating a general understanding of the underlying relationships and then applying that general knowledge to make a prediction about a new molecule, for example, whether it will cross the blood–brain barrier. If the ultimate goal is to make predictions for a finite set of observations, then the rationale behind transduction is that the inductive learning step is not necessarily needed. Transduction skips the inductive learning step and goes directly to the prediction of the future examples. A nice heuristic explanation of this is given by Vapnik (1998, page 355). The general model that is best when applied to a universe of observations may not be the model that is best for the specific subset of observations under current scrutiny.

6.4 Model Selection and Assessment

Usually a variety of different models can be applied to the same data set, each model capturing part of the structural information relating explanatory variables to responses and also part of the noise. The objective of model selection may be considered, in a general sense, to be that of optimizing the quality of inference. In practice this can take several forms including discovering the "true" model, interpreting or understanding what natural process is driving a phenomenon, or simply choosing the model that gives the most accurate predictions on new data. In the QSAR drug discovery context, this latter objective is most often the appropriate one.

It is important to distinguish between algorithms and models. An algorithm creates a model, given data and tuning parameters as input. The model is a static entity. At Lilly we perform studies to select the best algorithm for a data set as well as the best model for a given algorithm, and finally to assess the error for a given model. A crucially important issue in model selection is the issue of model complexity, because training set error tends to decrease and test set error tends to increase with increasing model complexity; see, for example, Hastie et al. (2001), pages 194–199.

For the example of variable subset selection in multiple linear regression, the R^2 statistic increases monotonically as the number of variables added to the regression model increases, leading to the situation dubbed as *overfitting*. Various methods have been devised for avoiding an overfitted model. Some methods are simple adjustments to the familar R^2 statistic, such as the adjusted R^2 (R^2_{adj}) which adds a simple penalty for the number of covariates included in the model. Other popular methods include the Bayesian Information Criterion (*BIC*), the Akaike Information Criterion (*AIC*), and Mallow's C_p; see, for example, Burnham and Anderson (2002). In the context of high- or medium-throughput screening, when little is known about a target or an SAR, and designed experiments are not possible, there are no a priori models that can be assumed and, in any case, the key interest in early stage screening is in predictive accuracy of models rather than inference about model parameters.

TABLE 3. Size of the model space for multiple linear regression (MLR) with h descriptors and for binary tree models.

k	MLR	Tree model
1	2	2
2	4	9
3	8	244
4	16	238,145
5	32	283,565,205,126

When the model space is large, the problem becomes extreme. One solution is model averaging, in line with Breiman (1996). One good use for this approach is in recursive partitioning or tree-based models. The model space for recursive partitioning is huge. Consider the special case of binary descriptors and an algorithm that iteratively partitions data into two parts, depending on descriptor values. Once a descriptor is used to split the data, it can never be used again. Thus the model space is much smaller than when the descriptors have more than two values. For binary descriptors, the number of possible tree models $T(h)$ for a data set with h descriptors can be computed from a simple recursive formula:

$$T(h) = 1 + h \cdot [T(h-1)]^2, \tag{2}$$

where $h = 0$ corresponds to the case of no descriptors where the tree model is the null model composed of the overall mean. For multiple linear regression and a simple additive model, there are 2^h possible models for h descriptors. There is a rough analogy between the choice of parameters in the regression model and the choice of cutpoint along each descriptor in recursive partitioning. Table 3 shows the size of the model space for multiple linear regression and for binary descriptor, two-way split tree models, for up to five descriptors.

With great flexibility in model choice comes great power but also great danger of misuse. As the model space spanned by tree models is huge for $h \geq 4$, there is need for both a computationally feasible way to search the space and for some way to guard against finding spurious relationships in the data. The bagging method of Breiman (1996) was a key advance in this area.

For regression models, one metric that we use for sorting the molecules by their predicted activity, which is considered proprietary at Lilly, is similar to a weighted variant of Spearman's ρ. This metric, labeled S, ranges from -1 to $+1$ and compares the predicted and actual responses. The weights are higher earlier in the sorted list to emphasize that, in practice, it is the top of the sorted list that will identify the molecules selected for testing, and that accuracy farther down the list is not nearly as important. We have very little interest in accurately distinguishing the relative activity levels of molecules that are all considered inactive, but a great deal of interest in the degree to which actives will rise to the top of a sorted list of molecules. Quality assessments have been assigned to various values of S, but these levels in isolation are not meaningful; a very high value of S (or R^2, or any other metric) can easily be obtained for observations that are in the training set of a model, but does not predict how the model will perform on untested molecules. The thresholds established are based on appropriate test hold-out results, as described below. With this in mind, a value of zero is equivalent to random (the mean value resulting from scrambling the predicted responses and computing S many times). A value of 0.40 is considered a minimum standard for a model to be used for decision making at Lilly. Such a model would be considered weak and would not be used at all by some scientists. A model with an S value of 0.60 is considered a solid and useful model and an S value above 0.80 indicates a very good model. A difference in values of less than 0.05 is not considered to be meaningful. Thus, if one were doing model selection and two competing models were statistically significantly different but the difference in mean S were below 0.05, the two models would be treated as equivalent. At Lilly, we couple the concept of significant differences with the concept of meaningful differences.

6.5 *Example: Blood–Brain Barrier Penetration*

We examine a data set of 750 molecules with blood–brain barrier penetration measurements. An important aspect of drug design is the consideration of the potential for penetration of the blood–brain barrier by any new candidate drug molecule. Whether the goal is for the potential drug to cross or not to cross the blood–brain barrier, the ability to estimate the blood–brain ratio is an essential part of the drug design process. Determination of this aspect of a molecule is a low-throughput operation and thus having the ability to prioritize molecules in silico through the use of predictive models adds considerable value to the drug discovery process.

The penetration of a compound across the blood–brain barrier is measured experimentally as the ratio BB of the concentration of the compound in the brain to that in the blood. This ratio is thought to be related to local hydrophobicity, molecular size, lipophilicity, and molecular flexibility (Crivori et al., 2000), but no explicit mathematical relationship has been given. The 750 available results, from an in situ experiment with rats, are responses known as *Kin* values. These are intended to be related to the BB ratio of these compounds in humans. The current goal of the analysis is to select a subset of descriptors (covariates) from about 1000 possibilities and a modeling method that gives good predictive accuracy of the Kin. At Lilly, the modeling methods used do the subset selection intrinsically. In this example we compare just two methods: one is based on partial least squares (PLS) with a model-averaging strategy, and the other is the random forest algorithm of Breiman (1999).

We split the data set into two equal parts at random. We train our algorithms on one half and score the other half as test data. The idea is to study how the methods behave on new untested molecules. Whether a 50/50 split is the best choice is discussed shortly; here we consider the question of how many repetitions are needed. We started with 200 repetitions, each a random 50/50 split of the 750 data points into equal-sized training and test sets, where after each split the model was retrained and the test hold-out set was scored. The conclusion favored the random forest model over partial least squares, and a natural question arises as to whether smaller tests would lead to the same conclusion.

6.6 Training and Test Set Sizes

Two key questions in model selection are what proportion of molecules to use for a training set versus a test set when doing random splits of the data, and how many different training/test splits should be analyzed to obtain reliable inferences about model performance. The number of repetitions needed is surprisingly low and often the same decisions are made whether the number of training/test splits used was 200 or 20 or 10.

Miller (2002, pages 148–150) recommended fivefold to tenfold validation, so that effectively 80% to 90% of the data should be in the training set. Another recommendation is that $n^{3/4}$ of the data should make up a training set (randomly selected) and the rest predicted as test hold-out data; see Shao (1993) for details. However, it is easy to show that use of the $n^{3/4}$ rule does not perform well in settings such as drug discovery where prediction accuracy, rather than selection of the true model, is the objective. We are sometimes better off with a model that is not the true model but a simpler model for which we can make good estimates of the parameters (leading to more accurate predicted values).

In order to choose the model that predicts most accurately for the test data, we need a new rule or a new information criterion. The usual criteria, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), Leave One Out (LOO or Qsquared), and so on, are all insufficient for our needs. This motivated Kerry Bemis to propose a new measure which he called predictive R^2 or pR^2 (described below).

Although this is an area of ongoing research, the current opinion at Lilly is that, when comparing candidate modeling methods in a model selection exercise, it is best to look at the entire learning curve (leave 90% out up to leave 10% out) and make a judgment about learning algorithms based on the performance across the whole curve. This we call a *learning curve* (but note that the phrase is used in other contexts with other meanings). Figure 10 shows the performance of the two candidate modeling methods applied to the BBB data set of Section 6.5. We generated 20 sampling runs for each level of Ptrain, where Ptrain is the proportion of the data assigned randomly to the training set, and used the two methods over a broad profile of training set sizes. The two lines connect the means of the values of S obtained for the two methods at each Ptrain level. The same train/test splits were used for both the PLS and the random forest methods. Thus a paired or block analysis was done. Here, we could ask whether the random forest method is superior over all Ptrain levels and use a test such as Tukey's HSD (Honestly Significant Difference; see Tukey, 1997, Kramer, 1956). This is perhaps conservative in that we are not interested in all of the pairwise comparisons but, as we can see from simply looking at the plot, any formal comparison is going to give an unambiguous result for this data set.

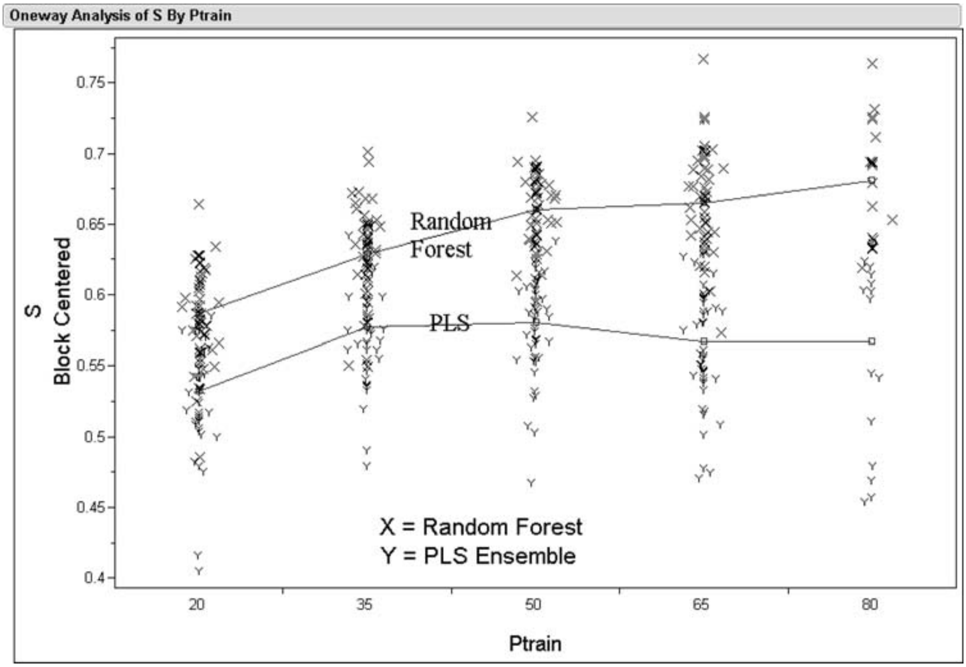

FIGURE 10. Model selection and assessment diagnostic: performance measure *S* for random forest and partial least squares (PLS) methods applied to the BBB data for various percentages of the data (Ptrain) in the training set.

There is a minimum size of training set necessary for a statistical model to be able to reveal links between vectors of descriptor values and biological activity. This has been called “statistical traction” by Young et al. (2002). Suppose a particular pharmacophoric feature is important for the binding of molecules to a receptor. Having one molecule that binds and has that feature is not sufficient for that feature to be detected as significant. Several examples of molecules that bind and contain that feature are needed before the statistical algorithm can detect it. In the model selection stage, it is possible to place a downward bias on the estimate of the predictive power of an algorithm by selecting for the training set a subset of the data that is too small. There may be a lack of “statistical traction” in the training subset that would not exist when the model is trained on all the available data. On the other hand, when the proportion of data selected for the training set is very large, and the test set is correspondingly small, it is more likely that a given test set molecule has a very similar “neighbor” in the training set and this gives an upward bias to the estimate of predictive power of the model.

Once the choice of modeling method has been made, all available data are used to train a final *scoring model* (to be used in the third step of Figure 7). Sometimes sampling issues arise here; for example, the data sets can be very large, and for classification data there is usually a huge imbalance in the number of examples of one class compared with another. In drug hunting, there may be 400,000 examples of inactive compounds and as few as 400 active compounds. If the available modeling methods do not deal well with this situation, there may be motivation either to create a training set that alleviates such a mismatch, or to create a smaller training set to reduce the computational burden. Either of these issues may or may not be related to the problem of model selection. One strategy for selecting a subset of available data for training a model is as follows.

1. Select all the active compounds.
2. Select a small subset of the inactive compounds whose nearest neighbor among the active compounds is a relatively short distance (by some distance measure such as those of Section 4.5). The motivation here is to preserve the boundary between classes.
3. From the remaining inactive compounds, select a maximally diverse subset (as described in Section 5). This augments the space beyond the boundary with an optimal exploration of the chemical space represented by inactives.

At Lilly we have focused on predictive accuracy in most of our project work. Predictive accuracy and interpretability tend to be inversely proportional. An active area of research at Lilly is an investigation of the question of ways in which the model can help us design a better molecule. This may involve interpretation, and there are excellent tools that can be used for this, such as partial dependence plots. It can also be approached through virtual screening—a scientist proposes a scaffold or series and the model provides an evaluation of the prospects of that idea.

6.7 The Predictive R^2 of Bemis

In the area of linear models, Bemis has proposed a "predictive R^2" or pR^2. Until 2004 this criterion was treated as a trade secret at Lilly. The pR^2 does not involve training/test split cross-validation, but rather uses an information-theoretic criterion motivated by ideas of Shi and Tsai (2002). For a model with h parameters,

$$pR_h^2 = 1 - \exp\left[\frac{RIC_h}{k-h-1} - \frac{RIC_0}{k-1}\right],$$

where RIC_0 corresponds to the Shi and Tsai RIC (residual information criterion) for the null model. To give more clarity, we give an alternative notation of the pR^2, building up from the familiar R^2 to the adjusted R^2 and finally to the pR^2. For a linear model with h parameters:

$$R_h^2 = 1 - \frac{SSE_h}{SSE_0} = 1 - \frac{SSE_h/k}{SSE_0/k} = 1 - \frac{\widehat{\sigma}_h^2}{\widehat{\sigma}_0^2},$$

$$adjR_h^2 = 1 - \frac{SSE_h/(k-h-1)}{SSE_0/(k-1)} = 1 - \frac{\tilde{\sigma}_h^2}{\tilde{\sigma}_0^2} = 1 - \exp\left[\log\left(\tilde{\sigma}_h^2\right) - \log\left(\tilde{\sigma}_0^2\right)\right],$$

$$pR^2 = 1 - \exp\left[\left\{\log\left(\tilde{\sigma}_h^2\right) + bias_h\right\} - \left\{\log\left(\tilde{\sigma}_0^2\right) + bias_0\right\}\right],$$

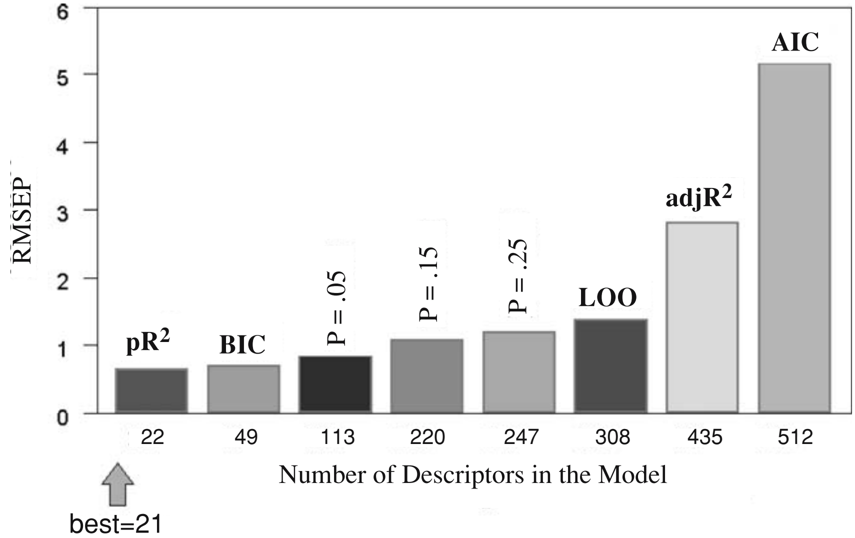

FIGURE 11. Performance of competing criteria: the number of descriptors in the model, for various criteria versus the root mean squared prediction error (RMSEP) in forward selection. (Reproduced with permission from the author.)

where

$$bias_h = \frac{h+1}{k-h-1}[\log(k) - 1] + \frac{4}{(k-h-1)(k-h-3)}.$$

Figures 11 and 12 illustrate the performance of the pR^2 compared with several of the currently popular criteria on a specific data set resulting from one of the drug hunting projects at Eli Lilly. This data set has IC50 values for 1289 molecules. There were 2317 descriptors (or covariates) and a multiple linear regression model was used with forward variable selection; the linear model was trained on half the data (selected at random) and evaluated on the other (hold-out) half. The root mean squared error of prediction (RMSE) for the test hold-out set is minimized when the model has 21 parameters. Figure 11 shows the model size chosen by several criteria applied to the training set in a forward selection; for example, the pR^2 chose 22 descriptors, the Bayesian Information Criterion chose 49, Leave One Out cross-validation chose 308, the adjusted R^2 chose 435, and the Akaike Information Criterion chose 512 descriptors in the model. Although the pR^2 criterion selected considerably fewer descriptors than the other methods, it had the best prediction performance. Also, only pR^2 and BIC had better prediction on the test data set than the null model.

6.8 *Common Errors*

Predictive modeling, statistical modeling, and machine learning are very open areas in the sense that the barrier to admission is very low. All that is needed to

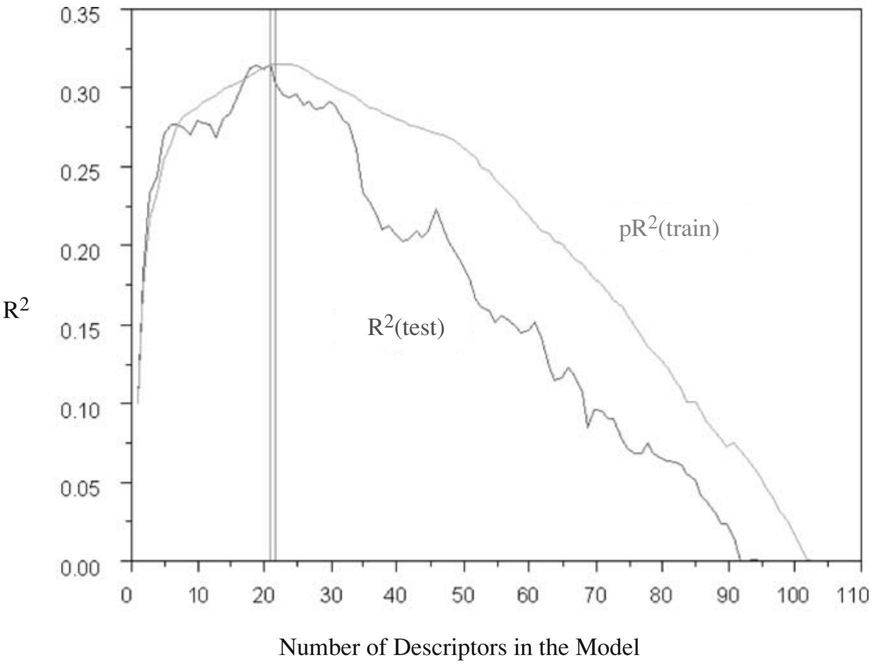

FIGURE 12. The Bemis pR^2 for an example data set. The "true" observed R^2 based on the test set, and the pR^2 estimated only from the training set. (Reproduced with permission from the author.)

start experimenting in this area is a PC and a data analysis package. Below is a list of the most frequent errors as they occur in this field.

1. Belief that a very small *p*-value for a predictor (for example, a biomarker) is more likely to occur with high predictive accuracy. The multiple testing problem must not be ignored, and the false discovery rate (FDR) controlled; see also Chapter 6.
2. Failure to pre-process and clean the data. Sometimes even data with missing values are jammed through a learning algorithm with little thought.
3. Use of an unsupervised algorithm to do the job of a supervised algorithm. For example, a cluster analysis or self-organizing map is used in combination with a post hoc analysis to do prediction.
4. Failure to evaluate a method on test data.
5. Test data set too small, with these consequences:
   a. Prediction error cannot be accurately estimated on each hold-out part.
   b. The test sample and the training sample are likely to be similar in their descriptor values.
6. "Cheating" in model assessment: using the whole training set to select descriptors, and then splitting the data into train and test sets, and training the model on the descriptor set selected from the whole training set.

7. Comparison of methods with the same type of feature selection forced on all methods, rather than letting each method do what it does best.
8. Confusion of model selection with model assessment. If one chooses the model with the lowest cross-validated error among competing models, that error is not a valid estimate of the prediction error of that model (selection bias).

7 Iterative Medium-Throughput Screening

Researchers who use high-throughput screening (HTS) methods are troubled with many obstacles such as poor data quality, misleading false-positive and false-negative information, and the need to confirm and expand the SAR of the identified lead candidates. Additionally, HTS strategies lead to the large-scale consumption of valuable resources such as proteins and chemicals from the inventory, and may not be applicable to all targets (Major, 1999). The problem is fueled, in particular, by the prospects of expanding universes of targets—an increase by a factor of 10 is expected (Drews. 2000)—that will lead to an explosion of costs. As a consequence, there is a need not only to increase the scope of screening, but also the efficiency of each screening experiment. Hybrid screening strategies have been suggested that unite in silico and in vitro screening in one integrated process.

Iterative medium-throughput screening (MTS) starts with a small (200 to 20,000) and "diverse" subset of compounds. This initial sample is subjected to a primary screening where the main objective is to gather SAR data for predictive model building. This is a key distinction from the older paradigm where the primary objective is to obtain an initial set of screening hits, and any subsequent model building is an added bonus. Based on this first SAR, the corporate inventory is screened in silico in order to identify a further, more focused set of compounds, the focused library, for a second round of MTS. Several cycles of testing–analyzing–testing can be applied aimed at either refining the SAR model(s) or the identification of more active compounds. Abt et al. (2001) studied the influence of the size of the focused sample and the number of cycles on the effectiveness of the computational approaches.

One factor that plays a role in the decision on how compounds are chosen in a given cycle is the stage of a project. An early stage project may require more diversity to be built into the selection process, whereas a later stage lead optimization effort would draw much more heavily on predictive modeling and expert judgement.

Active learning is a strategy that is iterative and where the selection of compounds to test in the next iteration is based on all the currently available data (see, for example, Warmuth et al., 2003 and Campbell et al., 2000). A key distinction between active learning and other modeling strategies is that, in active learning, the primary objective for selecting the next batch of compounds for testing is to improve the model optimally, whereas in drug screening programs the primary goal is to find as many potent (and usually novel) compounds as possible. This distinction and the consequent effects are dramatic. In active learning, the most interesting compounds are the ones for which the model has had difficulty in the assignment of a clear classification whereas, in a typical drug hunting program, the most interesting compounds are the ones that are scored the most unambiguously as active. This has downstream implications on what will come out of future iterations of screening. The traditional business-driven approach will find good compounds faster, but the active learning approach will generate better models faster, and eventually lead to better exploration of chemical space, resulting in finding the best compounds.

There is also an analogy with ancillary efforts such as toxicity testing. A drug hunting project tends to focus on finding compounds with potency and selectivity for the target of interest. When interesting compounds are found, they are submitted for toxicity testing so that a small set of structurally related compounds is tested for the toxic endpoint of interest. This places a handicap on any toxicity modeling effort. If the goal is to develop a good toxicity model (which would reduce the need for animal testing and reduce cycle time in the project), then compounds that are not interesting from a potency standpoint would need to be tested for toxicity. This would mean that the interesting potent compounds must wait their turn due to limited capacity in toxicity testing. The long view, both in terms of active learning for potency and for toxicity, might be to strike some balance between immediate and future gains.

8 Virtual Screening and Synthesis

Virtual screening is a simple concept, arising from the need to break out from the confines of the currently available set of in-house chemical libraries. It is a simple matter to construct representations of molecules using computers, and this can be done in a combinatorial manner. Usually one or more "scaffolds" are chosen—these are the "backbone" of the molecule. Then a number of "substituents" are chosen; these can be thought of as the "appendages" that gets attached to the backbone at various (preselected) locations. All (reasonable) combinations of scaffolds and substituents can be made in silico and these structures form a virtual library. The library may contain millions of molecules but it is more typical to see something of the order of 500,000 structures. This is because most virtual screening efforts are knowledge driven; something is known about the SAR before the virtual screen is attempted. Most of the molecules in the virtual library will not exist in the corporate molecular stores. This virtual library is then the subject of a modeling effort whereby the virtual library is prioritized and rank ordered, with the most promising structures at the top of the list. The biological screening is done virtually through the use of the predictive models applied to the virtual library.

Some of the high-ranking structures may be very similar to structures that have already been tested. These are removed from the list using molecular diversity methods such as a Leader algorithm (Hartigan, 1975). In this context, the Leader algorithm is not providing a cluster analysis, but simply a post-processing of a rank ordered list. From what is left, a relatively small number of molecules are then synthesized and tested for biological activity. Because this is a relatively expensive part of the process, it is usually important that some knowledge about the SAR has been gained before the virtual screening is done. That knowledge could come from literature sources or from prior early-stage screening.

9 Pooling

Pooling strategies can take numerous forms, as discussed in Chapter 3. In the drug hunting screening context, chemical compounds can be pooled. Ten compounds may be pooled together in a well and tested as a mixture. If the mixture is potent, the individual components can then be tested. If the mixture shows no potency, it might be assumed that the individual components are each inactive. This assumption may sometimes be incorrect, as compounds may exert an antagonistic (or conversely, synergistic) effect on each other. For the use of orthogonal arrays in the design of a pooling study see Phatarfod and Sudbury (1994; and also Dorfman, 1943).

The design and deconvolution of pools in drug discovery screening has been approached in different ways by a number of companies. In a highly specialized experiment at Merck, Rohrer et al. (1998) pooled a staggering 2660 compounds per well. The deconvolution of these results was done using chemical technology rather than the informatics approach one might use following Phatarfod and Sudbury (1994).

An interesting informatics strategy involves pooling covariates in a variable subset selection context. Suppose one has a data set with hundreds of thousands of covariates (descriptors), as happens in the drug discovery setting, and perhaps one does not have a data analysis package capable of handling so many columns of data. If the covariates are sparse binary, meaning that each column is mostly zeros with a few ones (a typical scenario), one strategy for data reduction is to pool columns together. One could take batches of, say, 100 columns and simply add them, creating a "pooled covariate." This data set is now 100-fold smaller, and a forward selection method might be run to fit a model on the reduced data set. Variables selected by such a procedure can then be collected and the individual covariates unpacked and a second stage of variable selection performed on this reduced data set.

10 Expectations for Discovery of Rare Events

The hit rate within a set of molecules selected by a virtual screen is primarily determined by two parameters: the unknown proportion of p hits that exist in the set of molecules scored and the false positive error rate (α) of the classifier used for virtual screening. To a large extent, the statistics of rare events (true hits within a large compound collection) leads to some initially counterintuitive results in the magnitude of a hit rate within a set of molecules selected by a model.

Most pharmaceutical companies expect to see hit rates in the 0.1% to 1% range for a high-throughput screen. In the virtual screening context, when the hits are a rare event (of the order of 0.1%) even very good predictive models cannot be expected to lead to arbitrarily high hit rates for the molecules selected. It is quite likely that marginal to good virtual screen models will result in no hits identified in a subset of molecules selected by virtual screening.

The virtual screen can be considered as a classifier that makes a prediction about whether a molecule is likely to be active or inactive in a biochemical assay. It can be constructed from training data (for example, a QSAR model) or constructed from a model of a binding site. For a given molecule in a virtual library, let the null hypothesis be that the molecule is not a hit. Then, using the notation of Section 6.2,

$$P(\mathrm{A}) = p, \qquad P(\hat{\mathrm{A}} \mid \mathrm{A}) = 1 - \beta, \qquad P(\hat{\mathrm{A}} \mid \mathrm{I}) = \alpha,$$

$$P(\mathrm{A} \mid \hat{\mathrm{A}}) = \frac{P(\hat{\mathrm{A}} \mid \mathrm{A})P(\mathrm{A})}{P(\hat{\mathrm{A}} \mid \mathrm{A})P(\mathrm{A}) + P(\hat{\mathrm{A}} \mid \mathrm{I})P(\mathrm{I})} \tag{3}$$

$$= \frac{(1-\beta)p}{(1-\beta)p + \alpha(1-p)}. \tag{4}$$

Equation (3), which is an application of Bayes theorem, is referred to as the "Positive Predictive Value." The parameter p is unknown but believed to be very small (<0.01) for large virtual libraries. $1 - \beta$ is the power (or 1 – type II error, where β is the false negative error rate) and α is the type I error, also called the "size" of a test in the hypothesis testing context, or the false positive error rate. The last equation defines the probability that a molecule is determined to be a hit in a biochemical assay given that the virtual screen predicts the molecule to be a hit. This probability is of great interest because it is valuable to have an estimate of the hit rate one can expect for a subset of molecules that are selected by a virtual screen.

The values of parameters p, α, and β can be varied to observe the effect on equation (3). It is straightforward to verify that the "power" of the classifier $(1-\beta)$ has relatively little effect on the hit rate observed in the subset of molecules selected by a virtual screen. The influence of power is greatly reduced as the probability of a hit existing in the set of compounds being scored decreases (the low prevalence effect) and, for rare events, the relative importance of α is greatly intensified. Even for less rare events, say a hit rate of 10% (disturbingly high in drug discovery, suggesting nonspecificity in the assay), the effect of α dominates.

11 Drugability of Molecules: ADME, Solubility, Toxicity

The word *drugability* is often used to cover all aspects of a molecule beyond initial potency. A potential drug compound must overcome many challenges in order to be a successful therapeutic. Critical components of drug design include absorption, permeation, distribution, metabolism, stability, specificity (does it do more than intended?), and toxicity (related but not identical to specificity). In this section, some of these issues are discussed in more detail.

11.1 ADME

Many of the compounds entering clinical trials are discontinued, often due to issues directly related to *ADME*: absorption, distribution, metabolism, and elimination/excretion of a drug. *Absorption* is of paramount importance, being the extent to which an intact drug is absorbed from the gut lumen into the portal circulation. *Distribution* is important because the drug will not work if it is not transported to the intended site. A compound may have potent effects in vitro screens involving cells or enzymes, but in a living organism the compound may have no effect because of a distribution problem. This can be due to a number of things; for example, the compound may bind so tightly to proteins in the bloodstream that it does not leave the bloodstream until it is eliminated by the liver. The opposite extreme can be a problem as well, because proteins in the blood can be important as transport mechanisms. In addition, the unbound drug may penetrate the wall of the blood vessel so that a certain amount of protein binding is desirable. Most pharmaceutical companies have models that predict the protein binding affinity of compounds. Distribution is only one problem that can confound an SAR effort when transitioning from in vitro to in vivo screens.

Two endpoints important to distribution are *oral bioavailability* and *first pass clearance*; see Birkett (1990, 1991). Oral bioavailability is particularly important because a drug that has, say, only 10% oral bioavailability would require a 10-fold higher dose when given orally as compared with being given intravenously. Orally administered drugs, after absorption through the gut lumen into the portal circulation, must then pass through the liver before reaching the systemic circulation. Pre-systemic or first pass extraction refers to the removal of drugs during this first pass through the liver. *First pass clearance* is the extent to which a drug is removed by the liver during its first passage from the portal blood on its way to the systemic circulation. Oral *bioavailability* is the fraction of the dose that reaches the systemic circulation as intact drug. It is apparent that this will depend both on how well the drug is absorbed and how much escapes being removed by the liver. In fact, the simple equation for bioavailability is

$$Ba = \text{fraction absorbed} \times (1 - \text{extraction ratio}),$$

where the extraction ratio is the proportion removed by the liver. Thus if drug $\mathcal{A}$ has 80% absorption and 75% extraction ratio, then the bioavailability of $\mathcal{A}$ is 20%. The 20% alone does not tell us anything about the metabolism or the absorption of the drug.

Because there are many ways to achieve a given level of bioavailability, it makes sense to consider using a compartmental model to predict bioavailability rather than simply training a model on a set of bioavailability results. The role of metabolism tends to dominate most often and variability in drug response is greatly influenced by this. Drugs that are efficiently eliminated by the liver often have high variability in the plasma levels both within and between individuals because, in that case, slight changes to the extraction ratio can cause large changes to the resulting bioavailability.

Treated as a special case of distribution is the ability of a molecule to cross the blood–brain barrier (BBB). This fact is important to know, both for central nervous system (CNS) drugs and for drugs that do not target the central nervous system. There has been a flurry of research attempts to model and/or predict the BBB propensity of molecules. Many of these efforts are statistically destitute; for example, a research group may examine only a set of molecules that do cross the BBB. Proper inference must involve examples of compounds that do not cross the BBB as well as compounds that do and this falls in the domain of predictive modeling and machine learning (see Section 6). The BBB is formed by the highly selective capillaries of the central nervous system. Passage of drugs through the BBB may occur by passive diffusion or from various specific uptake mechanisms, many of which are there to supply nutrients to the brain. There are also mechanisms for transporting substances out of the brain. P-glycoprotein (or Pgp) is an efflux pump that removes many drug compounds from the brain. Thus BBB transport is a complex phenomenon and modeling this is a challenging and ongoing research topic in most pharmaceutical companies.

Metabolism is another critically important aspect for determining the fate of a drug. If a drug is metabolized quickly, it may be excreted in the urine before it has a chance to reach the intended site, but the full story is much more complicated than this. Most successful drugs are lipid-soluble and are reabsorbed from the kidney back into the bloodstream. These compounds undergo metabolism, which is a way for the body to break down and ultimately eliminate a substance. The liver uses a number of different enzymes to break a compound down into smaller parts, called metabolites. A metabolite may either be pharmacologically similar to the parent compound or harmless, but not pharmacologically active, or may possess life-threatening toxicity. Thus it is essential to know into which of these categories a drug falls and it is desirable to control this aspect in a favorable way. Ideally a compound would metabolize at a moderate rate, neither too slowly nor too quickly. Because humans are genetically diverse, the same compound will be metabolized differently in different people. All of these issues are interdependent and are illustrated in the following examples.

A major group of enzymes, not just in the liver but also in the intestines, lung, kidneys, and brain, is known as the Cytochrome P-450 isoenzymes, often abbreviated as *CYP450*. Some drugs interact with these enzymes. A drug with a high affinity for an enzyme will slow the metabolism of any low-affinity drug; for example, grapefruit juice inhibits a number of CYP450s which results in higher than expected levels in the body of the drugs that are metabolized by those CYP450s.

The inhibition of CYP450 isoenzymes by grapefruit juice lasts about 24 hours and occurs in all forms of the juice—fresh fruit and fresh and frozen juice. There is the potential for dangerous arrhythmias for patients taking cisapride, astemizole, and terfenadine. Other substances may induce the opposite effect, that is, upregulate the levels of the enzymes and result in faster metabolism, for example, smoking and the ingestion of charbroiled meats may induce isoenzymes, resulting in increased clearance of drugs (such as theo-phylline). The herb known as St. John's wort causes an increase in the Cytochrome P450 enzymes, especially CYP 3A4, which are responsible for the metabolism and elimination of many drugs. This is why the birth control pill is rendered less effective by St. John's Wort. But in addition to being an inducer of CYP 3A4, St. John's Wort is also an inhibitor of CYP 2D6. Hence patients taking St. John's wort are likely to experience an increase in blood levels of therapeutic drugs that are metabolized by the 2D6 family (this includes beta blockers, antidepressants, antipsychotics, cough suppressants, codeine, and others) as well as a decrease in blood levels of drugs that are metabolized by the 3A4 family of CYP 450s (which includes antibiotics, HIV protease inhibitors, antihistamines, calcium channel blockers, and others).

Just two decades ago, the FDA was uninvolved in issues regarding CYP450 metabolism, but currently there are stringent guidelines that must be met to ensure that the metabolic fate of a drug is under control. There are genetic polymorphisms in some of the genes expressing CYP450 subfamilies. For example, 5 to 10 percent of Caucasians have polymorphic forms of the 2D6 subfamily; such individuals are called "slow metabolizers." There is a large list of drugs metabolized by 2D6 that can pose a risk to slow metabolizers and dosing must be done carefully.

11.2 Solubility

Solubility plays a critical role in the absorption of a drug. A compound with poor solubility may not achieve high enough levels in the stomach and intestine to be absorbed well. However, it is generally true that highly soluble compounds lack sufficient lipophilicity to cross the blood–brain barrier and so, if the compound is an intended CNS drug, a balance must be maintained; see Amidon et al. (1995).

11.3 Toxicity

Toxicity is often related to ADME; for example, when a compound cannot be broken down and eliminated by the body it builds up toxic levels in the system. Some other toxicity issues that have recently received heightened attention are discussed below. Phospholipidosis (an adaptive storage response to drug administration) and cardiomyopathy (a pathologic condition of the heart muscle) have been reasons for the recent FDA withdrawal of drugs. Another issue concerns the need for additional assurance of the absence of any potential for QT prolongation (an effect on electrical impulse conduction in the heart). Many classes of drugs induce QT prolongation, including antihistamines, antibiotics, antipsychotics, and macrolides. QT prolongation can lead to sudden death. At least four drugs have been taken off the market due to QT prolongation alone: Terfenadine, Sertindole, Astemizole, and Grepafloxacin. Acquired long QT syndrome (LQTS) occurs as a side effect of blockade of cardiac HERG K+ channels by commonly used medications. These issues have inspired modeling efforts aimed at predicting these effects and using those predictions to filter compounds in the early screening stages. Because these models are less than perfect, false negative and false positive rates are an issue.

A common approach to lead optimization is now parallel optimization, which is discussed in the following section. This is an extremely challenging undertaking because it requires the simultaneous control of several medicinal chemistry components. Furthermore, many of these components are not independent and, in fact, may even be negatively correlated. To do parallel optimization, it is also necessary to generate drugability related information in the early stages of lead optimization.

12 Multi-Objective Optimization Methods

Decisions in drug discovery are almost always multidimensional. Numerous criteria must be managed in order to develop a successful drug: potency, selectivity, toxicity, and ADME characteristics, and these tend to have conflicting trends so that difficult decisions are forced on scientists. For example, Zyprexa is an excellent antipsychotic drug but it causes weight gain in most people, a side effect of almost all antipsychotic drugs. Why this happens is still being investigated and there are at least six different hypotheses given in the literature. It is possible to modify an antipsychotic drug so that it does not produce weight gain, but such modifications may reduce the potency of the drug or introduce other side effects which may be even worse. A common side effect, for example, of many antipsychotic drugs is "extrapyramidal side effects" (EPS) which produce symptoms such as tremors, rigidity, and slowness of movement. These are deemed by most to be worse than weight gain. Less clear-cut trade-offs might involve the propensity for a molecule to cross the blood–brain barrier versus the therapeutic effect desired. For example, Benadryl is still a popular drug because, in spite of its tendency to induce a feeling of somnolescence, it is an extremely potent histamine (H1) blocker. Specificity is a problem faced by virtually every project team in drug discovery. Potency is desired at one receptor but not at another.

The old paradigm in drug discovery, which might be labeled "sequential search," generally fails. With this paradigm, one would optimize each objective independently and in succession. Finding a lead compound corresponds to searching on one landscape. Optimizing the lead corresponds to searching additional landscapes starting with the results of searching the first. With more than two objectives, the likelihood of failure increases exponentially. What is needed is a holistic approach with a mathematical framework for considering trade-offs between objectives. A variety of algorithms exists for finding the best possible trade-offs; these are used surprisingly seldom despite their utility.

One strategy involves restricting a search to only those solutions that are *Pareto optimal*. A solution is Pareto optimal if there is no other solution that is better under one criterion without being worse for the other criteria. It is often true that not every response has the same importance; for example, avoiding EPS symptoms might be 50-fold more important to a team than avoiding weight gain. Although Pareto optimality provides more than one solution, it does not allow different weightings on different criteria, as this is difficult to manage with more than two dimensions. Another useful approach is Derringer's desirability function which does allow weights to be assigned to each criterion (Derringer, 1980). The desirability function involves transformation of each criterion to a desirability value d, where $0 \leq d \leq 1$. The transformation is done in such a way that the value of d increases as the "desirability" of the corresponding criterion increases. This transformation may be linear, quadratic, step function, and so on. In the terminology of decision theory, these are monotonic utility functions. The individual desirabilities are then combined using a geometric mean, which is an overall assessment of the desirability of the combined response levels. It can be a weighted mean where the weights reflect relative importance of the criteria.

13 Discussion

Drug discovery is a challenging endeavor that involves many disciplines in the life sciences and informatics. There are a great many interesting and diverse problems that need to be solved. This chapter has given an overview of a number of them while omitting many others. Areas that are increasing in research intensity include the areas of genomics, gene chip microarrays (see Chapter 5), proteomics, metabolomics, and other technologies that involve spectral analysis. There are a host of interesting and challenging problems in these areas and, currently, there is great interest in merging these disciplines with the cheminformatics-related disciplines that have been the focus of this chapter.

The future will see dramatic changes in drug discovery and development processes. Within the next decade, researchers will almost certainly find most human genes and their locations. Explorations into the function of each one is a major challenge extending far into the next century and will shed light on how faulty genes play a role in disease causation. With this knowledge, commercial efforts will shift towards developing a new generation of therapeutics based on genes. Drug design will be revolutionized as researchers create new classes of medicines based on a reasoned approach using gene sequence and protein structure information rather than the traditional trial-and-error method. The drugs, targeted to specific sites in the body, will not have the side effects prevalent in many of today's medicines. Over 150 clinical gene therapy trials are now in progress in the United States, most for different kinds of cancers.

The road map of human biology generated by the human genome project will supply an enormous store of genes for studying, and ultimately curing, the ills that beset us. As the factors underlying the maladies of the human condition slowly come to light, the challenge will be to use the information effectively and responsibly.

Acknowledgments. I would like to thank Rick Higgs, Kerry Bemis, Bruno Boulanger, and Philip Iversen for fruitful discussions.

Appendix
Tools of the Trade

Robots: Used in a number of processes: screening compounds for biological activity, inoculating microbial cultures, and filling compound libraries.

High-throughput screening (HTS): Technology where robotics is used to test many compounds rapidly in an effort to identify novel inhibitors of receptors or enzymes. Usually 100,000 to 200,000 compounds are screened.

Medium-throughput screening (MTS): Similar to HTS but with only modest throughput requirements which implies more careful usage of robotics and higher quality of data. Typical MTS throughput is 1000 to 10000 compounds.

Combinatorial chemistry (Combi Chem): Used to make thousands of variants of a compound. Consider a compound with a six-membered aromatic ring and a chlorine atom attached at a certain position. One might change the location of the chlorine to any of the other five positions, or change the chlorine to a fluorine or a bromine, and/or make the same changes at all the other five positions. To enumerate all the possible combinations is to make a combinatorial library.

Genomic information: Used to identify possible protein therapies and targets, to develop biomarkers, and to understand more deeply how a given compound interacts with a complex living system.

X-ray crystallography, nuclear magnetic resonance (NMR): Used in exploring the physical properties/shape of a molecule and/or a receptor target. If the structure of the target is known, docking studies can be done to assess how well a molecule may "fit" in one of the receptor's binding sites.

Bioinformatics tools: Used to search enormous volumes of biological information, for instance, to find the best genomic match of a nucleotide sequence or learn the chromosomal location and disease linked to a particular gene. We may know that a compound evokes a biological response but with genomics and bioinformatics tools we can examine which proteins are affected by the compound.

Cheminformatics tools: Used to explore the relationship between the structure of a compound and the biological response it evokes (the SAR), with a view toward predicting what will happen with new, as yet untested compounds. Also used to model the docking of a small molecule (or ligand) to a protein or receptor.

References

Abt, M., Lim, Y., Sacks, J., Xie, M., and Young, S. S. (2001). A sequential approach for identifying lead compounds in large chemical databases. *Journal of Biomolecular Screening*, **16**, 154–168.

Amidon, G., Lennernäs, H., Shah, V., and Crison, J. (1995). A theoretical basis for a biopharmaceutic drug classification: The correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharmaceutical Research*, **12**, 413–420.

Bejamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

Birkett, D. J. (1990). How drugs are cleared by the liver. *Australian Prescriber*, **13**, 88–89.

Birkett, D. J. (1991). Bioavailability and first pass clearance. *Australian Prescriber*, **14**, 14–16.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, **26**, 801–849.

Breiman, L. (1999). Random forests, random features. *Technical Report*, University of California, Berkeley.

Breiman, L. (2001a). Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, **16**, 199–231.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. CRC Press, New York.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. Springer-Verlag, New York.

Campbell, C., Christianini, N., and Smola, A. (2000). Query learning with large margin classifiers. *Proceedings of ICML2000*, 8.

Comprehensive Medicinal Chemistry (2003). MDL Informations Systems, California.

Cook, R. D. and Nachtsheim, C. J. (1982). Model robust, linear-optimal design: A review. *Technometrics*, **24**, 49–54.

Crivori, P., Cruciani, G., Carrupt, P., and Testa, B. (2000). Predicting blood-brain barrier permeation from three-dimensional molecular structure. *Journal of Medicinal Chemistry*, **43**, 2204–2216.

Crum-Brown, A. and Fraser, T. R. (1869). On the connection between chemical constitution and physiological action. Part I. On the physiological action of the salts of the ammonium bases derived from strychnine, brucia, thebaia, codeia, morphia and nicotia. Part II. On the physiological action of the ammonium bases derived from atropia and conia. *Transactions of the Royal Society of Edinburgh*, **25**, 151–203; 693–739.

Cummins, D. J., Andrews, C. W., Bentley, J. A., and Cory, M. (1996). Molecular diversity in chemical databases: Comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *Journal of Chemical Information and Computer Sciences*, **36**, 750–763.

Dasarathy, B. (1991). *Nearest Neighbor Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA.

Derringer, G. and Suich, R. (1980). Simultaneous optimization of several response variables. *Journal of Quality Technology*, **12**, 214–219.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Science*, **14**, 436–440.

Drews, J. (2000). Drug discovery: A historical perspective. *Science*, **287**, 1960–1964.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

Engels, M. F. M., Thielemans, T., Verbinnen, D., Tollenaere, J. P., and Verbeeck, R. (2000). Cerberus: A system supporting the sequential screening process. *Journal of Chemical Information and Computer Sciences*, **40**, 241–245.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Technical Report 4*, U.S. Air Force, School of Aviation Medicine, Texas.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

Goldberg, J. and Wittes, J. (1978). The estimation of false negatives in medical screening. *Biometrics*, **34**, 77–86.

Hansch, C., Maolney, P. P., Fujita, T., and Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*, **194**, 178–180.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley and Sons, New York.

Hastie, T. and Tibshirani, R. (1996a). Discriminant adaptive nearest-neighbor classification. *IEEE Pattern Recognition and Machine Intelligence*, **18**, 607–616.

Hastie, T. and Tibshirani, R. (1996b). Discriminant adaptive nearest neighbor classification and regression. In *Advances in Neural Information Processing Systems*. Editors: D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, volume 8, pages 409–415, MIT Press, Cambridge.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hawkins, D. M., Basak, S. C., and Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, **43**, 579–586.

Higgs, R., Bemis, K., Watson, I., and Wikel, J. (1997). Experimental designs for selecting molecules from large chemical databases. *Journal of Chemical Information and Computer Sciences*, **37**, 861–870.

JMP (2003), Version 4.0.4. SAS Institute, North Carolina.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximum distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Kennard, R. and Stone, L. (1969). Computer aided design of experiments. *Technometrics*, **11**, 137–148.

Kramer, C. Y. (1956). Extensions of multiple range tests to group means with unequal numbers of replications. *Biometrics*, **12**, 309–310.

Leach, A. R. and Gillet, V. J. (2003). *Introduction to Chemoinformatics*. Kluwer Academic, Boston.

Maccs Drug Data Report (2003). MDL Informations Systems, California.

Major, J. (1999). What is the future of high-throughput screening? *Journal of Biomolecular Screening*, **4**, 119–125.

Miller, A. J. (2002). *Subset Selection in Regression*, second edition. Chapman & Hall/CRC, New York.

Phatarfod, R. M. and Sudbury, A. (1994). The use of a square array scheme in blood testing. *Statistics in Medicine*, **13**, 2337–2343.

Rohrer, S. P., Birzin, E., Mosley, R., Berk, S. C., Hutchins, S., Shen, D., Xiong, Y., Hayes, E., Parmar, R., Foor, R., Mitra, S., Degrado, S., Shu, M., Klopp, J., Cai, S. J., Blake, A., Chan, W. W. S., Pasternak, A., Yang, L., Patchett, A., Smith, R., Chapman, K., and Schaeffer, J. (1998). Rapid identification of subtype-selective agonists of the somatostatin receptor through combinatorial chemistry. *Science*, **282**, 737–740.

Rusinko, A., III, Farmen, M. W., Lambert, C. G., Brown, P. L., and Young, S. S. (1999). Analysis of a large structure/biological activity data set using recursive partitioning. *Journal of Chemical Information and Computer Sciences*, **39**, 1017–1026.

SAS System (2003), Version 8.2. SAS Institute, North Carolina.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486–494.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—A residual likelihood approach. *Journal of the Royal Statistical Society B*, **64**, 237–252.

Sittampalam, G. S., Iversen, P. W., Boadt, J. A., Kahl, S. D., Bright, S., Zock, J. M., Janzen, W. P., and Lister, M. D. (1997). Design of signal windows in high throughput screening assays for drug discovery. *Journal of Biomolecular Screening*, **2**, 159–169.

Tukey, J. W. (1994). Reminder sheets for "Allowances for various types of error rate". In *The Collected Works of John W. Tukey, volume VIII, Multiple Comparisons: 1948–1983.*. Editor: H. I. Braun, pages 335–339, Chapman & Hall, New York.

Tukey, J. W. (1997). More honest foundations for data analysis. *Journal of Statistical Planning and Inference*, **57**: 21–28.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience, New York.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*, second edition. Springer Verlag, New York.

Warmuth, M. K., Liao, J., Ratsch, G., Mathieson, M., Putta, S., and Lemmenk, C. (2003). Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, **43**, 667–673.

Weston, J., Perez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., and Schölkopf, B. (2002). Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics*, **1**, 1–8.

Wikel, J. H. and Higgs, R. E. (1997). Point: Applications of molecular diversity analysis in high throughput screening. *Journal of Biomolecular Screening*, **2**, 65–66.

World Drug Index (2002). Thompson Derwent, London.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–131.

Young, S. S., Ekins, S., and Lambert, C. G. (2002). So many targets, so many compounds, but so few resources. *Current Drug Discovery*, 1–6 (www.currentdrugdiscovery.com).

Zemroch, P. J. (1986). Cluster analysis as an experimental design generator. *Technometrics*, **28**, 39–49.