Title
Mining high-throughput screening data to accelerate drug lead discovery

Author
Shelat, Anang A

Publication Date
2005

Peer reviewed|Thesis/dissertation

Mining High-Throughput Screening Data to Accelerate Drug Lead Discovery

by

Anang A. Shelat

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry and Chemical Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Date University Librarian

Degree Conferred:.....

Acknowledgements

I am deeply indebted to a number of collaborators who supplied much of the experimental data used in this work. Hong Yang and Vadiraj Gopinath were responsible for the ΔF508 potentiator results described in Chapter 1. The enzyme-based and DLS screens for promiscuous inhibition reported in Chapter 3 was primarily conducted by Brian Feng. Raw HTS data for the Assay Explorer detailed in Chapter 4 was provided by Leggy Arnold, Zachary Mackey, Jennifer Weisman, Ally P. Liou, and Nick Mills. From the Bay Area Screening Center, Brian Wolff, Mike Mike Uehara-Bingen, and Janice Williams provided technical expertise.

Mark Segal was instrumental in providing me with an excellent guide to the world of computational learning algorithms; I appreciate his thoughtful advice and patience. My two advisors, Tack Kuntz and Kip Guy, have not only taught me the art of science, but have also been fine examples of how to balance being a successful scientist with living a rich and complete life.

I am most grateful to Nandini Gandhi for her skillful editing, infinite patience, and boundless friendship. Finally, I would like to thank Ben, Abhi, Mom, and Dad for their continuous love and support.

Mining High-Throughput Screening Data to Accelerate Drug Lead Discovery

Anang A. Shelat

Data mining has two main objectives: (a) to describe patterns and relationships among existing pieces of information and (b) to predict these patterns and relationships in future data. The pioneering work of Hansch and Fijuta during the 1960s represents the first application of this technique to quantitatively express biological activity as a function of chemical properties (Wermuth, 2003). These quantitative structure activity relationships, or QSARs, provided valuable insight into the contributions of chemical moieties, and quickly became an important tool for drug discovery and design.

In this work, we explore how data mining methods can be applied to high-throughput screening (HTS) data in order to accelerate the discovery of suitable lead compounds. In Chapter 1, we describe the construction of a naïve Bayes classifier to identify active molecules in a screen for potentiators of ΔF508 CFTR, one of the mutant forms of the CFTR gene responsible for Cystic Fibrosis. This work, originally reported in Yang et al (Yang, 2003) is expanded and updated to reflect how the interpretability of the model helped inspire pharmacophore-based hypotheses which guided subsequent optimization.

Chapters 2 and 3 describes the development and application of a consensus model for predicting promiscuous inhibitors—molecules that nonspecifically disrupt HTS assays. High-throughput experimental assays and a preliminary computational model were reported earlier by Feng et al (Feng, 2005). We present new analysis that questions the reliability of one of the high-throughput models, and detail a novel consensus modeling method that achieves greater predictive performance by aggregating the results

from two philosophically different algorithms—support vector machines and generalized boosting machines.

In Chapter 4, we discuss the need to approach drug discovery as a multivariate optimization problem and describe the Assay Reporter—an informatics platform that integrates both chemical and biological information in order to identify "good" molecules suitable for further development. Future plans to introduce elements of computational learning into the Assay Reporter framework are detailed in Chapter 5.

Table of Contents

List of Tables

List of Figures

Chapter I: Elucidation of the Structure-Activity Relationship for Tetra-substituted Thiophenes That Target the ΔF508 Cystic Fibrosis Gene using a Naïve Bayes Classifier

Modified from Yang H, Shelat AA, Guy RK, Gopinath VS, Ma T, Du K, et al. Nanomolar affinity small molecule correctors of defective delta F508-CFTR chloride channel gating. *J Biol Chem.* 2003 Sep 12; 278(37):35079-85.

In early lead discovery, understanding how the properties of molecules influence biological activity is sometimes more useful than building the most accurate predictive model. Knowledge of the role played by chemical moieties or physical attributes helps chemists formulate verifiable hypotheses which guide the efficient exploration of chemical scaffolds.

In this chapter, we revise and update our contribution to the work in Yang et al that reports the discovery of small molecule activators targeting the ΔF508 Cystic Fibrosis gene (Yang, 2002). Specifically, we provide an extended discussion of the application of a *naïve* Bayes classifier to mine high throughput screening (HTS) data from that drug discovery project. This particular model structure was chosen for two reasons. First, the binary nature of the HTS output ("active" vs. "inactive") necessitated the use of a classifier. Second, and more importantly, the algorithm reported the contribution of each descriptor as an additive log-odds ratio for favoring activity. By limiting the independent variables in the study to well-known medicinal chemistry properties and chemical functional groups, our model provided chemists with an understandable road-map to further define structure-activity relationships (SAR) for the validated tetra-substituted thiophene hit. A description of the hypotheses generated, the evidence accumulated using data mining techniques, and conclusions regarding the current state of the SAR for thiophene potentiators are reported here.

1.1 Background on Cystic Fibrosis

The Cystic Fibrosis Transmembrane Receptor (CFTR) mediates cyclic-AMP (cAMP) dependent Chloride ion secretion in the apical membrane of cells lining mammalian airways and other luminal surfaces. Mutations in the CFTR gene cause Cystic Fibrosis (CF), the most prevalent lethal hereditary disease among Caucasians (Yang, 2002). According to the CF Genetic Analysis Consortium, approximately 1400 mutations altering Cl^- secretion in sweat glands, the pancreas, intestines, reproductive organs, and airways have been identified (http://www.genet.sickkids.on.ca). However, 90% of all patients with CF have at least one allele containing ΔF508, a mutation which causes protein misfolding, disrupts protein trafficking to the membrane, and prevents proper channel gating (Bobadilla, 2002). CFTR activity in these patients is severely diminished, and most succumb to respiratory failure due to fluid buildup in the lungs and subsequent infection.

Thus, ΔF508 Cystic Fibrosis presents two targets for small molecule therapeutics: "correctors" which stabilize the mutant protein and increase transport to the membrane, and "potentiators" which resolve the channel gating defect. Compounds possessing these types of activity, though rare, have been identified in other systems. Loo and Clark described substrates of mutant human P-glycoprotein that corrected defective protein kinesis; interestingly, both P-glycoprotein and CFTR are members of ATP Binding Cassette (ABC) protein superfamily (Loo, 1997). Indeed, compounds such as alkylxanthines and the isoflavone genistein have been shown to activate ΔF508 and other mutant forms of CFTR; however, the required concentrations far exceeded reasonable

physiological values (1 mM and >50 μM, respectively) and the maximum achievable activities (V_{max}) were less than wild-type (Schultz, 1999).

Unfortunately, the complex nature of CFTR regulation in cells has confounded mechanistic studies of existing CFTR modulators and hindered the development of better, more specific compounds. The activity of the protein can be perturbed by modifying and of the following (Schultz, 1999): (a) the G-protein coupled receptor pathway upstream of the channel (i.e., binding at the receptor, activation of adenyl cyclase, inactivation via phosphodiesterases), (b) the protein kinases and phosphatases important in CFTR regulation, (c) the Na^+-K^+-ATPase and Na^+-K^+-$2Cl^-$ channels that work together to maintain the Cl^- concentration gradient, (d) the K^+ channels that maintain the membrane potential necessary to drive Cl^- out of the cell, and (e) the overall metabolic state which governs the ATP:ADP ratio (nucleotide binding and hydrolysis are required for channel gating). Furthermore, CFTR currents must be distinguished from the current conducted by other Cl^- channels.

Despite these difficulties, Yang et al were able to develop an HTS assay for small molecule potentiators of ΔF508 (Yang, 2002). The experiment entailed transfecting Fischer Rat Thyroid (FRT) cells with the truncated CFTR gene and a halide-sensing version of Green Fluorescent Protein (YFP-H148Q/I152L) (Galietta, 2001). Cells were grown at reduced temperature (27°C) to correct for CFTR misfolding and trafficking problems. A stable line with high fluorescence and suitable levels of CFTR at the membrane was cultured and then transferred into 96 well plates. The mutant CFTR was activated by adding forskolin (20 μM final concentration), a compound which raises cyclic AMP (cAMP) via adenyl cyclase induction, prior to screening. Under these

conditions any influx of halide, as measured by the change in fluorescence intensity of YFP-H148Q/I152L, should be attributable to the CFTR channel.

A collection of 100,000 commercially available compounds from the ChemBridge company were screened in this manner at 2.5 μM. Seventy-five compounds representing six distinct chemical scaffolds were identified as "strong" potentiators (halide influx > 0.1mM/s), 252 were classified as "weak" potentiators (detectable change in halide influx), and the rest were inactive. The 75 strong hits were then subject to secondary analysis. None of these compounds stimulated halide influx in FRT cells transfected with YFP-H148Q/I152L alone or ΔF508 expressing cells in the absence of forskolin stimulation. The halide current for each compound was blocked by the CFTR inhibitor, $CFTR_{inh}$-172 (Ma, 2002). Dose response studies yielded 32 compounds that activated the mutant CFTR channel with V_{max} greater than 50 μM genistein and $EC_{50} < 1$ μM, indicating the potential for therapeutic benefit at acceptable physiological concentrations. Short circuit current analysis was performed for these 32 molecules to verify that the Cl^- influx occurred through the apical membrane: 13 of the 32 compounds produced currents comparable to genistein, but with $EC_{50} < 2$ μM.

An additional 1000 analogs to the 6 structural classes were purchased from ChemBridge to further define structure-activity relationships. However, only one class, the tetra-substituted thiophenes, afforded active compounds. Representatives of the six scaffolds identified earlier and the best thiophene analogs underwent additional secondary screens. None of these compounds produced a significant increase in cellular cAMP, inhibited phosphatase under conditions where the phosphatase inhibitor okadiac

acid inhibited activity > 90%, or resulted in significant cellular toxicity as measured by the dihydrorhodamine assay (Wang, 2002) and by unimpaired cell growth.

Thus, considerable evidence suggested that the tetra-substituted thiophenes were *bona fide* ΔF508 potentiators. The activity of compounds belonging to the remaining five scaffolds was unclear and warrants further investigation. The absence of activity in chemically similar structures cast doubt on the initial findings, although this result could be attributable to an incomplete or limited analog series. Nevertheless, the thiophene scaffold was selected for further development. In order to efficiently guide medicinal chemistry efforts, a computational analysis using a naïve Bayes classifier was performed on this series.

1.2 Naïve Bayes Classifiers[1]

For any generic classification problem given an unknown object with measurement vector **x** (e.g., a molecules with descriptors), the probability of belonging to class k out of M possible classes can be estimated using Bayes theorem:

$$p(c_k \mid \mathrm{x}) \propto p(c_k)p(\mathrm{x} \mid c_k), \text{ where } 1 \leq k \leq M \quad (1\text{-}1)$$

Here, $p(c_k)$ is the "prior" probability, or simply the proportion of all objects in class k. If the class distribution is unknown, then these values can be estimated from a sample of the population (i.e., the training set), or a "flat" prior can be used to assign equal probability to all classes. However, the class conditional probability, $p(\mathbf{x} \mid c_k)$, is much harder to

[1] The equations in this section are adapted from Hand, 2001.

calculate because it requires knowledge of the joint distribution of the p variables in **x** for each class:

$$p(\mathrm{x} \mid c_k) = \prod_{j=1}^{p} p(x_{k,j} \mid x_{k,i \neq j}), \text{ where } 1 \leq k \leq M \qquad (1\text{-}2)$$

In general, the density of variable x_j will be conditional on the other variables in **x**, as reflected in the right-hand expression in Equation 1.2. The number of parameters describing these densities increases as $O(M^p)$, or exponentially as a function of the number of descriptors. For example, a two-class problem with a measurement vector **x** comprised of 10 binary descriptors requires $2^{10} = 1024$ parameters to derive the conditional probabilities for every variable. Most real world applications lack sufficient data to estimate values in this way.

To avoid the explosion in parameters as p increases, the *first-order* or *naïve* Bayes algorithm assumes that each descriptor is conditionally independent within a class (Hand, 2001). Equation 1-2 then reduces to:

$$p(\mathrm{x} \mid c_k) = \prod_{j=1}^{p} p(x_j \mid c_k), \text{ where } 1 \leq k \leq M \qquad (1\text{-}3)$$

Under these conditions, the number of parameters required to estimate all class conditional probabilities increases as $O(Mp)$, or linearly in M and p. Computing these probabilities for discrete variables simplifies to counting the number of occurrences of each value and dividing by the total number of occurrences for each class. Density estimation techniques, such as approximating the data as a Gaussian distribution or using

non-parametric kernel functions, can be applied to derive the class conditional probabilities for continuous variables.

By combining Equations 1-1 and 1-3, the probability that an unknown object described by **x** belongs to class k becomes:

$$p(c_k \mid \mathrm{x}) = p(c_k)\prod_{j=1}^{p} p(x_j \mid c_k), \text{ where } 1 \le k \le M \qquad (1\text{-}4)$$

Moreover, the likelihood of belonging to class 1 relative to class 2 is equal to the ratio of the conditional probabilities: $p(c_1 \mid \mathbf{x}) / p(c_2 \mid \mathbf{x})$. Alternatively, this ratio can be expressed as log-odds:

$$\log\frac{p(c_1 \mid \mathrm{x})}{p(c_2 \mid \mathrm{x})} = \log\frac{p(c_1)\prod_{j=1}^{p} p(x_j \mid c_1)}{p(c_2)\prod_{j=1}^{p} p(x_j \mid c_2)} = \log\frac{p(c_1)}{p(c_2)} + \sum_{j=1}^{p}\log\frac{p(x_j \mid c_1)}{p(x_j \mid c_2)} \qquad (1\text{-}5)$$

Equation 1-5 underscores the intuitive nature of the naïve Bayes algorithm: the likelihood of belonging to one class relative to another is equal to the sum of a constant, which reflects the baseline probability for that class, and the log-odds of having a particular value for each descriptor. The additive, probabilistic nature of the descriptor contributions facilitates interpretation. Positive values for the second term on the right-hand side of Equation 1-5 indicate that that value for the descriptor is more likely to be found in objects from class 1 relative to class 2; negative values indicate the opposite, and values close to zero suggest negligible or no influence on classification. Thus, the domain of each descriptor is partitioned into regions favoring one class over another. In

the context of medicinal chemistry, these regions provide targets for identifying new compounds or further optimizing the activity of an existing molecule.

Unfortunately, the assumption that variables are conditionally independent within each class is usually not valid; in fact, chemical descriptors are often highly correlated. This characteristic can artificially inflate the conditional probabilities in the model, leading to an inaccurate assessment of the covariate log-odds and an over-estimation of predictive power. To preserve interpretability, the descriptor set could be limited to fairly orthogonal variables, as was done in this study. Alternatively, if the latter problem is more important, then variable selection schemes such as the one described in Chapter 3 could be useful. With proper manipulation, then, naïve Bayes classifiers will generally afford satisfactory performance. The models are resistant to the influence of outliers and noise due to the smaller number of required parameters (these models have less variance as described in Chapter 2) (Hand, 2001). Furthermore, errors in estimating the conditional probabilities are less important, because most classifications depend on the sign—not the magnitude—of the log odds on the left-hand side of Equation 1-5 (Hand, 2001).

Thus, the naïve Bayes classifier offers powerful interpretive and predictive power. The following sections describe an application of the method to elucidate SAR information for the lead compound series identified from an HTS for ΔF508 CFTR potentiators.

1.3 Data Modeling Methods

Data manipulations, property calculations, and model building were performed using Pipeline Pilot v. 4.0 (Scitegic, Inc.). All graphs were created in Microsoft Excel or the R statistics package (version 2.1.1). The data set for modeling consisted of all tetra-substituted thiophenes from the 100,000 ChemBridge library and 1,000 ChemBridge analogs screened in a follow-up study ($N = 3025$). Of the 3025 molecules, forty were classified as "active" based on their performance in the HTS assay and subsequent secondary analysis; the remaining molecules were labelled "inactive."

The "Learn Good Molecules" component in Pipeline Pilot provided an algorithm for the naïve Bayes classifier. This protocol estimated the class conditional probabilities for continuous variables by first partitioning the range into bins (in a process similar to constructing a histogram), and then calculating the log-odds values for each bin. For discrete descriptors, a bin was created for each value in the data set. The Laplacian correction was applied to avoid bins with probabilities that are either equal to zero or skewed by small sample sizes. The algorithm also assumed an equal prior probability for all classes. The model included six physical property descriptors (molecular weight, surface area, polar surface area, number of H-bond donors, number of H-bond acceptors, and AlogP), and binary variables ("bits") derived from Pipeline Pilot's functional class fingerprints with a diameter of 6 bonds (FCFP_6). Each fingerprint bit represents a unique chemical pattern derived from the training set molecules and is set to either 0 or 1 depending on whether the pattern is absent or present.

Four-fold cross-validation (75% training, 25% test) was employed to train the classifier to distinguish between active and inactive tetra-substituted thiophenes, and

produced four models. Each model output a score proportional to the probability of belonging to the active class. The Mann-Whitney statistic for non-parametric two-group comparisons was used to assess the likelihood that the distributions of model scores for active and inactive tetra-substituted thiophenes in the test sets represented different populations. Following cross-validation, a final model using all 3025 molecules was generated and used to analyze the contributions of the descriptors. A table of log-odds values for each descriptor was supplied by the "Learn Good Molecules" component. Favourable and unfavourable chemical bits were translated into structural elements using Pipeline Pilot's fingerprint manipulation routines. A congeneric series for structure-activity analysis was generated by removing the R-group from each active compound, and using the resulting scaffold to perform substructure queries on the entire tetra-substituted thiophene set. Student's T-test was employed to assess whether differences in the physical property distributions of R-groups from active and inactive thiophenes were significant.

1.4 Results and Discussion

As a first step in lead optimization, a computational model relating ion transport activity to structural and physico-chemical parameters of the tetra-substituted thiophene class of ΔF508-CFTR potentiators was generated using a naïve Bayesian classifier methodology. All four models constructed during cross-validation clearly segregated active and inactive compounds (Mann-Whitney $p < 0.00001$ and Receiver-Operator Characteristic [ROC] AUC > 0.98, regardless of originating training set). The

distribution of model scores and the ROC curve for the poorest performing model are shown in Figure 1-1.

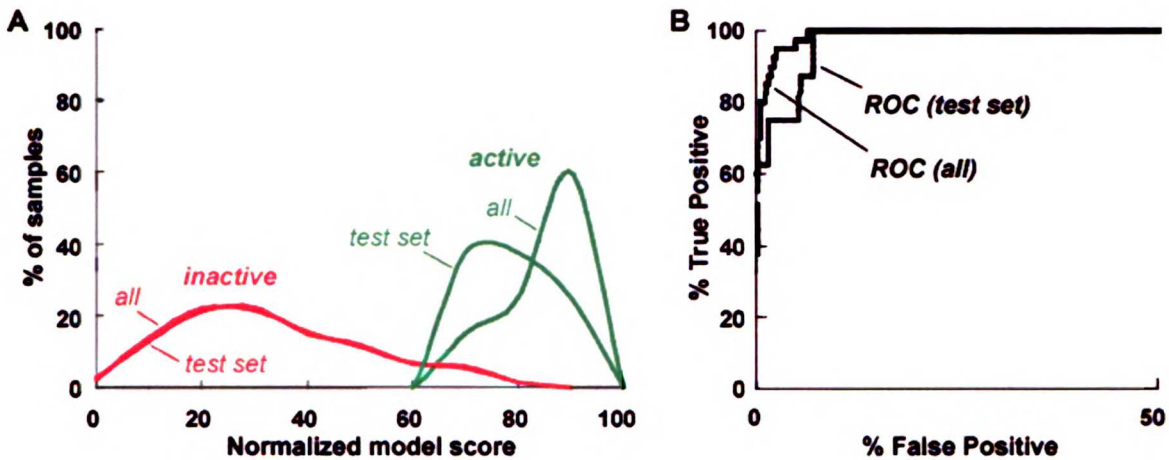
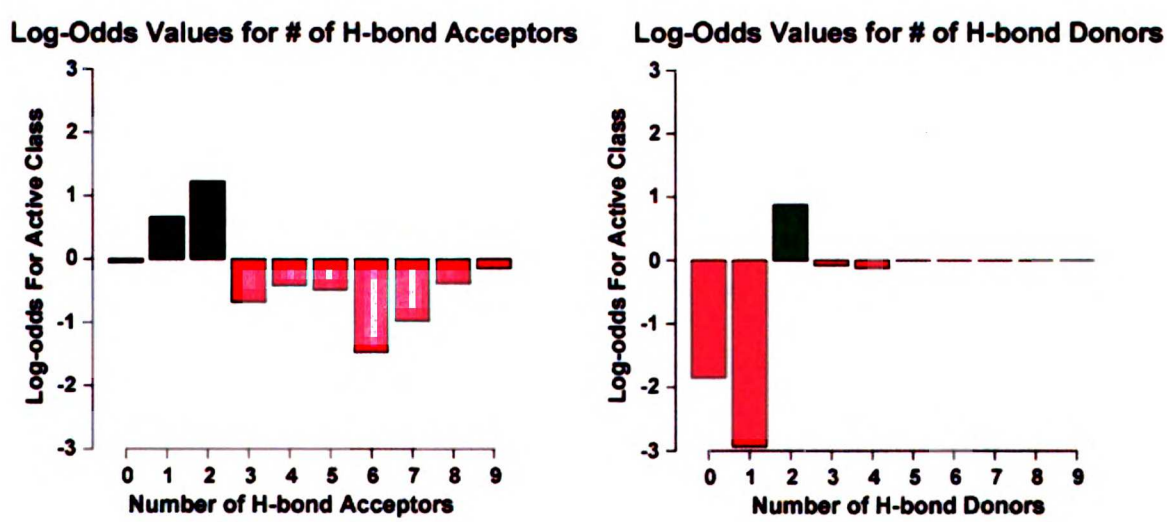


Figure 1-1. Performance of the naïve Bayes classifier. (A) The poorest performing model effectively differentiated active and inactive tetra-substituted thiophenes in the test set and for all tetra-substituted thiophenes studied (Mann-Whitney, $p < 0.00001$). The model score has been normalized to the range 0-100. (B) The AUC of the Receiver-Operator Characteristic (ROC) for the test set and all tetra-substituted thiophenes are 0.98 and 0.99, respectively. The ROC AUC for random and perfect models is 0.50 and 1.0, respectively.

A preliminary SAR emerged following an analysis of the contributions of physical properties and binary descriptors to the naïve Bayesian model. As reported below in Figure 1-2, five of the six physical property descriptors possessed narrow, well defined regions that were enriched for potentiators.



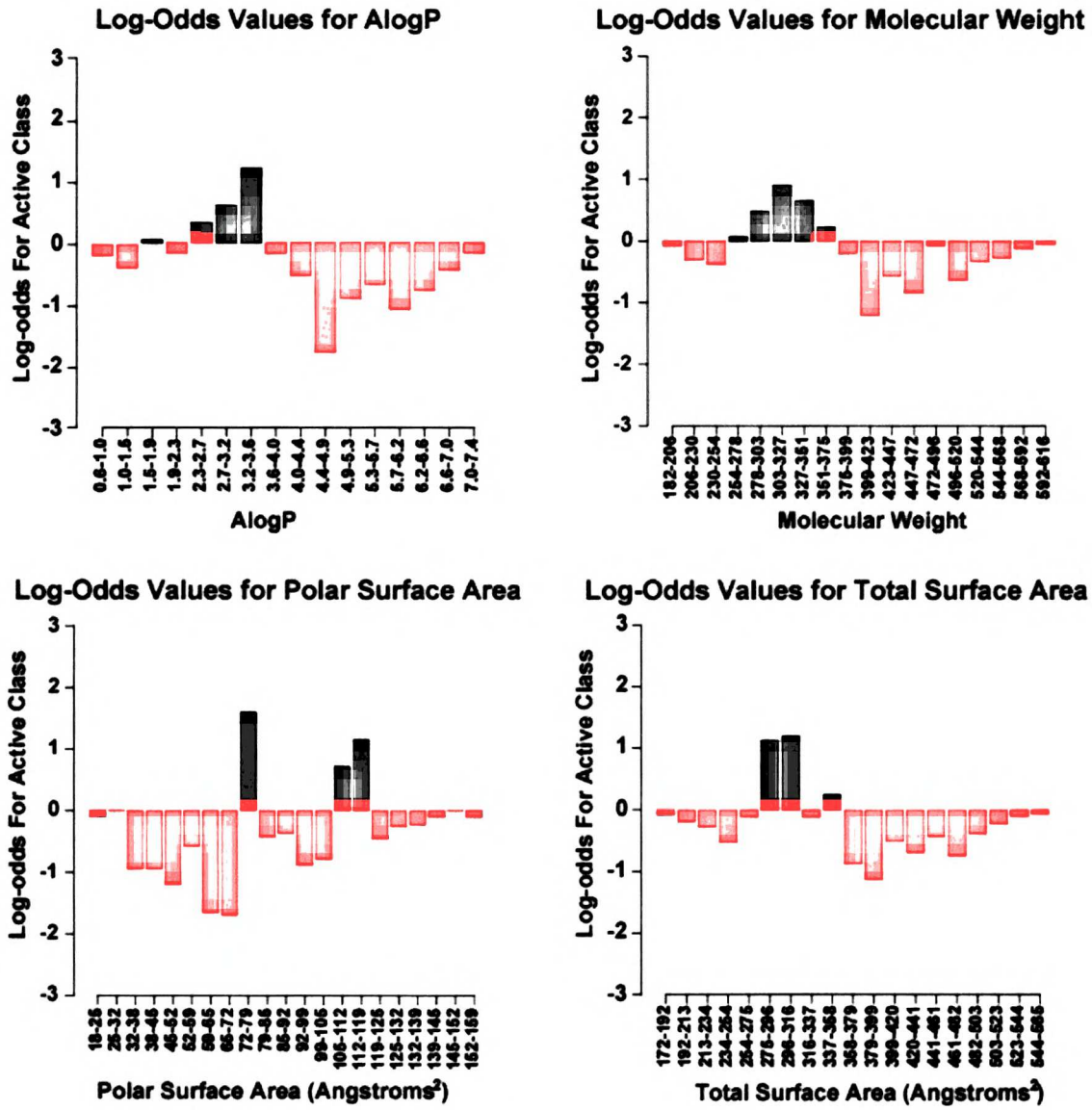


Figure 1-2. Log-odds values for the active class computed for the six physical properties explored in the naïve Bayesian model. Compounds falling within a range with log-odds values greater than zero (the green bars) are more likely to be active; those falling within ranges with values less the zero (the red bars) are less likely to be active. Property ranges with log-odds close to zero suggest either no influence on activity or a lack of information due to small sample size.

Active tetra-substituted thiophenes are more likely to have no more than two H-bond acceptors (top right), at least two H-bond donors (top left), AlogP between 2.3 and 3.6 (center right), molecular weight between 254 and 375 Daltons (center left), and total surface area between 275 and 358 $Å^2$ (bottom right). The log-odds plot for polar surface area (bottom left) suggests that regions favoring activity follow a bimodal distribution

with peaks around 72-79 and 105-119 $Å^2$. Interestingly, inspection of the molecules in the 105-112 and 112-119 bins reveals that a single functional group—a nitro, sulfonamide, or phthalate belonging to the R-group at the 2-position on the thiophene ring—accounts for the higher values. Thus, the amount of polar surface area in the core thiophene scaffold appears to be constrained to the range 72-79 $Å^2$.

Decomposing the contribution of the FCFP_6-based descriptors to the naïve Bayesian model revealed favorable and unfavorable structural elements. Figure 1-3 depicts some of the interesting fingerprint bits as colored spheres superimposed onto representative structures to illustrate their meaning. For dark green and dark red spheres, the pattern contains the element type and hybridization of the underlying atom; the light green and light red spheres allow for any heavy atom.

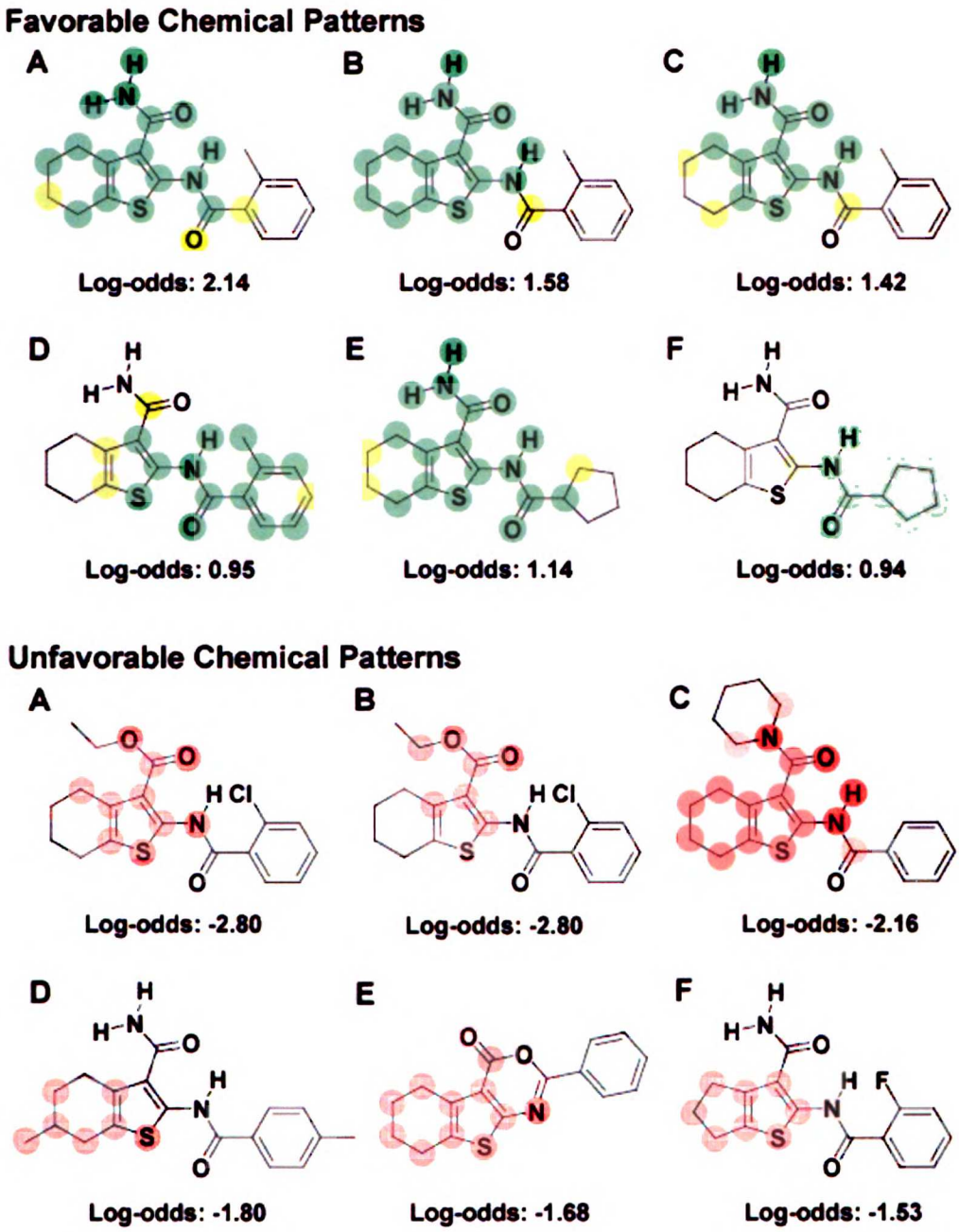


Figure 1-3. Favorable and unfavorable structural elements superimposed onto representative molecules. The chemical pattern for dark green or red colored spheres follows the element type and hybridization of the underlying atom; light green or red spheres implies that any heavy atom is acceptable.

The Favorable bits A-C suggest that thiophene rings possessing a secondary amine at the 2-position, an unsubstituted amide at the 3-position, and an aliphatic fused ring at the 4

and 5 positions are more likely to be active. Favorable patterns D-F indicate that an aromatic or aliphatic substituted carboxamide at C2 promotes activity. In contrast, unfavorable elements A-C signal that molecules with either an ester or a di-substituted amide at C3 are more likely to be inactive. Furthermore, active molecules are less likely to possess a cyclopentyl or branched aliphatic ring fused to positions 4 and 5 (patterns D and F), or a cyclization between R-groups at positions 2 and 3 (pattern E).

Thus, the naïve Bayes model afforded early insight into how physical properties and chemical structures influence thiophene potentiator activity. These findings provided a roadmap for defining more concrete structure-activity relationships. Using knowledge of the favorable and unfavorable elements, hypotheses regarding the thiophene pharmacophore were proposed. For each supposition, supporting or contrary evidence was collected by generating congeneric series (molecules that differ at only one position) via complex substructure query techniques applied to the entire thiophene data set. The hypotheses, exemplar data, and conclusions are presented below:

Hypothesis 1: H-bond donor and acceptor functionality is required for activity at the 3-position on the thiophene ring.

Data Mining Results for Hypothesis 1:

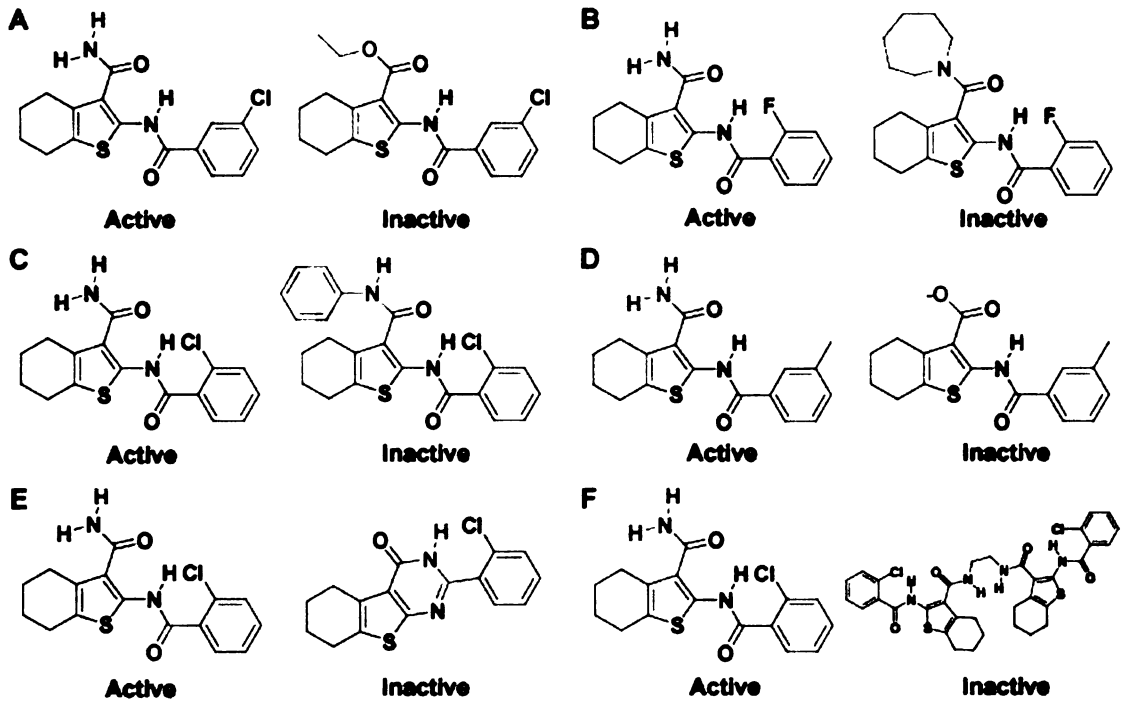


Figure 1-4. Examples of congeneric series exploring the 3-position of the thiophene ring. The variable positions in the molecules are delineated by bluish-gray spheres.

Unsubstituted amides at the 3-position—but not di-substituted amides or esters—favor activity according to the Bayesian model. Indeed, data mining revealed numerous congeneric series similar to panels A and B in Figure 1-4 that support this assertion. The lack of H-bond donor capability might account for the loss of activity; however, inactivity could also be the result of the additional steric bulk. For example, both the mono-substituted amide in panel C and the bis-compound in panel F are inactive, despite providing an H-bond donor. The inactive carboxylic acid in panel D suggests that at least one H-bond donor is required at C3 of the thiophene ring, although desolvation penalties for burying a negative charge may also be important. Thus, the evidence regarding the necessity of an H-bond donor at C3 is substantial, though not conclusive.

The orientations of the H-bond donor and acceptor also appear to be important. Molecules in which the 3-position amide has cyclized with a carbonyl group at the 2-position (panel E) possess H-bond donor and acceptor functionality that is conformationally locked. The constraints imposed by the new ring might account for the inactivity, although the importance of other changes to the scaffold (e.g., loss of the carbonyl at C2) cannot be ruled out.

Unfortunately, no evidence was available to assess the significance of the H-bond acceptor. The methyl ketone and the alcohol resulting from the reduction of the ketone would be logical analogs to assay. The ketone, which is neutral and isosteric with the amide, could provide conclusive evidence regarding the H-bond donor requirements. The alcohol could test the requirement for an H-bond acceptor.

In summary, the C3 side-chain must be roughly isosteric with an amide group and is likely to require an H-bond donor. The function of the carbonyl group is unclear. Two additional molecules, the methyl ketone and its reduced alcohol, were proposed to confirm the H-bond donor and acceptor requirements.

Hypothesis 2: The 4 and 5 positions must provide sufficient hydrophobic bulk.

Data Mining Results for Hypothesis 2

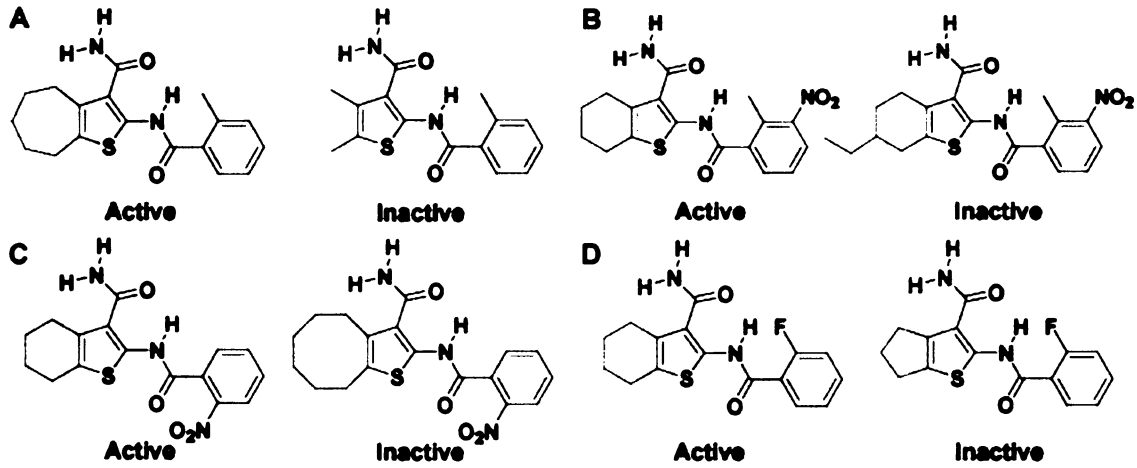


Figure 1-5. Examples of congeneric series exploring the 4 and 5-positions of the thiophene ring. The variable positions in the molecules are delineated by bluish-gray spheres.

The Bayesian model suggests that a six-membered, unsubstituted aliphatic ring fused at C4 and C5 of the thiophene ring promotes activity, whereas a cyclopentyl ring is unfavorable. Further evidence from data mining indicates that either a six or seven-membered, unsubstituted ring is favorable at this position (Figure 1-5). The loss of activity associated with the eight-membered ring (panel C) and the substituted, six-membered ring (panel D) could be rationalized in terms of excessive steric bulk. However, the cyclopentyl system in panel D and the dimethyl substituted thiophene in panel A are difficult to interpret. The three dimensional structures of these two molecules are unlikely to produce additional van der Waals clashes relative to the larger ring systems. Perhaps the hydrophobic binding surface provided by these molecules is not sufficient for activity. Alternatively, issues resulting from the cell-based nature of the HTS, such as limited cell permeability or faster metabolism, might be important.

Interestingly, the data set contained no benzothiophenes or thiophenes with heterocycles at the 4 and 5 positions. If the hydrophobic pharmacophore hypothesis is correct, then ether oxygen or sulfur in the fused rings might be tolerated; in contrast,

nitrogen atoms would be unfavorable because they would be positively charged and/or possess unsatisfied H-bond donors or acceptors. The benzothiophenes could be a promising scaffold because they satisfy the hydrophobic requirements and provide greater flexibility in terms of synthetic options and functionality.

Thus, the 4 and 5 positions of the thiophene ring appear to provide a hydrophobic binding surface that is optimal for unsubstituted six and seven-membered rings. No evidence was available for the effect of heteroatoms in these rings, or for the activity of benzothiophenes.

Hypothesis 3: The 2-position requires a hydrophobic binding moiety and an H-bond acceptor.

Data Mining Results for Hypothesis 3:

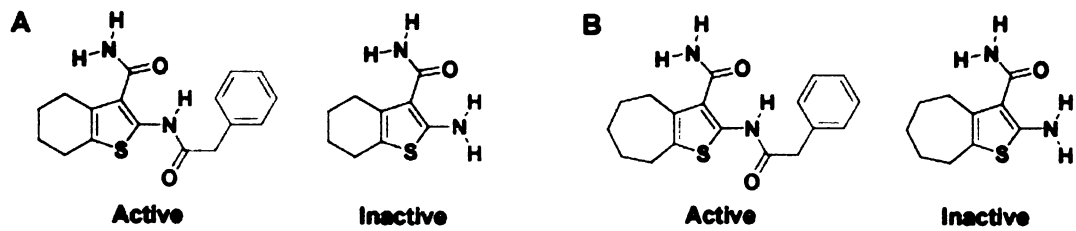


Figure 1-6. Examples of congeneric series exploring the 2-position of the thiophene ring. The variable positions in the molecules are delineated by bluish-gray spheres.

The naïve Bayes model predicts that carboxamides substituted with aliphatic or aromatic groups are favorable at C2 of the thiophene ring. Unfortunately, data mining revealed that the thiophene set contains few variations other than carboxamide. Figure 1-6 reports the only congeneric series obtained for this position. An unsubstituted amine is not tolerated at the 2-position, suggesting that the additional binding contacts afforded by the carbonyl and/or its substituent are essential for activity. The nature of the

carboxamide R-group was further explored by analyzing the distribution of physical properties for active and inactive molecules. Figure 1-7 describes histograms for the molecular volume and polar surface area of inactive, active, and the best active tetra-substituted thiophenes (molecules with sub-micromolar EC_{50}). The carboxamide side chains of active compounds appear to have smaller volumes and less polar surface area relative to inactive molecules (T-test $p<0.04$ and $p<0.0025$, respectively). These trends are even more apparent for the "best active" compounds. As noted earlier, the second peak in the polar surface area histogram for active molecules results from the presence of a single, uncharged functional group (a nitro, sulfonamide, or phthalate moiety), and might be considered an outlier.

In summary, C2 of the thiophene ring is not well explored. Data mining suggests that an H-bond acceptor and a relatively non-polar moiety (PSA < 20 $Å^2$) with volume <120 $Å^3$ are favorable for activity.

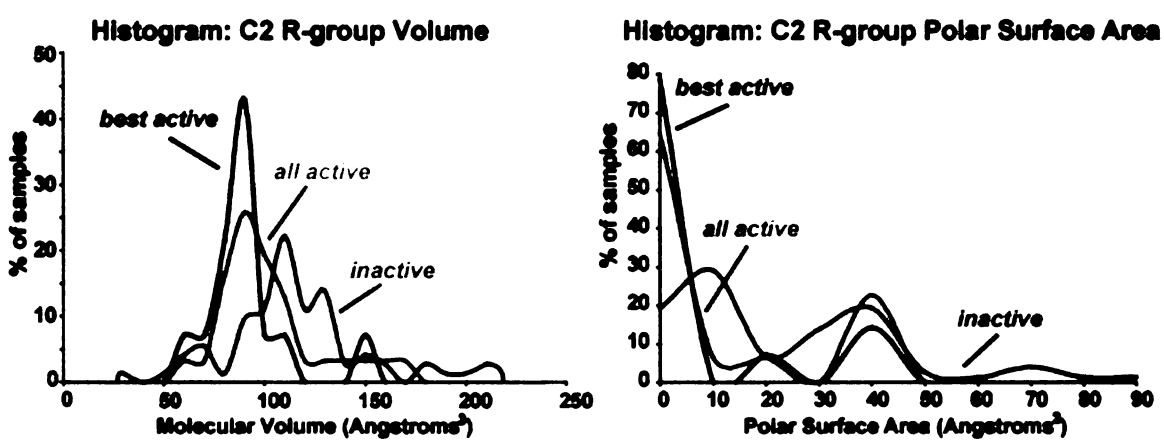
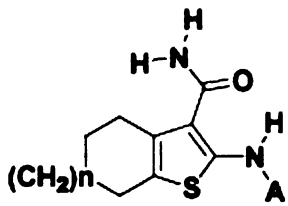


Figure 1-7. Distributions of molecular volume and polar surface area for the C2 R-group of tetra-substituted thiophenes. The "all active" and "inactive" tetra-substituted thiophenes are statistically distinguishable (T-test $p<0.04$ and $p<0.0025$, respectively). Only molecules with sub-micromolar EC_{50}s were included in the "best active" set. The peak centered around 40 $Å^2$ in the polar surface area histogram for active molecules results from a single nitro, sulfonamide, or phthalate moieties.

1.5 Consensus Model for Thiophene ΔF508 Potentiators

The minimal consensus substructure and optimal physical property ranges for active tetra-substituted thiophenes, based on an analysis of the naïve Bayes model and evidence collected from subsequent data mining, is described in Figure 1-8.



n = 1 or 2

Physical Property Range
2.3 ≤ ALogP ≤ 3.6
254 ≤ MW ≤ 375 daltons
275 ≤ surface area ≤ 358 A^2
72 ≤ polar surface area ≤ 119 A^2
H-bond acceptors ≤ 2
H-bond donors > 1

Figure 1-8. Extracted minimal consensus substructure and optimal physical property ranges for active tetra-substituted thiophenes.

The substructure allows for variation in the composition of the aliphatic ring fused at the 4 and 5-positions of the thiophene ring, but requires an amide at C3 and a relatively non-polar aliphatic or aromatic group appended to the nitrogen at C2. The physical properties of active tetra-substituted thiophenes fall within a narrow subset of the classic Lipinski parameters.

1.6 Conclusion

Computational learning models based on the naïve Bayes classifier algorithm offer facile interpretability and reasonable predictive performance. In this chapter, we described an application of the method to the tetra-substituted thiophene class of ΔF508

CFTR potentiators, yielding cross-validated models that successfully differentiated active and inactive molecules. Knowledge of the importance of the physical properties and structural descriptors extracted from the model helped formulate hypotheses about the thiophene pharmacophore. Subsequent data mining efforts established a concrete SAR that will efficiently guide the synthesis of new compounds.

Chapter II. Data Mining in Higher Dimensions: A Review of Theory and Algorithms

In the previous chapter, we described an application of the naïve Bayes algorithm that afforded both reasonable predictive accuracy and powerful interpretability. By focusing on a limited set of descriptors known to be relevant in SAR studies *a priori* (e.g., molecular weight, AlogP, number of hydrogen-bond donors, etc), we built a model predicting ΔF508 activity that was easily understood and communicated. This aspect facilitated the development of pharmacophore hypotheses, and helped guide further synthesis and screening.

However, our modeling strategy was simplified by an important aspect of the ΔF508 data set: all molecules were constrained within the tetra-substituted thiophene scaffold. The lack of chemical diversity effectively held many variables constant or limited their range, allowing us to better discern the contributions of our descriptors. In the next two chapters, we detail the construction and performance of computational learning models that reliably predict promiscuous inhibition across multiple, diverse scaffolds. Since the physical basis for the chemical phenomena was poorly understood, we lacked knowledge of the suitable variables to use in our model. As a result, we employed all available chemical descriptors—a set with a size on the on the order of the number of experimental data points. Furthermore, the diverse nature of the training data suggested the likelihood of nonlinear relationships between the descriptors. Under these model building conditions, two fundamental concerns in statistical learning became paramount: the *curse of dimensionality* and *model bias*.

In this chapter, we discuss the nature and consequences of these two concepts, and explore how modern statistical learning algorithms attempt to mediate them. The

equations employed here are derived from standard references (Hastie, 2001; Hand, 2001), whereas the interpretations and commentary reflect our understanding unless otherwise noted.

2.1 The Curse of Dimensionality

Consider a K nearest-neighbors approach to property prediction: the value of an unknown molecule is calculated by averaging over the values of nearby, known molecules. Here, a metric such as Euclidean distance, is applied to the p descriptors of points i and j to calculate "closeness":

$$d(i,j) = \left(\sum_{k=1}^{p} (x_k(i) - x_k(j))^2\right)^{1/2} \qquad (2\text{-}1)$$

Assume that N molecules are uniformly distributed in a p-dimensional, unit hypercube, so that the K nearest-neighbors of any point lie in a volume that corresponds to a fraction r of the unit volume, where r is K / N. The edge length of this volume is described by: $e(r) \sim r^{1/p}$. In order to average over 5% of the data in a nearest-neighbors model with ten descriptors, the edge length of the "neighborhood" must be $\sim 0.05^{1/10}$, or 0.74. Thus, over 74% of the entire range for every descriptor must be captured in order to form a local average using only 5% of the data. The meaning of "local average" in this high dimensional space falls apart. Thus, as the number of descriptors in the analysis grows, the curse of dimensionality demands that the amount of data necessary to make reliable predictions increases exponentially (Hastie, 2001).

One solution might be to decrease r dramatically, which corresponds to reducing the value of K in our algorithm. As the sample size becomes smaller, however, uncertainty in the estimated mean increases. The presence or absence of a few points can cause significant deviations from the true average. Because real world data often lacks

adequate coverage in large areas of the descriptor space, predictions in these sparse regions suffer.

2.2 Model bias

Consider an alternative to the *K*-nearest-neighbors algorithm: instead of allowing only a small subset of points in the training set to govern predictions in their local vicinity, a global function is constructed based on all points. This process introduces assumptions about the functional form of the model and how best to fit it to the training data. One popular choice is the linear regression model, which has the form:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \qquad (2\text{-}2)$$

In Equation 2-2, x_j is the jth descriptor variable, β_j is the jth model parameter, and β_0 is the model offset. If the data is truly derived from a linear system that fluctuates with random, Gaussian noise, then the best linear model based on all descriptors is found by minimizing the residual sum of squares (RSS) criteria:

$$\mathrm{RSS} = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 \qquad (2\text{-}3)$$

Thus, the parameters of the model are determined by all points in the training set. In general, linear models are more stable than *K* nearest-neighbors because the absence or presence of individual data points will have less impact on the estimation of the model

parameters. Unfortunately, this stability comes at a price. Even if the original independent variables are transformed via arbitrary functions to achieve greater flexibility as in Equation 2-4, the additive nature of our model prevents interactions *between* descriptors:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} f(x_j)\beta_j, \qquad (2\text{-}4)$$

The number of interactions grows combinatorially as a function of the number of descriptors, so adding extra terms to the model augments the number of dimensions and forces confrontation with the curse of dimensionality. Thus, imposing structure on the data introduces a model bias; predictions suffer if the model assumptions fail to reflect the truth.

2.3 Bias-Variance Tradeoff

To summarize, as the number of descriptors in a model increases, the ability to adequately describe the local environment around known points suffers due to the curse of dimensionality. The resulting prediction function becomes rougher as it attempts to fit the fluctuations inherent in the training set. The model can be made less prone to outliers by imposing a less complex structure on the data; however, such assumptions impose a bias if they do not properly describe reality. These opposing forces constitute the *bias-variance tradeoff*. More formally, suppose a population is described as below (Hand, 2001):

$y = f(\mathbf{x}; \theta) + \varepsilon$, the observed property y is a function of descriptors $\mathbf{x}$ and model parameters θ, plus some experimental random noise ε.

$\mu_y = E[y|\mathbf{x}]$, the true value of y given the descriptors $\mathbf{x}$. The expectation, E[ ], is employed in order to average over the experimental noise.

$\hat{y} = f(\mathbf{x}; \theta')'$, the estimate of the property y using a model fit with parameters θ'

The mean squared error (MSE) of the prediction given $\mathbf{x}$ is:

$$\mathrm{MSE}(\mathbf{x}) = E[\hat{y} - \mu_y]^2 \qquad (2\text{-}5)$$

$$= E[\hat{y} - E(\hat{y})]^2 + E[E(\hat{y}) - \mu_y]^2 \qquad (2\text{-}6)$$

The first term in Equation 2-6 assesses the contribution of variance to the error, and represents how much the prediction fluctuates when the model is fit using different data. Each training set approximates the true parameters of the population, θ, as θ'. Errors accumulate as the number of estimated model parameters increase, yielding greater deviations in $\hat{y}$. The second term describes how far the model prediction deviates from the true value, and results from the innate bias of the function, $f(\mathbf{x}; \theta')'$. Models that are too simple are subject to bias; overly complex models with many parameters suffer from high variance. Therefore, in order to maximize predictive performance, model complexity must be tuned to balance the tension between bias and variance.

2.4 Model Complexity

Low complexity models fail to describe either the training set or the test set. As complexity increases, the training set error can be driven to zero; however, at some point, over-training occurs, and the fitted model generalizes less and less well to external data.

As demonstrated in Figure 2-1, the optimal model lies somewhere in between these two extremes (image adapted from Hastie, 2001).

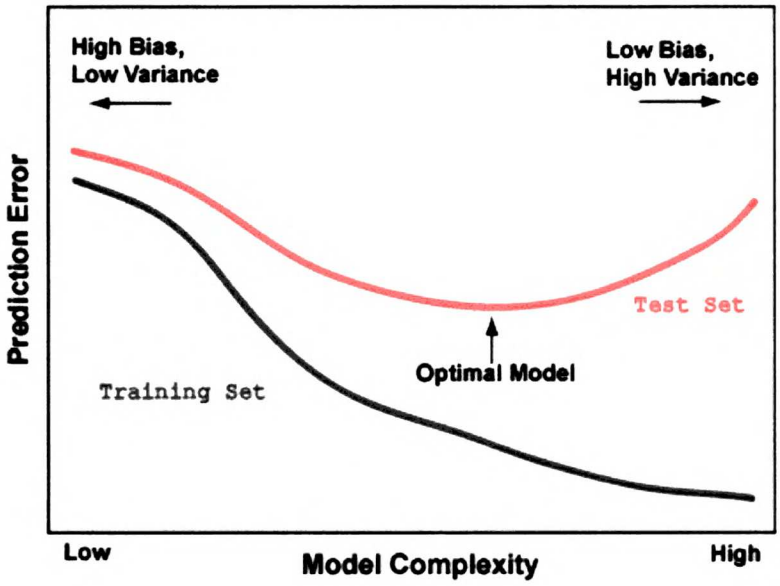


Figure 2-1. Training set and test set error as a function of model complexity

Complexity is a function of the type of model structure employed and the number of descriptors. Most modern statistical learning algorithms are adaptations of either linear regression or K nearest-neighbors that modulate the level of complexity. For example, the nearest-neighbor method can be modified to calculate a local average using a kernel, which weights the contribution of points via a function that goes to zero with increasing distance. Though the influence of local points remains dominant, the prediction function is smoother and has less variance. Flexibility can be introduced into linear models by fitting basis expansions of the original input variables, thereby decreasing model bias.

The algorithms used in the study of promiscuous inhibitors described in Chapter 3 are introduced below, with an emphasis on how complexity is adjusted to balance the variance-bias tradeoff.

2.5 Statistical Learning Algorithms

The promiscuous inhibitor data set, described in the next chapter, contains percent inhibition measurements for 1030 compounds. Criteria for classifying compounds as "active" (promiscuously inhibiting) or "inactive" were based on the distributions of known controls. The goal was to construct a binary classifier for distinguishing promiscuous inhibitors ($\hat{y} = 1$) from normal compounds ($\hat{y} = 0$).

Background

Classifiers generally follow three paradigms (Hand, 2001):

- Class-conditional approaches attempt to model the densities of input variables in each category explicitly, and then use Bayes rule to derive the posterior class probabilities. The predicted class is the class with the highest posterior probability. The naïve Bayes algorithm is an example of this type of classifier.
- Discriminative approaches attempt to model the decision boundary in the descriptor space. Molecules mapped to one side of the boundary are classified to one group, and vice-versa. Some methods attempt to identify linear boundaries using transformed versions of the input vectors, resulting in more complex, non-linear boundaries in the original descriptor space. Examples include support vector machine classifiers and some decision trees.
- Regression approaches attempt to model the posterior class probabilities explicitly; again, class assignment is made based on the maximum probability. Traditionally, such algorithms are applied to model categorical data and yield log-odds predictions for each class. However, models trained on continuous response data can also be included if categories are assigned using threshold

values on the continuous prediction. Regression algorithms following this paradigm include support vector machine regression, some decision trees, general boosting machines, principal components regression, partial least squares, and least angle regression.

Naïve Bayes

The naïve Bayes algorithm was discussed in Chapter 1. A variable subset selection procedure, such as backwards stepwise selection, is one way model complexity can be adjusted. In this method, p+1 models are built (one model with all variables and p additional models lacking one variable). Performance is assessed via cross-validation. The variables are ordered according to how they affect predictive power when absent from the model; a percentage of the lowest ranking covariates is then removed, and the procedure repeats. The optimal subset of the descriptors yields the most parsimonious model performing equally or better than more complex models.

Least Angle Regression (LARS)

LARS (Efron, 2003) is an enhancement to ordinary least squares (OLS). In the first step of OLS, the descriptor vector $\mathbf{x}_j$ is regressed onto the response y to yield its univariate regression coefficient, β_j. The next vector, $\mathbf{x}_{j+1}$, is made orthogonal with respect to all prior vectors, and is then regressed onto the residual of the response to yield the next coefficient, β_{j+1}. The linear regression model is then:

$$\hat{y} = \sum_{j=1}^{p} x_j \beta_j \quad (2\text{-}7)$$

However, this is a greedy process: the regression of the current orthogonalized vector might eliminate the contribution of other covariates that happen to be correlated with the residual of the response (Efron, 2003). With LARS, the descriptor vectors are fractionally added to the regression equation. The first vector is chosen such that is has maximum correlation with y. It is partially regressed onto y until the residual of the response is equally correlated with another covariate. Instead of regressing onto an orthogonal projection of this new variable, LARS uses a vector that is equiangular to all prior descriptors in the model. This bisecting vector is then regressed onto the residual of the response, until the resulting residual is equally correlated with another descriptor, and so on. In this way, the regression equation is constructed by incrementally adding contributions from the independent variables. Furthermore, the equiangular constraint on the regression vectors parallels conjugated gradient techniques for minimization: each "downhill" step taken to reduce the response residual is weighted by previous steps. These modifications afford greater opportunity for weakly correlated variables to contribute to the model.

For p descriptors, the algorithm has a maximum of $M = p$ steps, whereby the OLS solution is returned. The LARS procedure essentially provides all solutions to the Lasso, a least squares variant that imposes an additional constraint such that the sum of the absolute magnitudes of all regression coefficients is less than the L1 norm, the length of the coefficient vector from the OLS solution (Efron, 2003). The upper bound on the sum forces a "competition" among the descriptors, whereby the contributions of poorly correlated covariates are driven to zero. Thus, LARS can be regarded as a smooth version of a variable subset selection procedure. Complexity can be adjusted by selecting

a model derived from an intermediate number of steps (or equivalently, a fraction of the L1 norm) that minimizes test set error as estimated by cross-validation.

Principal Component Regression (PCR)

PCR uses a subset of the principal components (PCs) of the p input descriptors as regression terms in a linear model constructed using RSS minimization. The PCs are the eigenvectors of the descriptor covariance matrix; they describe orthogonal directions of variance in the original descriptor space (Hastie, 2001). The model is built by adding principal components to the regression equation in order of decreasing eigenvalue, until all p eigenvectors are added. Complexity is tuned by identifying the ideal $M \leq p$ number of principal component terms to use in the regression via cross-validation. Using less than p terms in the regression function reduces or eliminates the contributions of some descriptors, thereby lowering variance in the predictions.

Partial Least Squares (PLS)

Like PCR, PLS builds a linear regression model using transformations of the input descriptors. However, these transformations are performed with respect to the matrix of input vectors and the response vector, y, in order to identify directions in the descriptor space that have high variance, *and* high correlation with the dependent variable (Hastie, 2001). The algorithm begins by regressing descriptor vector, $\mathbf{x}_j$, onto y, yielding the univariate regression coefficient φ_j. A derived input, $\mathbf{z}_m$ is constructed as the sum:

$$z_m = \sum_{j=1}^{p} x_j \varphi_j \qquad (2\text{-}8)$$

The vector $\mathbf{z}_m$ is now regressed onto y, yielding another univariate regression coefficient θ_m. Each $\mathbf{x}_j$ is orthogonalized with respect to $\mathbf{z}_m$; the resulting $\mathbf{x}_j$ residuals are used to derive $\mathbf{z}_{m+1}$ as before, and $\mathbf{z}_{m+1}$ is then used to obtain θ_{m+1}. The prediction function after step M is then given by:

$$\hat{y} = \sum_{j=1}^{p} x_j \beta_j(M),$$

$$\text{where } \beta_j(M) = \sum_{l=1}^{M} \varphi_{jl}\theta_l \qquad (2\text{-}9)$$

The process continues until $M = p$ derived inputs are constructed. Using $M < p$ terms in the prediction function reduces the effective number of descriptors as in PCR.

Support Vector Machines (SVM)

SVM classifiers attempt to define a separating boundary between two populations of data where the closest distance to any point in either group is maximized (Hastie, 2001). Geometrically, this can be interpreted as creating a buffer space or *margin* around the decision boundary, countering the effects of over-training by providing "slack." Mathematically, the problem is formulated as:

$$\max C,$$

$$\text{subject to } y_i(\mathbf{x}_i^T \beta + \beta_0) \geq C, \text{ for } i \rightarrow 1..\mathrm{N} \text{ and } y \in \{1,-1\} \qquad (2\text{-}10)$$

In Equation 2-10, $x^T\beta + \beta_0$ describes a linear hyperplane (i.e., a line in two dimensions, a plane in three, etc.) which acts as the decision boundary, and C is the size of the margin. An unknown point i is assigned a class label according to **sign**$(x_i^T\beta + \beta_0)$. Any misclassified point will appear on the wrong side of the decision boundary, and yield a negative value for the term $y_i(x_i^T\beta + \beta_0)$. Points within the margin have $y_i(x_i^T\beta + \beta_0) < C$. In real world applications, the data may never be perfectly separable into two populations, so the equation is modified to tolerate some rogue points:

$$\max C,$$
$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq C(1-\xi_i), \text{ for } i \rightarrow 1\ldots N \text{ and } y \in \{1,-1\}$$
$$\text{where } \zeta_i \geq 0 \text{ and } \sum_{i=1}^{N} \xi i \leq \text{some constant} \qquad (2\text{-}11)$$

The slack variables ζ_i allow for some points to fall within the margin, or even on the wrong side of the boundary if $\zeta_i > 1$. The amount of error is constrained by setting an upper bound on the sum over all ζ. This set of equations and constraints can be formulated as a convex optimization problem solvable via Lagrange multipliers (Hastie, 2001). The solution is:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

subject to the constraints:

$$\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi i)] = 0$$
$$y_i(x_i^T\beta + \beta_0) - (1-\xi i) \geq 0 \qquad (2\text{-}12)$$

Only a subset of the training inputs contribute to the solution for β, because α will be nonzero only when $y_i(x_i^T\beta + \beta_0) - (1 - \zeta_i) = 0$. These points are the *support vectors*, and they completely define the decision boundary (Hastie, 2001).

The math to this point has assumed a linear form of the decision boundary, $x_i^T\beta + \beta_0$. In order to increase model flexibility, the descriptor space can be transformed via a basis set expansion. A linear boundary in this new space can be nonlinear in the original space, thus affording better separating power (Hastie, 2001). A popular choice for the expansion is the radial basis function kernel (RBF):

$$K(x, x_i) = \exp(-\lambda\|x - x_i\|^2), \text{ where } \lambda > 0 \quad (2\text{-}13)$$

This kernel generates a basis centered at each point *i* which falls off exponentially in every direction. The parameter λ can be modified to control the smoothness of the function; for example, a large value of λ will produce peaks centered at the individual points, resulting in a bumpier curve. Again, Equation 2-11 employing an RBF can be solved as a constrained optimization problem, yielding the following solution for the decision boundary:

$$f(x) = \sum_{i=1}^{N} \alpha y K(x, x_i) + \beta_0 \qquad (2\text{-}14)$$

SVM complexity can be modulated by modifying the cost of errors (i.e., the upper bound on the sum over all ζ_i) and the kernel function parameters (λ for the RBF), and limited

the number of descriptors available to the model. Large cost values discourage errors on the margin, leading to rougher decision boundaries with higher variance (Hastie, 2001). As noted above, λ controls the smoothness of the decision boundary.

The SVM can be adapted for regression purposes. In the case of the classifier, points far away from the margin do not contribute to the solution. Similarly, in a regression context, points with residuals less then a user-defined threshold are ignored (Hastie, 2001). Thus, SVM regression minimizes residual error by amplifying the influence of points contributing the greatest deviance according to the error function. As before, complexity is adjusted via variable selection and the cost and λ parameters. Here, the cost relates to the threshold of the error function, as opposed to the size of the margin.

Decision Trees

A decision tree is formed by recursively partitioning the training data using binary splits on single variables in order to maximize a measure of class purity in the resulting two nodes (Hand, 2001). Trees continue growing until the purity of the resulting nodes fails to increase, or some pre-defined threshold for minimum node size or maximum depth are reached. Predictions on external data are made by beginning at the root node and sequentially traveling down the tree, choosing the appropriate branch at each junction. The unknown molecule is assigned the majority class (or weighted majority) of the terminal node or "leaf".

In a classification context, RSS error is not an appropriate metric for measuring the utility of a binary split. Instead, the following measures are employed to quantify class purity at node m with K classes (Hastie, 2001):

- Misclassification error: $1 - \rho_m k(m)$, the probability of not belonging to the class assigned at node m
- Gini Index: $\sum_{k=1}^{K} \rho_{mk}(1 - \rho_{mk})$, the probability of belonging to class k at node m multiplied by the misclassification probability at that node
- Cross-entropy or deviance: $\sum_{k=1}^{K} \rho_{mk} \log \rho_{mk}$

The node splitting process at each branch point follows a greedy algorithm, because the selected variable and threshold are chosen only to maximize class purity in the immediate child nodes. As a result, single trees are often prone to over-fitting. In order to decrease variance, the predicted response is calculated by aggregating over multiple trees in a process called *bagging* (Hastie, 2001). These additional trees are trained on bootstrap or otherwise perturbed versions of the original data (i.e., subsets of the original descriptors and/or data points), yielding different models than the original tree. The average prediction from multiple bagged trees is more stable than the estimate from an individual tree. Another method to generate additional trees called *boosting* is described in the *general boosting machines* section below.

The complexity of decision trees can be tuned by modifying the number of trees and the maximum numbers of nodes per tree.

Generalized Boosting Machines (GBM)

Earlier, Equations 2-3 and 2-4 described a general approach to constructing a linear model. It was noted that transformations of the original descriptors afforded greater flexibility; for example, PCR uses principal components. However, such methods are limited to a finite number of functions that are fit simultaneously. In contrast, LARS

and Decision Trees seek new functions to add to the model based on the negative gradient of the loss function, and the addition of new terms does not affect previously determined parameters. GBMs can be viewed as a generalization of these two algorithms: they build models by adding terms one-by-one from a potentially infinite set of functions derived from the original descriptor set. More formally, a GBM can be described as a function, $f(x)$, such that (Hastie, 2001):

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \lambda_m) \qquad (2\text{-}15)$$

In Equation 2-15, $b(x; \lambda_m)$ describes any additive expansion of the original descriptors (such as a neural network, a regression spline, a decision tree, etc.), β_m acts as a weighting factor, and M describes the current iteration of the GBM. At step m of model construction, the current term, $\beta_m b(x; \lambda_m)$, is adaptively chosen to maximize overlap with the negative gradient of the loss function.

Any loss function can be driven to zero by following the gradient and adding enough terms. Unfortunately, this result almost certainly leads to an over-trained model. Two strategies are used to prevent this scenario: (a) adding limited basis functions that are unable to perfectly match the negative gradient, (b) shrinking the contribution of terms to the model (Ridgeway, 2005). By using constrained basis functions, the GBM only approximates the negative gradient, and therefore moves in the right direction with imprecision or "slack". Decision trees are a popular choice for these functions, as the node structure imposes limits on how well the tree can match the gradient. Down-weighting the contribution of each term to the model slows the rate of learning. Although

the GBM will require more terms to achieve the same reduction in the loss function, the resulting model is more likely to have explored the effects of weakly predictive descriptors.

By applying these two criteria, the GBM can be viewed as a collection of weak classifiers (Hastie, 2001). By virtue of following the gradient, the GBM emphasizes the most poorly fitting data points in subsequent rounds of model building. This process is known as *boosting*, and has been shown to dramatically improve prediction accuracy (Hastie, 2001). Moreover, GBMs can utilize different loss functions that are optimized for certain prediction problems. For example, exponential and Bernoulli loss functions are ideal for classification problems, with the latter providing more robust solutions (Hastie, 2001). Gaussian and Laplace loss functions are more appropriate for regression problems (Hastie, 2001).

The GBM complexity is modulated via the constraint places on the basis functions (e.g., the maximum depth of a decision tree), the number of terms in the model, and the shrinking factor. Practical experience suggests using the smallest possible shrinking factor given computational and time constraints (Ridgeway, 2005). The ideal number of terms and the type of basis functions employed are then chosen via cross-validation.

2.6 Selecting the Best Algorithm

Model performance is a function of the structure and quality of the underlying data. However, issues besides prediction accuracy, such as the ability to handle multiple data types and scalability, are also important to consider when selecting a computational

learning method. Table 2-1 summarizes the relative benefits and limitations of the algorithms detailed in this chapter.

Algorithm	Naïve Bayes	LARS	PCR	PLS	SVM	Decision Trees	GBM
Ease of handling mixed data types	●	●	○	○	○	●	●
Scalability (for large N)	●	○	○	○	●	●	●
Computational burden	○	●	●	●	●	○	○
Flexibility of the model structure	●	●	●	●	●	○	○
Predictive performance	●	○	●	●	●	○	●
Interpretability	●	●	○	●	●	●	○

Table 2-1. Relative benefits and limitations of select statistical learning methods (● = poor, ● = fair, ● = best)

In general, the Naïve Bayes, Decision Tree, and GBM algorithms handle multiple data types (categorical, ordinal, interval, and ratio) without requiring significant pre-processing. The other methods require binarization of categorical variables, as well as scaling, centering, and additional normalization. The SVM scales poorly with the number of data points due to the demands of the kernel function, whereas the Naïve Bayes, Decision Tree, and GBM methods scale linearly. Computational burden refers to the time spent optimizing the complexity parameters via cross-validation. Only a single parameter needs to be investigated for LARS or PCR and PLS (either the fraction of the L1 norm or the number of components, respectively). The naïve Bayes and SVM methods often employ an external variable selection method that requires parameter

optimization at each step. Moreover, the cost and gamma values of the SVM are usually explored using an expensive grid search. On the other hand, the SVM offers the most flexible model structure via the kernel function. The basis functions employed by Decision Trees and GBM also allow for complex interactions between variables, whereas the remaining algorithms only afford additive relationships. The SVM's ability to explore nonlinear spaces translates to better predictive performance, though at the expense of model interpretability. Indeed, naïve Bayes, Decision Trees, and LARS offer the best interpretability because they do not transform the descriptors. The GBM also uses the original variables, but is less comprehensible because it aggregates hundreds or thousands of basis functions.

Thus, the selection of the best computational learning algorithm depends on the nature of the data, time and resource constraints, and the goals of the project.

2.7 Consensus Models

One alternative to focusing on a single, "best" algorithm is to amalgamate the results from a collection of satisfactory models derived by different methods. This process is a logical extension of the bagging procedure described earlier. Each method imposes a structure that might be particularly adept at learning a portion of the data; however, the bias and variance inherent in every model also introduce errors. By combining different predictions, a consensus model attempts to reinforce the "good" while averaging out the "bad." A critical assumption is that the original models performed well: combining the results from poor estimators will likely increase error. Indeed, O'Brien and de Groot recently reported improved specificity and selectivity for

predicting hERG channel blocking and Cytochrome P450 2D6 inhibition by aggregating results from a neural network and a naïve Bayes classifier (O'Brien, 2005). Thus, consensus models benefit from the advantages of multiple learning algorithms, and therefore, might enjoy significant improvements in predictive performance.

2.8 Conclusion

In this chapter, we covered some of the theory behind data mining in high dimension descriptor spaces. A brief description of popular statistical learning algorithms was presented. The key feature of these methods is the ability to finely tune model complexity. In essence, identifying the best predictive model entails traveling down the training error curve presented in Figure 2-1, and determining an optimal point to stop based on an estimate of test set error. Unfortunately, trial-and-error is often the only way to determine which algorithm to use. In the next chapter, we explore an application of these statistical learning algorithms to a real world problem: promiscuous inhibition in high-throughput screening assays.

Chapter III. Data Mining in Higher Dimensions: Computational Models to Predict Promiscuous Inhibition

Modified from Feng B, Shelat A, Doman T, Guy RK, Shoichet BK. High-throughput assays for promiscuous inhibitors. *Nat Chem Bio*. 2005;1(3):146.

In this chapter, we describe methods to detect promiscuous inhibitors—compounds that nonspecifically disrupt the action of biological macromolecules by putatively forming large aggregates in solution. Experimental high-throughput enzyme and Dynamic Light Scattering (HT DLS) assays and preliminary computational models to identify such molecules were reported earlier (Feng, 2005). We extend and revise that work by first showing how HT DLS correlates poorly with the enzyme-based screen. We conclude by describing the development of a consensus computational model that efficiently classifies promiscuous inhibitors with good specificity and selectivity. Interpretation of the model offers insight into the nature of this intriguing phenomenon.

3.1 Background on Promiscuous Inhibition

High-throughput screening is commonly-used to discover drug leads in industrial settings, and is increasingly penetrating academic and non-profit institutions as well. However, many hits identified using HTS often show up in multiple, unrelated screens and are therefore not useful in subsequent development. These "frequent hitters" or "promiscuous inhibitors" usually display noncompetitive activity and flat structure-activity relationships.

A number of criteria have been proposed to identify such molecules, including the presence of reactive functionality (Rishton, 1997), physical properties that disrupt HTS detection methods (Roche, 2002), and "privileged" scaffolds (Roche, 2002). Recently, the Shoichet group provided strong evidence for a mechanism based on compound

aggregation. Dynamic Light Scattering (DLS), an experimental technique used to measure light scattering in solution, detected aggregates in the solutions of 15 compounds that were reported as inhibitors of one or more proteins or nucleic acids and of a diverse panel of model enzymes, (McGovern, 2002). Subsequent work using transmission electron microscopy (TEM) provided visual evidence of the physical association of enzyme with these structures, suggesting that the proteins were sequestered and unable to catalyze reactions (McGovern, 2003a). The addition of 0.01% Triton X-100 removed promiscuous inhibition, consistent with an adsorption hypothesis (McGovern, 2003a).

This phenomenon has been widely reported in other compound collections, such as the in-house library of a pharmaceutical company, commercially available kinase inhibitors, and known drugs (McGovern, 2003b; Seidler, 2003). Thus, the rapid detection of promiscuous inhibitors would be considerably useful to the screening community.

3.2 HTS and initial computational methods for identifying promiscuous inhibitors

Feng et al proposed two different HTS screens for identifying promiscuous compounds: (a) directly measuring nonspecific inhibition using a model β-lactamase system, (b) detecting aggregation using HT-DLS (Feng, 2005). In the β-lactamase, inhibition was measured in the presence and absence of 0.01% Triton-1000; the loss of activity upon the addition of a small amount of detergent is a hallmark of promiscuous inhibition. Known active and inactive compounds identified earlier via biochemical and physical methods validated the HTS assay. Based on the distribution of activities from controls, compounds inhibiting β-lactamase with >23.8% were deemed promiscuous,

compounds with activity < 10.9% were considered inactive, and compounds with activities between these bounds were classified as ambiguous. The HTS was further verified by applying the method to a set of 1030 unknown compounds (described below), and confirming the activities of a subset of predicted actives and inactives via low-throughput enzyme assays.

It appeared logical to pursue high-throughput DLS, which measures the amount of light scattering by particles in solution, due to the strong evidence linking promiscuous compounds and aggregation. Forty-nine known aggregating, promiscuous inhibitors and known non-aggregating, inactive molecules were used to calibrate the HT DLS-based classifier. Light scattering intensity was converted to a probability of aggregation as follows:

The probability of being an aggregator given a scattering intensity $\mathbf{x}$, $p(c_{AGG} \mid \mathbf{x})$, was estimated using Bayes' Theorem:

$$p(c_{AGG} \mid \mathrm{x}) = p(AGG)p(\mathrm{x} \mid c_{AGG}) / ((p(AGG)p(\mathrm{x} \mid c_{AGG}) + p(NAGG)p(\mathrm{x} \mid c_{NAGG})) \quad (3\text{-}1)$$

The prior probabilities, $p(AGG)$ and $p(NAGG)$, were set to 0.5 assuming a flat prior distribution. The class conditional probabilities $p(\mathbf{x} \mid c_{AGG})$ and $p(\mathbf{x} \mid c_{NAGG})$ were derived from the densities of the known aggregators and non-aggregators as a function of HT DLS intensity. Examination of the raw and log-transformed scattering intensity histograms for both populations revealed a significant departure from normality; thus, modeling the two probability distributions as normal was not appropriate. Instead, the distributions were estimated for the range of the log-transformed intensity values using

the *density* function in the R statistics package (v. 2.0.1, http://www.r-project.org/) with a Gaussian kernel and bandwidth="nrd0". For smoothing purposes, the 'adjust' parameter was set to 1 for the aggregator density and 2 for the non-aggregator density. The probability for a given distribution was set to one past the mean of that distribution to avoid edge effects.

The distributions for $p(c_{AGG} \mid \mathbf{x})$ and $p(c_{NAGG} \mid \mathbf{x})$ (calculated as $1 - p(c_{AGG} \mid \mathbf{x})$) as a function of log-transformed light intensity based on the initial training set of 49 molecules are shown in Figure 3-1.

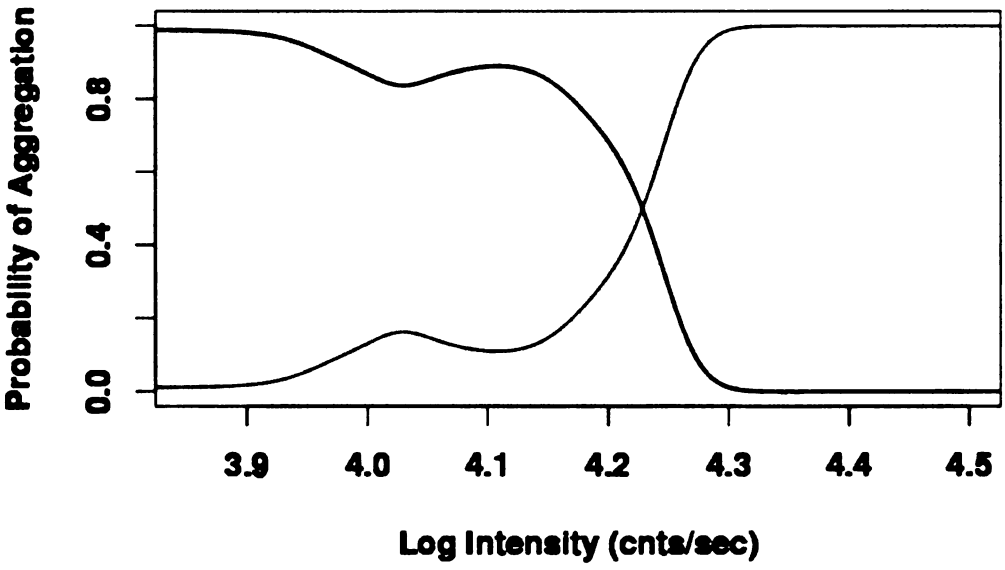


Figure 3-1. First generation of the HT-DLS Classifier. Distributions for $p(c_{AGG} \mid x)$ (red) and $p(c_{NAGG} \mid x)$ (blue) as a function of log-transformed light intensity. Molecules with $0.1 <= p(c_{AGG} \mid x) <= 0.9$ were assigned as ambiguous. Under these criteria, molecules with intensities < 10159 (cnts/s) were classified as non-aggregators; molecules with intensities > 63609 (cnts/s) were classified as aggregators; all other molecules were classified as ambiguous.

Unfortunately, this version of the HT DLS classifier failed to accurately distinguish between aggregators and non-aggregators from external data. Therefore, an additional 58 known molecules were added to the training set. The refined HT DLS classifier correlated better with measurements from low throughput experiments. The

distributions of $p(c_{AGG} \mid \mathbf{x})$ and $p(c_{NAGG} \mid \mathbf{x})$ for the updated model are presented in Figure 3-2.

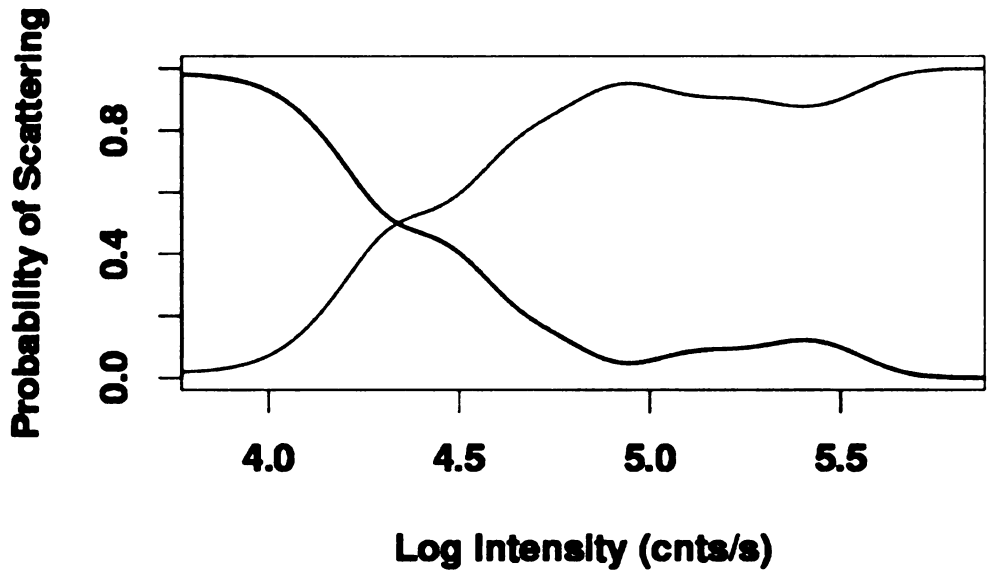


Figure 3-2. Refined HT-DLS Classifier. Refined distributions for $p(c_{AGG} \mid \mathbf{x})$ (red) and $p(c_{NAGG} \mid \mathbf{x})$ (blue) as a function of log-transformed light intensity. Molecules with $0.1 <= p(c_{AGG} \mid \mathbf{x}) <= 0.9$ were assigned as ambiguous. Using these criteria, molecules with light scattering intensity < 10991 (cnts/s) were classified as non-aggregators, and molecules with light scattering intensity > 310934 (cnts/s) were classified as aggregators.

Although the HT DLS classifier was trained to distinguish aggregators from non-aggregators, it was believed that promiscuous inhibitors could also be identified by proxy. To test this hypothesis, the HT DLS was applied to the 1030 "drug-like" molecules used to validate the HTS enzyme assay. This set was composed of 298 randomly chosen molecules and 732 molecules selected by two preliminary computational models that were trained on a total of 110 known promiscuous inhibitors and inactive compounds reported in the literature. Details of the compounds selection strategy are given in the section below.

Compound Selection Strategy

Compounds were purchased from Chemical Diversity, Inc. "Prediction" set molecules were classified as either promiscuous inhibitors or inactive by one of two models: a previously described recursive partitioning model (RP) or a naïve Bayesian model (NB, see below for details). The Prediction set contained 493 predicted inhibitors (200 Bayesian/298 RP) and 239 predicted inactive molecules (97 Bayesian/ 142 RP). The "Random" set contained 298 molecules. All compounds were prepared as 10 mM stocks in neat DMSO.

All purchased molecules satisfied the following Lipinski criteria: (Nitrogen count + Oxygen count) <= 10, molecular weight <= 500, number of H-bond donors <= 5, and AlogP <= 5.6 (Lipinski, 2001). An upper bound of 5.6 was more appropriate for the AlogP-based estimation of logP (Ghose, 1999). Common physical property distributions for the Prediction Set, the Random Set, and the Comprehensive Medicinal Chemistry (CMC) database (v. 2004, Elsevier MDL) were compared to further ensure that the test sets were reasonable representations of drug-like molecules. The CMC was filtered according to Ghose et al to remove compounds that were unlikely to be orally bioavailable, such as contrast agents, solvents, and pharmaceutical aids (denoted CMC*) (Ghose, 1999). As shown in Table 3-1, the interquartile ranges of chemical properties for both the Prediction and Random sets are similar to those of the CMC*.

Property	CMC* (N=7790)	Prediction Sets (N=732)	Random Set (N=298)
AlogP	1.13 - 3.96	2.62 - 4.35	2.12 – 4.02
Molecular Solubility (log ug/ml)	-5.72 – -2.82	-6.22 – -4.39	-5.88 – -3.76
Molecular Weight (Daltons)	261 – 411	303 – 400	309 – 411
Polar Surface Area ($Å_2$)	42 – 105	59 – 100	67 – 102
Total Surface Area ($Å_2$)	255 – 395	282 – 365	289 – 383
# H Donors	1 – 2	1 – 2	1 – 2
# H Acceptors	3 – 6	3 – 5	3 – 5
# N + O Atoms	3 – 7	3 – 6	4 – 6
# Rotatable Bonds	3 – 7	4 – 6	4 – 7

Table 3.1. Interquartile Ranges for common physical properties from the CMC*, and Prediction and Random Sets selected from Chemical Diversity, Inc.

Recursive Partition (RP) Model

The RP model has been described elsewhere (Seidler, 2003)

Naïve Bayesian Model

The initial Bayesian model employed the *naïve Bayesian classifier* component in Pipeline Pilot 4.5.0 (Scitegic, Inc). The algorithm, described in Chapter 1, calculates the sum of log-odds for the occurrence of a given feature in a set of molecules belonging to one of two classes ("promiscuous inhibitor" and "inactive" in this case).

The descriptors used in the initial model included bits from Scitegic's ECFP_6 molecular fingerprint and the first five principal components (PCs) calculated from a principal component analysis (PCA) of all 1D and 2D descriptors from MOE (v. 2002.3, Chemical Computing Group). These five PCs accounted for 92.5% of the variance in the MOE descriptor space.

The initial data set contained 110 known inhibitors and inactives (McGovern, 2002; McGovern, 2003a and b; Siedler, 2003). Ten percent of the data was withheld from the model building process (the "validation" set). Using the remaining data, models were trained on 80%, and scored on the other 20% using the Receiver-Operator Characteristic (ROC, calculated in Pipeline Pilot). One hundred models were generated using different partitions of the data and the top ten performing models were selected. The consensus model score, or CSCORE, was the mean of the 10 models.

The probability of an unknown compound promiscuously inhibiting was calculated using a method similar to that of the HT DLS classifier (see above). A flat prior distribution was assumed. The probability mass function was modeled as two normal distributions estimated using the CSCORE distributions from known actives and inactives in the validation set. Molecules with a posterior probability for nonspecific inhibition >= 0.6 were classified as active; molecules with a probability <= 0.4 were classified as inactive, and no classification was made for molecules in the probability range 0.4 – 0.6.

To assess the utility of the initial Bayesian model, a test set of predicted promiscuous inhibitors and predicted inactive molecules were selected from Chemical Diversity, Inc's compound library. In order to minimize uncertainty in the predictions, only putative active compounds with posterior probabilities > 0.99 were selected; likewise, the selected inactives had probabilities < 0.01. These constraints yielded 968 predicted inhibitors and 80,778 predicted inactives. Within each set, a maximum dissimilarity metric (implemented in the *Diverse Molecules* component in Pipeline Pilot)

was employed to select the most diverse set of 300 inhibitors and 300 inactives, of which 200 predicted inhibitors and 97 predicted inactives were purchased.

Accuracy of the DLS and Computational Models

For comparison, the RP and NB models were also applied to the Random set. Remarkably, both computational models outperformed the HT DLS classifier with respect to the misclassification rate (Table 3-2).

Model	Active Precision	Active Recall	Inactive Precision	Inactive Recall	Mis-classification Rate	Not Classified
Recursive Partioning	43% (44/103)	77% (44/57)	92% (160/173)	73% (160/219)	26% (72/276)	0%
Naïve Bayes	33% (13/39)	23% (13/57)	81% (180/222)	82% (180/219)	26% (68/261)	5% (15/276)
HT DLS Classifier	40% (37/93)	65% (37/57)	97% (33/34)	15% (33/219)	45% (103/200)	54% (149/276)

Table 3-2. Results from the HT DLS Classifier and initial computational models applied to the Random Set (57 Aggregators, 219 Non-aggregators; 22 compounds were ambiguous and removed from this study). Misclassification rate is defined as: total number of incorrect predictions / total number of prediction.

A number of reasons could have accounted for the unreliability of the HT DLS classifier. First, DLS in a high-throughput setting might be prone to significant noise. As noted in the rightmost column in Table 3-2, HT DLS failed to provide a classification for 54% of the Random set. This deficiency stems from the fact that many compounds fell within the ambiguous region of the probability distributions. Perhaps a larger training set would provide better resolution. To examine whether there was any detectable correlation between HT DLS measurements and promiscuous inhibition, a simple linear regression model was used to fit log(DLS intensity) as a function of %inhibition. A

scatter plot of the data and the least squares fit (red) are reported below (R^2=0.2441, slope = 0.014 (0.001), $p < 10^{-10}$):

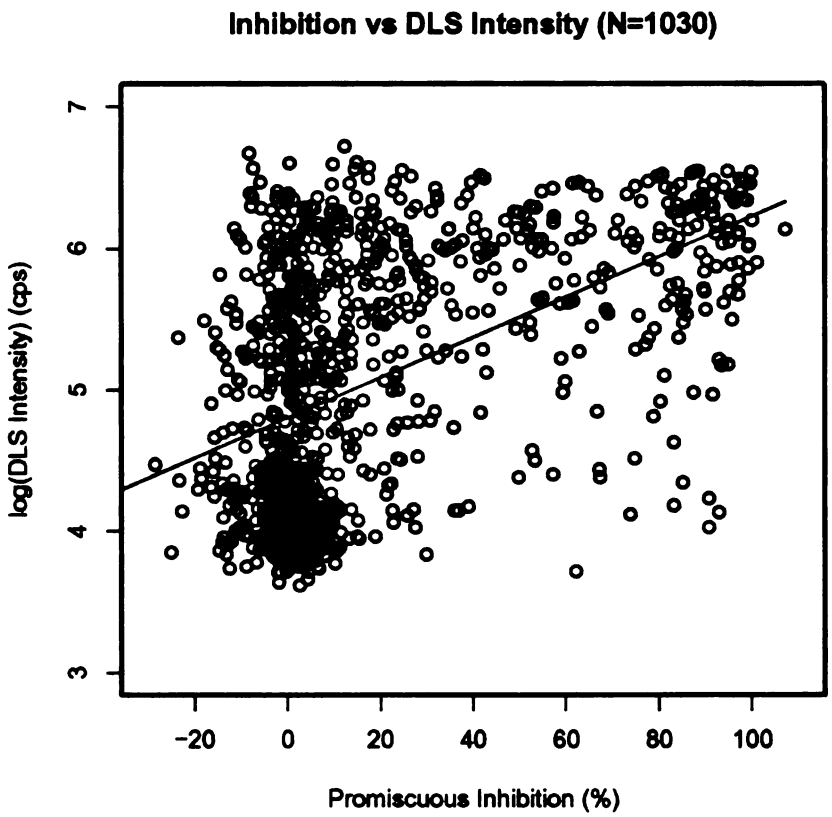


Figure 3-3. Scatter plot and least squares fit (red) of Promiscuous Inhibition vs. log (DLS Intensity) for all HTS data.

Although HT DLS failed to account for roughly 75% of the variance in promiscuous inhibition, a slight correlation did exist. However, it was noted that 732 of the 1030 molecules were chosen based on computational models trained on promiscuous inhibitors that were also aggregators and inactive molecules that were shown not to form particles. The learning algorithms could have selected for compounds that aggregated in addition to, or as a proxy for, promiscuous inhibitors. The Random set, however, did not

contain this bias. A least squares fit for the 298 randomly chosen molecules performed even more poorly (R^2= 0.1131, slope = 0.011 (0.002), $p < 10^{-8}$) (Figure 3-4):

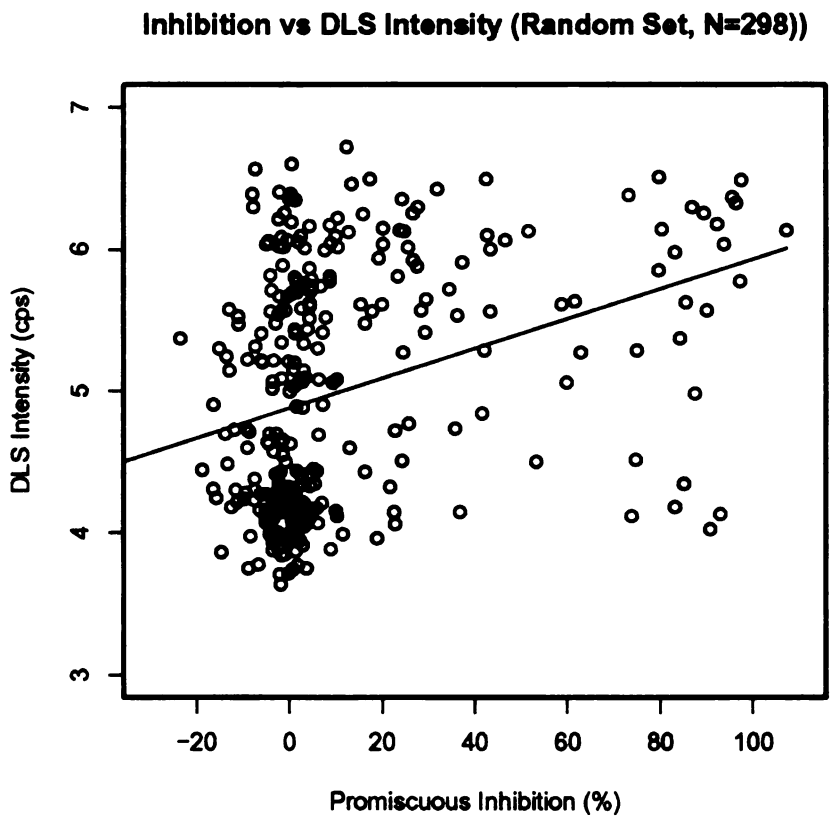


Figure 3-4. Scatter plot and least squares fit (red) of Promiscuous Inhibition vs. log (DLS Intensity) for HTS data from the Random set.

One explanation for the lack of correlation in Figures 3-3 and 3-4 was the occurrence of precipitates among inactive compounds. The high scattering intensities observed for these molecules might have been due to precipitation, a behavior that was distinguishable from aggregation (Feng, 2005). However, this rationale did not account for the high proportion of promiscuous inhibitors with low scattering intensities. On the other hand, promiscuous compounds could have optical properties that perturb HT DLS measurements. For example, the nonspecific inhibitor Congo Red can only be studied at

DLS concentrations much higher than its IC_{50} for β-lactamase, the model protein used in the enzyme-based assay (Feng, 2005). Alternatively, this phenomenon could have arisen from aggregates only after particle size has exceeded some threshold; such dependencies were not detectable via HT DLS intensities, which only reflected the number of scattering particles. Finally, it was conceivable that promiscuous inhibition and aggregation were distinct properties, or that multiple mechanisms existed by which molecules promiscuously inhibit.

Intrigued by the possibility of gaining insight into the physical basis behind promiscuous inhibition and by the desire to develop robust tools for virtual detection, we pursued refinements to the computational models. The training set was enlarged to include the Prediction set in addition to the 110 compounds identified from the literature while the Random set was withheld for validation.

3.3 Revised Models Predicting Promiscuous Inhibition

Prior Art

Roche et al completed the first computational assessment of "frequent hitters" (Roche, 2002). In their study, the definition of "promiscuous" encompassed molecules interacting via nonspecific mechanisms and compounds with physical properties that hampered detection during HTS. First, the authors selected compounds that repeatedly showed up as hits in HTS assays, were frequently requested from in-house libraries, or were specially submitted by medicinal chemistry groups. A group of expert medicinal chemists then examined the list to identify molecules that were likely to be frequent hitters and not compounds that were wrongly annotated, chemically degraded, or

significantly impure. Nine of the eleven chemists needed to concur in order to assign a compound as promiscuous. A selection of commercially available drugs was chosen as the control group. The resulting computational model performed well, correctly classifying 90% of all frequent hitters and 91% of all controls.

In the current study, the data set was limited to compounds known to inhibit promiscuously based on the model enzyme HTS, and did not preclude commercially available drugs as in the Roche study. This distinction was important because known therapeutics are often included in HTS experiments as controls or for profiling purposes. Any compound that inhibits promiscuously, whether it be a drug or an undistinguished molecule, can disrupt biological screening assays. Furthermore, though both data sets were "drug-like" according to such criteria as Lipinski rules, significant differences in composition existed. Thus, the promiscuous inhibitor model was likely to behave differently than the Roche model.

Refinement of the RP model (Random Forest)

A random forest model was constructed as described elsewhere (Feng, 2005)

Refinement of naïve Bayesian model (rNB):

Chemical fingerprint bits were derived as before, except now using Scitegic's FCFP_6 fingerprint. Physico-chemical descriptors were calculated using molconnZ (EduSoft LC, version 4.09). Initially, the training set contained 765 descriptors. Using the R statistics package, correlated descriptors or descriptors with zero variance were removed, yielding 540 covariates.

The initial Bayesian model used bits present in Scitegic's ECFP_6 fingerprint and the first five principal components derived from PCA of MOE descriptors. Reducing the dimensions of the MOE descriptor space effectively captured the variance of the data in fewer variables; however, PCA confounded interpretation of the contribution of individual descriptors. In the refined model, PCs were avoided; instead, a subset of the original variables was selected using the backward stepwise method (described in Chapter 2). The ideal set was determined by maximizing the Z-factor as a function of the number of descriptors. Models were trained on 80% of the data using all descriptors, then evaluated using the remaining 20% via the Z-factor metric (Zhang, 1999). Each descriptor was then individually removed, and the model was rebuilt and reassessed. The entire procedure was repeated 10 times, and then descriptors were rank ordered according to their average contribution to the Z-factor. The lowest performing one-half of all descriptors was removed, and the process was repeated until an optimal set was determined. The final model consisted of 43 descriptors. An ensemble of 25 models was generated using the 43 descriptors, and the probability of aggregation was determined using the mean of all model scores as described previously.

Results from the Refined Models

The results from the refined models are presented in Table 3-3:

Model	Active Precision	Active Recall	Inactive Precision	Inactive Recall	Mis-classification Rate	Not Classified
Random Forest	83% (34/41)	60% (34/57)	90% (212/235)	97% (212/219)	11% (30/276)	0%
rNB	50% (42/84)	74% (42/57)	95% (158/167)	73% (158/219)	20% (51/251)	9% (25/276)

Table 3-3. Results from the revised computational models applied to the Random Set (57 Aggregators, 219 Non-aggregators; 22 compounds were ambiguous and removed from this study).

The Random Forest was the most accurate model as reflected by the misclassification rate. However, it fails to identify 40% of the promiscuous inhibitors. The Refined Bayes correctly identifies a larger number of the inhibitors, albeit with more false positives. Thus, both models were unsuitable for virtual screening, necessitating additional refinement.

The contrasting performance of the models underscored the tradeoff between identifying inhibitors (active recall) and incorrectly classifying compounds (misclassification rate). Because the data set was imbalanced by a smaller number of promiscuous compounds, it was more difficult to correctly identify true positives than true negatives. To account for this characteristic, further optimizations and comparisons employed Cohen's Kappa in place of the misclassification rate (Equation 3-2):

$$\kappa = (\rho_o - \rho_e) / (1.0 - \rho_e), \qquad (3\text{-}2)$$

where ρ_o is the observed proportion of agreement (the classification rate) and ρ_e is the expected agreement from chance alone. Kappa expresses the agreement between a prediction and the truth as a proportion of the maximum possible agreement not due to chance.

3.4 Final Round of Modeling Promiscuous Inhibition

In the final round of modeling, we explored a larger descriptor space and a different set of statistical learning algorithms (see Chapter 2 for a theoretical background). The methods are described below.

Software

Unless otherwise stated, the following versions of software were used: Pipeline Pilot (v. 4.5.2, Scitegic, Inc.), the R program (v. 2.1.1), MOE (v. 2004.3, Chemical Computing Group), Volsurf (v. 4.1.3, Molecular Discovery Ltd.), Grid (v. 22a, Molecular Discovery Ltd.), and MolconnZ (v. 4.09, Edusoft, LC). Default parameters were used unless explicitly noted.

Data Preparation

For consistency, the data set for this new generation of models was limited to compounds screened in the HTS. The 732 molecules selected via computational methods were used as the training set, while the remaining 298 were withheld for validation. All molecules were standardized using a Pipeline Pilot protocol that removes all hydrogen atoms and salts, chooses a canonical tautomer, sets all atoms to formal charge, and selects a single topology for certain functional groups (such as nitro and carboxylate). Each molecule was then minimized into a low energy three-dimensional conformation and ionized to physiological pH in MOE. A total of 864 descriptors were computed, including all 2D descriptors in MOE, all Volsurf descriptors, a selection of the most relevant MolconnZ descriptors, and physical properties and binary descriptors

representing structural features in Pipeline Pilot. The binary features were derived from Pipeline Pilot's FCFP_6 fingerprint; to limit their number, only structural features present in ≥ 5% of all molecules, but ≤ 95% of all molecules were allowed. The data set was then filtered to remove descriptors with zero variance or with perfect correlation to other descriptors using the R program, resulting in a final set of 627 descriptors. Next, all variables in the data were centered to zero mean and scaled to unit standard deviation using the R *scale* function. The response variable was either the experimentally measured promiscuous inhibition at 30uM or a class label assigned using inhibition thresholds described elsewhere (Feng, 2005). Ambiguous molecules were removed from the training data when employing classification algorithms. For regression algorithms, no molecules were discarded from training; molecules with predicted activities greater than 23.8% were classified as predicted inhibitors and all others were predicted as inactive. Ambiguous molecules were removed when computing the test set kappa for all models.

Benchmark experiments to assess the scalability of the computational models were programmed within the Pipeline Pilot environment using a combination of the native Pilot Script language and external calls to R program scripts. All calculations were performed using a single 3.4 GHz CPU running Redhat Linux AS3 with 1 GB RAM.

Computational algorithms

The PCR (method="svdpc", numcomp="200") and PLS models (method="kernel", numcomp="200") were constructed and assessed using the *mvr* and

predict.mvr functions in the R pls package (v. 1.0-3; Wehrens and Mevik; 2005). The ideal number of components (numcomp) was estimated via cross-validation (see below).

The LARS (method="lar") model was constructed and assessed using the *lars* and *predict.lars* functions in the R lars package (v. 0.9-5; Hastie and Efron; 2004). The algorithms required all descriptors to be scaled to unit length. The optimal fraction of the L1 norm of the coefficient vector was estimated via cross-validation. This number ranges from 0 to 1, with 1 yielding the OLS solution (see Chapter 2).

The naïve Bayes model was constructed and assessed using the *naiveBayes* and *predict.naiveBayes* functions in the R e1071 package (v. 1.5-8; Dimitriadou, Hornik, Leisch, Meyer, and Weingessel; 2005). The class conditional probabilities were modeled as Gaussian distributions. An optimal number of descriptors were determined using the backward stepwise selection method described in Chapter 2. Initially, 90% of the variables were removed at each step of the procedure; selection occurred on individual descriptors once the range was narrowed to a workable space.

The SVM classification (scale="FALSE", type="C-classification", kernel="radial", cachesize="500") and regression models (scale="FALSE", type="eps-regression", kernel="radial", cachesize="500") were constructed and assessed using the *svm* and *predict.svm* functions in the R e1071 package (v. 1.5-8; Dimitriadou, Hornik, Leisch, Meyer, and Weingessel; 2005). Cross-validation was used to identify the ideal subset of descriptors using the backward stepwise selection method. For each subset of descriptors, the cost and gamma parameters were explored via a grid search. Initially, 50% of the variables were removed at each step of the procedure, and the grid search was coarse (cost ranged from 2^1, 2^2 .. 2^5; gamma ranged from 2^{-3}, 2^{-5} .. 2^{-11}). The granularity

of the exploration increased as the bounds of the optimal range for the SVM parameters decreased.

The GBM classification (distribution="bernoulli", n.trees="100000", interaction.depth="3", n.minobsinnode="5", shrinkage="0.001", bag.fraction="0.5", train.fraction="1.0") and regression models (distribution="gaussian", n.trees="50000", interaction.depth="3", n.minobsinnode="5", shrinkage="0.001", bag.fraction="0.5", train.fraction="1.0") were constructed and assessed using the *gbm* and *predict.gbm* functions in the R package gbm (v. 1.5-1; Ridgeway; 2005). Cross-validation was used to identify the optimal number of trees to use for predictions.

The Bagging (split method="Gini index", minimum samples in node="2", enrichment threshold="0.5", good bias="1", preserve minority="TRUE") and Boosting Decision Trees (split method="Gini index", minimum samples in node="2", enrichment threshold="0.5", good bias="1") were constructed and assessed using the Recursive Partitioning with Decision Trees components in Pipeline Pilot. The ideal number of trees and the tree depth were determined via cross-validation.

Cross-validation

As noted above, adjustable model parameters governing complexity were optimized using cross-validation. Models were trained on a random partition containing 85% of the 732 molecules (640 molecules when ambiguous compounds removed for classifier algorithms); test set error was estimated by applying the model to the remaining 15%. To avoid the influence of structure in the training set, the cross-validation was repeated between 10-200 times depending on the computational burden of the

calculation. For example, the naïve Bayes algorithm was repeated 200 times for each subset of variables. The SVM classification grid search used 10 repetitions in the early stages of variable selection, and 200 when identifying the ideal cost and gamma parameters after the optimal number of variables was identified. For each algorithm, the ideal set of parameters was determined by identifying the most parsimonious model with estimated test set error (kappa or mse) within one standard error of the mean of the best performing model.

Figures 3-5 – 3-7 below show graphs of test set error as a function of model complexity as assessed via cross-validation studies for the GBM Regression and the SVM classifier.

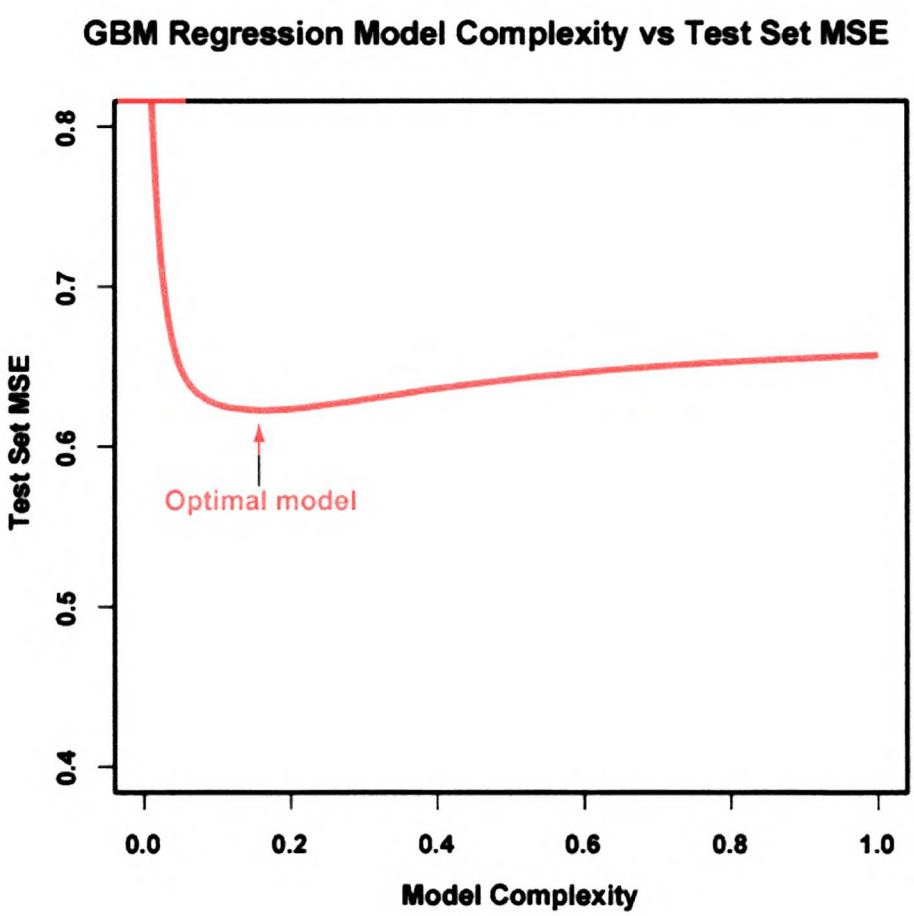


Figure 3-5. Test Set Mean Squared Error (MSE) as a function of model complexity for the GBM Regression (each data point averaged over 50 repetitions). Here, model complexity refers to the fraction of the 50,000 trees constructed by the algorithm. The optimal model contained 8,000 trees.

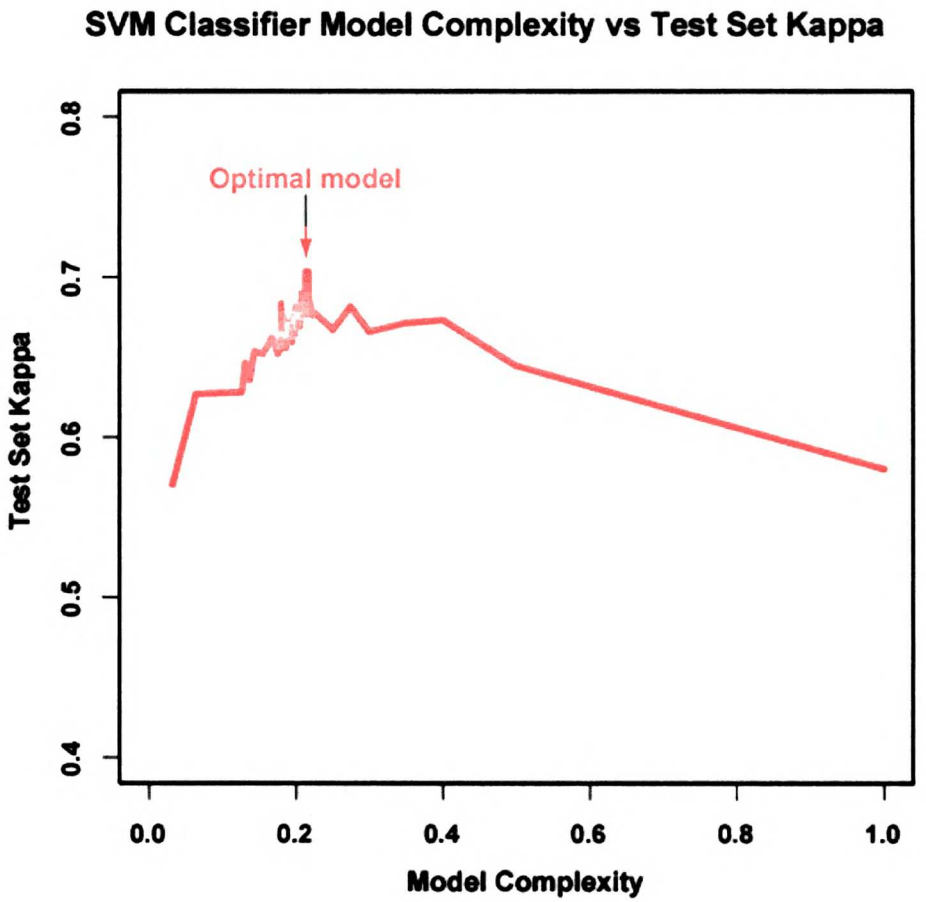


Figure 3-6. Test Set Kappa as a function of model complexity for the SVM Regression I (each data point averaged over >10 repetitions). Here, model complexity refers to the fraction of the 627 descriptors available to the algorithm. The optimal model employed 133 descriptors.

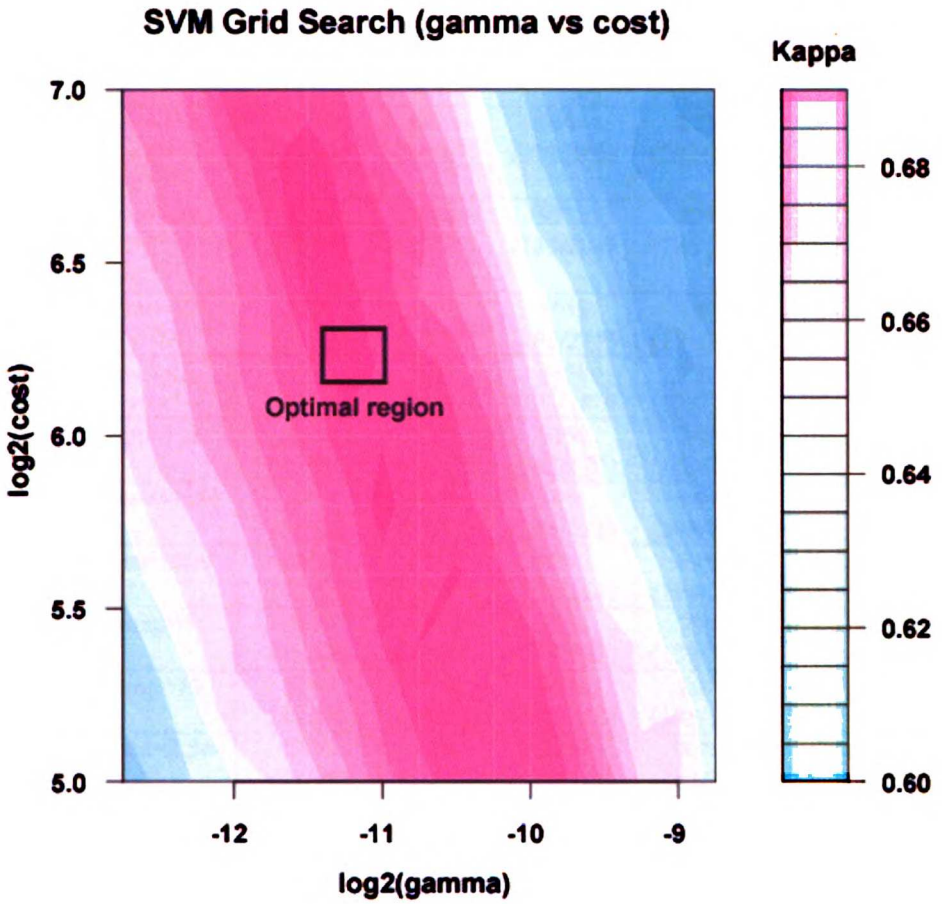


Figure 3-7. Test Set Kappa as a function of model complexity for the SVM Regression II (each data point averaged over 200 repetitions). Here, model complexity refers to the Cost and γ parameters of the algorithm. The optimal model had Cost=$2^{6.2}$ and $\gamma=2^{-11.25}$.

Modeling Results

Results from the final round of modeling are shown in Table 3-4.

Model	Parameters	Active Precision	Active Recall	Inactive Precision	Inactive Recall	Mis-classification Rate	Kappa
SVM Classifier	**ndesc=133, $C=2^{6.2}$, $\gamma=2^{-11.25}$**	**67% (34/51)**	**60% (34/57)**	**90% (212/235)**	**92% (202/219)**	**14% (40/276)**	**0.540**
SVM Regression	ndesc=109, $C=2^{2}$, $\gamma=2^{-6.75}$	56% (42/75)	74% (42/57)	93% (186/201)	85% (186/219)	17% (48/276)	0.524
DT Bagging	ntrees=32 ply=16	57% (35/61)	61% (35/57)	90% (193/215)	88% (193/219)	17% (48/276)	0.482
DT Boosting	ntrees=45, ply=45	49% (44/90)	61% (44/57)	93% (173/186)	79% (173/219)	21% (59/276)	0.462
Naïve Bayes	ndesc=172	42% (41/97)	72% (41/57)	91% (163/179)	74% (163/276)	26% (72/276)	0.368
GBM Classifier	**ntrees=84500**	**70% (31/44)**	**54% (31/57)**	**89% (206/232)**	**94% (206/219)**	**14% (39/276)**	**0.529**
GBM Regression	**ntrees=8000**	**62% (41/66)**	**72% (41/57)**	**92% (194/210)**	**89% (194/219)**	**15% (41/276)**	**0.572**
PLS	ncomp=5	48% (42/88)	74% (42/57)	92% (173/188)	79% (173/219)	22% (61/276)	0.439
PCR	ncomp=75	50% (44/88)	77% (44/57)	93% (175/188)	80% (175/219)	21% (57/276)	0.475
LARS	fraction=2.5e-07	53% (41/78)	72% (41/57)	92% (182/198)	83% (182/219)	19% (53/276)	0.484

Table 3-4. Results from the final round of computational modeling applied to the Random Set (57 Aggregators, 219 Non-aggregators; 22 compounds were ambiguous and removed from this study). The top three models are shown in bold.

The GBM Regression model performed the best according to the kappa metric, and yielded an active recall rate and misclassification rate similar to the naïve Bayes model and Random Forest, respectively, from the previous round of modeling. Interestingly, performance was improved by aggregating the predictions from the top two models—the GBM Regression (GBMr) and the SVM Classifier (SVMc)—such that any molecules classified as active in either model were designated promiscuous inhibitors. Indeed, the benefits of such consensus models were discussed earlier in Chapter 2. Adding the third best model, the GBM Classifier (GBMc), afforded only marginal improvement (Table 3-5).

Model	Active Precision	Active Recall	Inactive Precision	Inactive Recall	Misclassification Rate	Kappa
GBMr	62% (41/66)	72% (41/57)	92% (194/210)	89% (194/219)	15% (41/276)	0.572
GBMr + SVMc	58% (49/84)	**86% (49/57)**	96% (184/192)	84% (184/219)	16% (43/276)	**0.596**
GBMr + SVMc + GBMc	59% (50/85)	88% (50/57)	96% (184/191)	84% (184/219)	15% (42/276)	0.607

Table 3-5. Results from the aggregate models applied to the Random Set (57 Aggregators, 219 Non-aggregators; 22 compounds were ambiguous and removed from this study).

Thus, the consensus model derived from the GBMr and SVMc afforded both high active recall and low misclassification rate. The majority of promiscuous inhibitors in the Random set were identified while maintaining a reasonable false positive rate.

However, we were skeptical about this *post hoc* conclusion, as we suspected that the results might only be applicable to the Random set and not to external data. In order to assess the generality of the two-component model relative to the GBMr alone, we performed an additional round of cross-validation. In this experiment, the parameters of the component models were not re-optimized; instead, the values reported in Table 3-4 were used. Each model in Table 3-5 was trained on 85% of the 1030 molecules, and then applied to the remaining 15%. The procedure was repeated 200 times to yield the following results:

Model	Active Precision	Active Recall	Inactive Precision	Inactive Recall	Mis-classification Rate	Kappa
GBMr	67% (4.6%)	73% (6.8%)	88% (2.5%)	85% (3.1%)	19% (2.7%)	0.560 (0.064)
GBMr + SVMc	66% (5.7%)	83% (5.8%)	92% (2.5%)	82% (4.3%)	17% (3.4%)	0.609 (0.071)
GBMr + SVMc + GBMc	66% (5.7%)	82% (7.1%)	92% (2.9%)	83% (4.3%)	18% (3.5%)	0.603 (0.075)

Table 3-6. Results from cross-validation experiments using the consensus models applied to the 1030 molecules in the study (N=200, 85% train: 15% test). Standard deviations are reported in parenthesis.

As evident in Table 3-6, the two-component model appeared to perform better than the single component GBMr, and roughly equal to the three-component model. A One-Way ANOVA on kappa with respect to the three models rejected the null hypothesis that the differences in mean kappa across factors were insignificant (df=2, F=7.22, p < 0.001). The mean kappa for the GBMr was different from the GBMr + SVMc according to Student's t-test (p < 0.0005, 95% CI [-0.076, -0.022]). In agreement with the prior assessment, the difference in mean kappa between the two-component model and the three-component model was not statistically significant (p > 0.71, 95% CI [-0.024, 0.034]).

In conclusion, the best performing predictive model combined the GBM Regression and the SVM Classifier, such that any molecule predicted to be active in either model was classified as a promiscuous inhibitor.

3.5 Model Scalability

The computational efficiency of the consensus model was sufficient for annotating large virtual libraries (10^6 molecules per CPU-Day). However, peak performance was achieved by pre-processing the input into smaller sets of molecules. The R program functions that calculate the GBMr and SVMc models evaluate entire sets of molecules during a single call. As reported in Table 3-7, the time required to compute the consensus score per molecule degrades exponentially as the size of the input set increases. The ideal partition size, taking into account the overhead of creating subsets, was approximately ~10000 molecules.

Partition Size (molecules)	Time (seconds)	Molecules / Second
1000	51	19.6
10000	508	19.7
25000	1898	13.2
50000	4825	10.4

Table 3-7. Computational efficiency of calculating the consensus scoring model (see Methods section for details). The ideal partition size is ~10000 molecules.

3.6 Model Interpretation

In general, interpreting the effects of variables in GBM and SVM models is difficult due to non-linearity. GBM models, though additive in nature, are composed of hundreds or thousands of recursively-partitioned trees that can form highly complex response functions. SVM models convolve the relationships between descriptors by constructing the decision boundary in a transformed version of the original input space. Nevertheless, the importance of variables to the predictive performance of both learning methods can be estimated, albeit with some caveats. For example, the "relative

influence" of variables in GBMs (Friedman, 2001) can be assessed using the *summary* function in the R gbm package. Relative influence measures the amount a variable reduced the loss function during model training. The importance of a variable in an SVM model can be assessed by measuring the change in test error due to the absence of the descriptor; however, the true influence of the covariate can sometimes be masked by the presence of correlated variables that compensate for the missing contribution to the model.

Table 3-8 identifies the highest ranking variables from the GBMr and SVMc models in the present study according to relative influence and change in kappa, respectively.

GBMr Model Variables	Relative Influence	SVMc Model Variables	Δ Kappa
molalogp	**8.953**	Hamidine	-0.023
logP.o.w.	4.220	Nazo	-0.018
mollogD	3.569	SaasN	-0.017
HB2_Nxx	2.634	Emin1_OH2	-0.014
D4_DRY	2.344	BIT_-2090462286	-0.013
SHCsats	2.312	nXch7	-0.013
D2_DRY	**2.288**	Tm3	-0.013
SlogP	**2.265**	SdsN	-0.012
Hamide	1.913	vsa_base	-0.011
HB1_Nxx	1.702	BIT_436915834	-0.009
PEOE_VSA_N2	1.590	BV12_OH2	-0.007
PEOE_VSA_P2	1.445	BIT_-1090046377	-0.007
LogP	**1.424**	**molalogp**	**-0.007**
SHBd	1.369	SdssC	-0.006
D3_DRY	1.361	Sester	-0.004

Table 3-8. Important variables from the GBMr and SVMc models (list truncated to 15 for brevity). The relative influence values for the GBMr reflect an average over three independently-generated models. The SVMc variables are ranked according to amount kappa changed in the absence of the variable. Variables in boldface appear in the top 25 descriptors for the GBMr models and the 133 descriptors in the SVMc. Descriptors beginning with 'H' followed by a functional group indicate a sum over HE-states for the moiety. Variables beginning with 'S' indicate a sum over the E-states for a functional group or an atom in a chemical substructure; the 's','d', or 'a' refers to single, double, or aromatic bonds in the fragment. Descriptors beginning with BIT refer to individual bits from the Scitegic FCFP_6 fingerprint.

The table clearly identifies hydrophobicity as a dominant feature in promiscuous inhibition. Both models employed multiple direct measures of the octanol:water coefficient (molalogp, SlogP, LogP), and an indirect metric via D2_DRY, a descriptor describing the volume of interaction between a molecule and a non-polar probe (D1, D2, etc., refers to a particular energy level for the interaction). Correlation probably accounted for the lack of more hydrophobic terms in the top 15 list of the SVMc. This problem was less apparent in the GBMr method because it uses a bagging procedure, whereby only 50% of the variables were available for each round of tree construction. Under such conditions, variables correlated with other descriptors were given more of an opportunity to contribute to the model.

Further analysis of the GBMc covariates revealed three additional descriptors important for predictivity: PEOE_VSA_FNEG, SHCsats, and SHBd. PEOE_VSA_FNEG is the proportion of negatively-charged van der Waals surface area. SHCsats is the sum of the electrotopological states of all hydrogens (HE-states) on carbon atoms sp^3 bonded to saturated carbon. SHBd is the sum of HE-states for hydrogen bond donors. Hydrogen atoms bonded to or near an electronegative atom have high HE-state values. Histograms of these descriptors are shown in Figures 3-8 (molalogp is included for comparison).

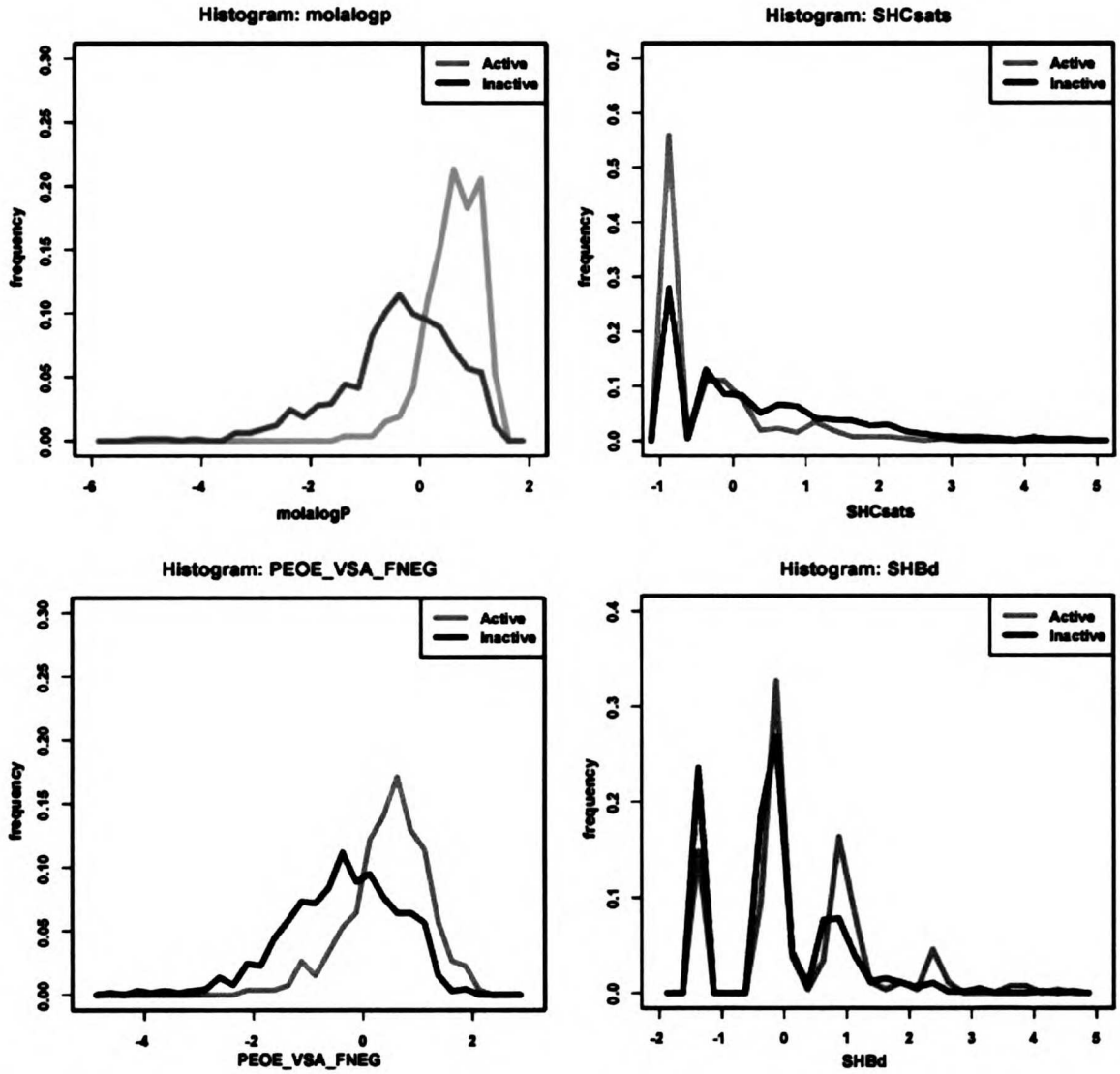


Figure 3.8. Histograms for influential descriptors in the GBMr model. In all graphs, the x axis is in standardized units (the descriptors values are mean-centered and scaled to unit variance).

As evident in the top left histogram of Figure 3-8, descriptors measuring logP discriminate between promiscuous inhibitors and inactive molecules well. The top right histogram demonstrates how inhibitors tend to have lower than average values for SHCsats, which is indicative of either a molecule with a non-polar environment around sp^3 carbons, or a lack of sp^3 carbons, or both. An interesting interpretation of this property is that an absence of tetrahedral centers might facilitate molecular packing,

consistent with the aggregator hypothesis for promiscuous inhibition. The bottom left histogram suggests active molecules have a larger portion of negatively charged van der Waals surface area. Other descriptors in both the SVM and GBM models support this assertion: inhibitors have larger values for PEOE_VSA_N0, PEOE_VSA_N1, PEOE_VSA_N2, and PEOE_VSA_N3 (proportion of van der Waals surface area with negative charge within a specified range); and larger values for PEOE_PC_N (total negative partial charge). Finally, the bottom left histogram associates promiscuous inhibitors with higher values for the SHBd descriptor. Higher SHBd values suggest either more polarized hydrogens available for H-bonding, a greater number of H-bond donors, or both. Other descriptors, such as HB1_Nxx and HB2_Nxx, support this claim (Figure 3-9).

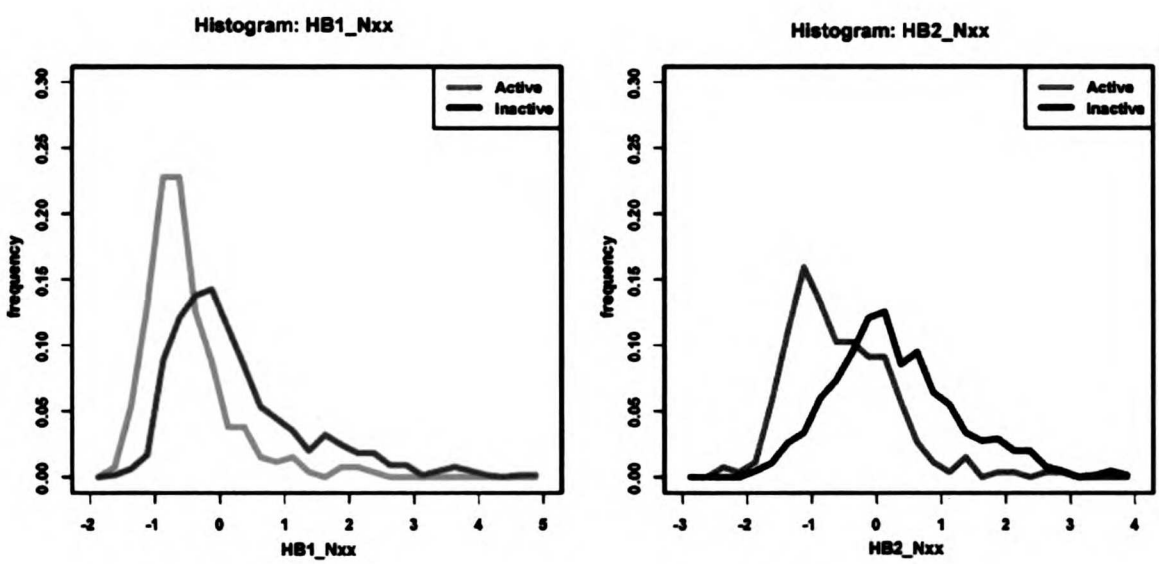


Figure 3-9. Histograms for HB1_Nxx and HB2_Nxx. In all graphs, the x axis is in standardized units (the values were mean-centered and scaled to unit variance). Only two of the HBX_Nxx series are shown as examples; in fact, the majority of HBX_Nxx metrics show discrimination between active and inactive molecules.

The HBX_Nxx series measures the difference in the volume of interaction (at energy range X) between a water probe (2 H-bond donors and 2 H-bond acceptors) and a

sp^2 N probe (0 H-bond donors and 1 H-bond acceptors) such as pyridine nitrogen. The water probe presents the optimal H-bond donor and acceptor for these calculations: any other probe will yield less favorable interactions. For promiscuous inhibitors, the HBX_Nxx metrics are lower than average, suggesting that there is less difference between the interactions of the sp^2 N probe and the ideal water probe. This fact supports the notion that activity is correlated with the strength of, or the number of, H-bond donors.

Thus far, the analysis has been limited to observing how a single variable segregates promiscuous inhibitors and inactive molecules. If the descriptors in the histograms from Figures 3-8 and 3-9 were independent, then their net effect towards predicting activity would be additive. However, the variables in most real world problems are often correlated to some degree. In such environments, the joint dependence of a descriptor—how the variable interacts with other variables in the context of activity—must be examined. Partial dependence plots attempt to assess these relationships for a subset of covariates by "averaging" out the contributions of all other variables. More formally, the partial dependence of a function of descriptors X, given a subset of interesting variables, X_S, and its complement, X_C, can be estimated as:

$$\bar{f}_S(X_S) = \frac{1}{N}\sum_{i=1}^{N} f(X_S, x_{iC}) \qquad (3\text{-}3),$$

where $\{x_{1C}, x_{2C}, \ldots, x_{NC}\}$ are the values of X_C in the training set (Hastie, 2001). This calculation is computationally efficient for GBMs and other tree based algorithms.

However, the number of covariates in X_S is often limited to two or three, as higher dimensional functions are difficult to visualize.

Figure 3-10 shows an example of a partial dependence plot for two perfectly correlated variables (in this case, the two variables are identical). Activity does not change in the vertical direction, indicating that the ordinate variable provides no additional effect on activity and can be completely accounted for by the abscissa variable. On the other hand, the checkered patterns in the graphs from Figure 3-11 reveal that both variables contribute to activity.

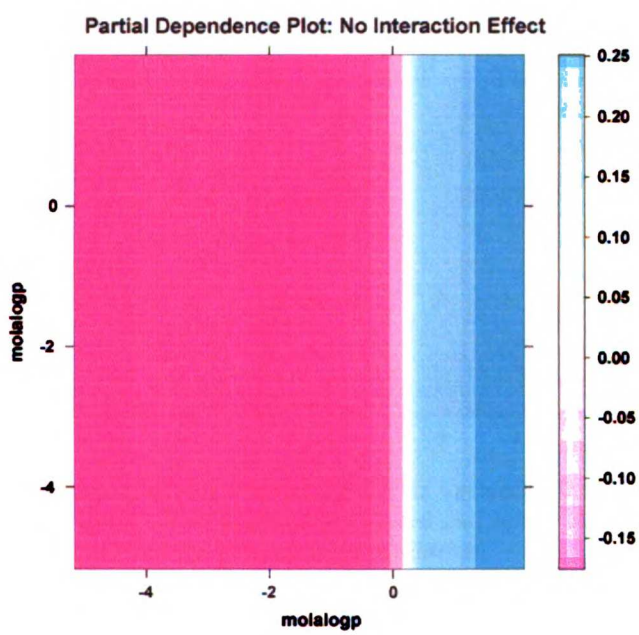


Figure 3-10. Partial dependence plots for two perfectly correlated variables. Cooler colors indicate increased favorability for promiscuous inhibition. The solid vertical bars indicate that the abscissa variable accounts for the entire effect on activity, making the ordinate variable redundant.

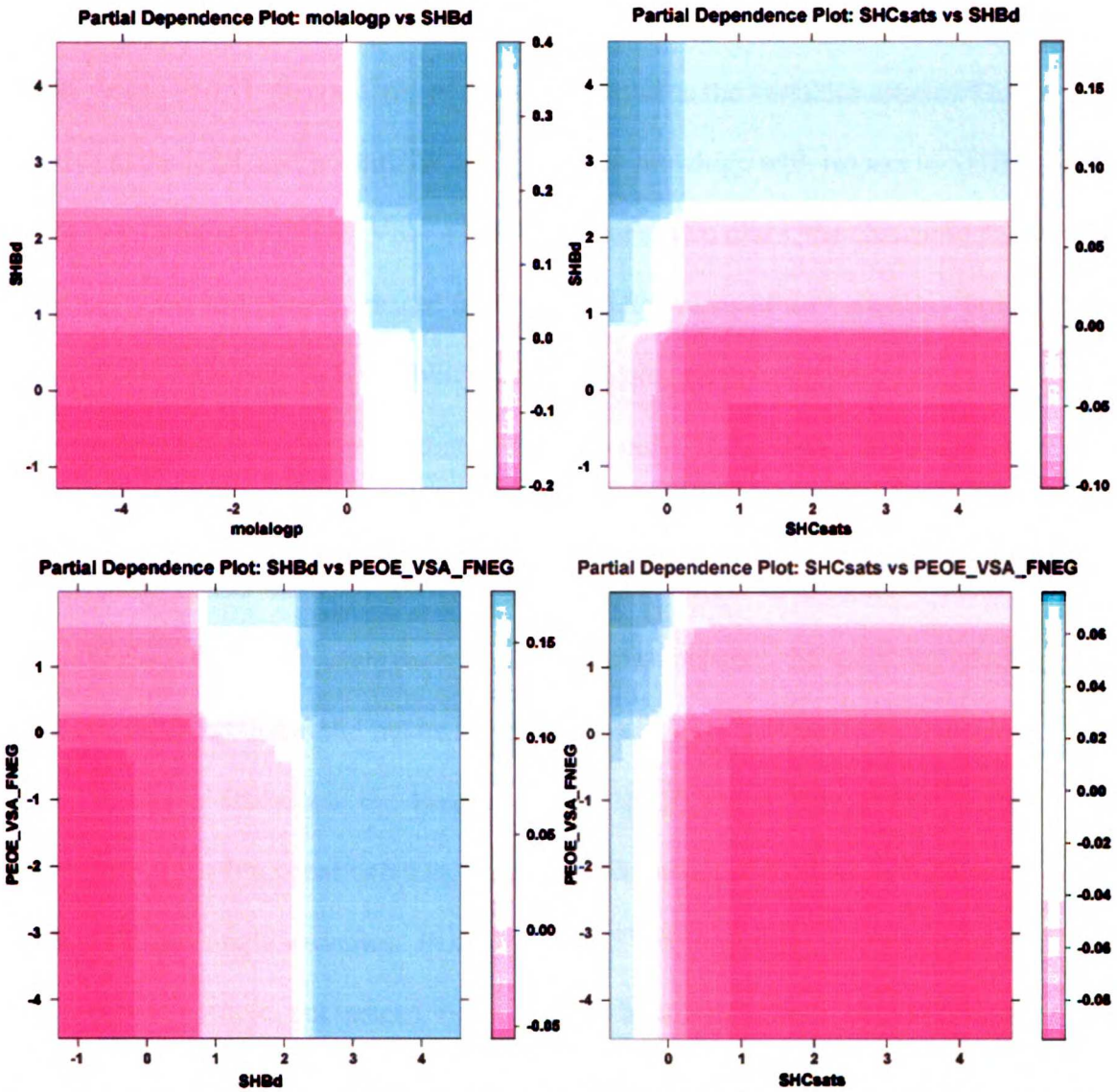


Figure 3-11. Partial dependence plots for interesting GBMr variables. In all graphs, the x axis is in standardized units (the descriptor values are mean-centered and scaled to unit variance). Cooler colors indicate increased favorability for promiscuous inhibition.

Additional insights emerged from Figure 3-11. First, it appeared that the four descriptors can be ranked in terms of their influence on activity as such: molalogp > SHBd > SHCsats > PEOE_VSA_FNEG. The plot of molalogp vs. SHBd indicated that a threshold value for the logP metric must be reached before inhibition is favorable, independent of the value of SHBd. This characteristic suggested that SHBd cannot overcome the influence of molalogP. Similar thresholds levels existed for SHBd with

respect to SHCsats and PEOE_VSA_FNEG, and SHCsats with respect to PEOE_VSA_FNEG. Second, the relationships between the variables appeared to be additive at the least, and potentially synergistic for molalogp with respect to SHBd and SHBd with respect to PEOE_VSA_FNEG. For these two cases, the change in the favorability for inhibition increased faster for certain values of the variables. For example, the difference in favorability at high molalogp across the entire range of SHBd was roughly $|0.40 - 0.15| = 0.25$; this change was only $|-0.20 - (-0.05)| = 0.15$ at low molalogp. Likewise, the change in favorability at high SHBd across the range of PEOE_VSA_FNEG was approximately $|0.17 - 0.09| = 0.08$, versus $|0.01 - (-0.05)| = 0.06$ at low SHBd. These two cases described interactions between the variables which produced an effect that could not be modeled via a linearly independent system.

A similar analysis of the descriptors from the SVMc model was more difficult. The GBMr algorithm constructed an additive model using decision trees built from binary splits on single variables. In contrast, the SVMc algorithm did not model the population as a whole, but instead focused on the boundary separating the two classes in a transformed version of the original input space. Histograms of the SVMc influential descriptors reported in Table 3-7 did not show significant separation between active and inactive populations. However, it was interesting that the SVM model had many descriptors pertaining to sp^2 or sp^3 conjugated nitrogen atoms: Hamidine (HE-state for amidine moiety), nazo (number of azo groups), SaasN (sum of E-states for nitrogen bonded to an aromatic ring), BIT_-2090462286 (Nitrogen containing substructure described by the SMARTS [*]=N[c]1:c:c:c:c:c:1), BIT_-1090046377 (Nitrogen containing substructure described by the SMARTS [*]C(=[*])NC(=O)[c](:c:[*]):c:[*]),

and SdsN (sum of E-states for nitrogen bonded another atom which contained a double bond). It was unclear how these variables influence promiscuous inhibition.

In summary, high values for logP, a low number of non-polar sp3 carbons, a high number of more polar hydrogen bond donors, and a larger proportion of negatively charged van der Waals surface area tended to favor promiscuous inhibition. Furthermore, the character of sp^2 or sp^3-conjugated nitrogen atoms also appeared to be an important predictor of activity.

3.7 Model Failures

Five of the 57 promiscuous inhibitors in the Random set failed to be correctly identified by any of the computational models, including the Random Forest.

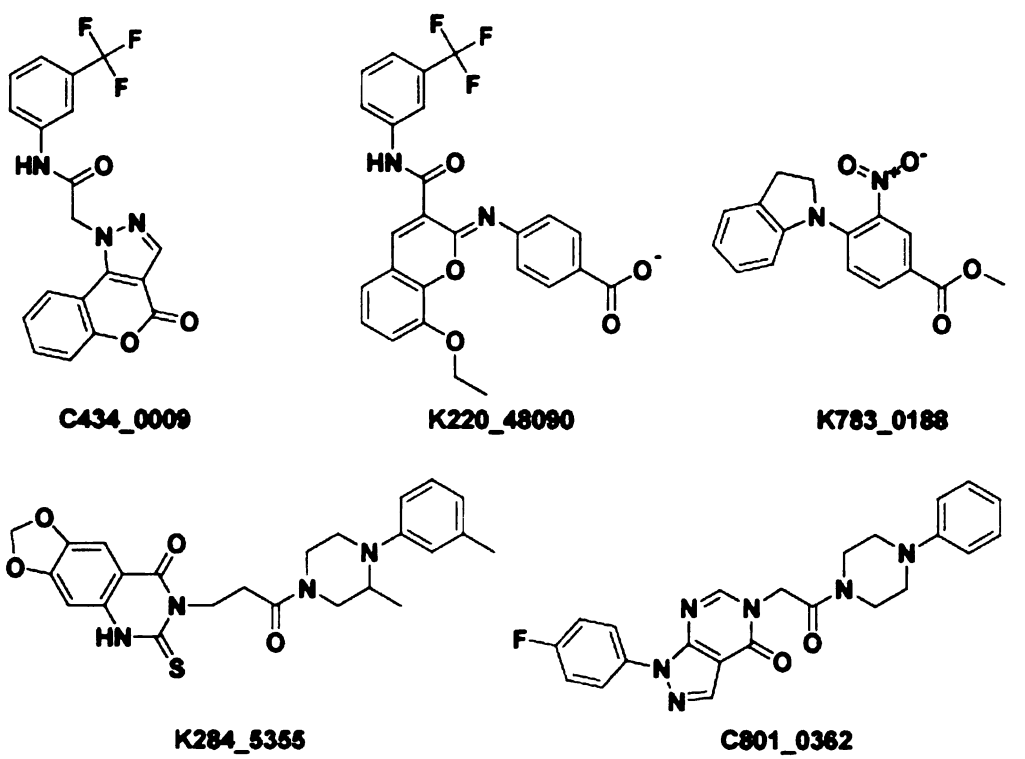


Figure 3-12. Molecules that failed to be classified as active in any computational model.

A closer examination of one of these molecules, K284_5355, highlights the difficulty in predicting promiscuous inhibitors (Figure 3-13). Four molecules similar to K284_5355 were retrieved from the set of 1030 compounds (Tanimoto Similarity > 0.3, Scitegic FCFP_6 fingerprint). Despite the fact that each similar molecule shared either the dioxolo-quinazolinone or the 4-phenyl-piperazine moieties present in the query molecule, none were active.

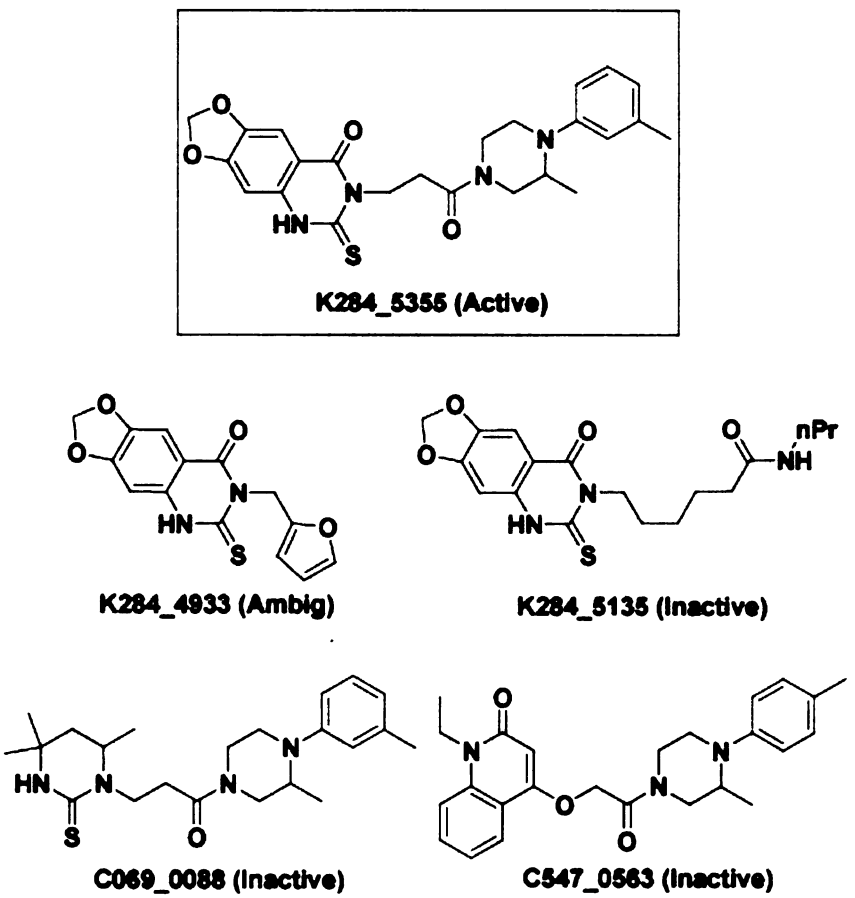


Figure 3-13. Molecules similar to K284_5355 (Tanimoto Similarity > 0.30, FCFP_6 fingerprint).

In some circumstances, then, promiscuous inhibition appears to be a complex phenomenon emerging only in particular molecular contexts.

3.8 Conclusion

In this chapter, we introduced the concept of promiscuous inhibitors and described how such molecules could be identified using a high-throughput enzyme-based screen. We then demonstrated how HT-DLS, a technique that measures particle formation in solution, failed to correlate with nonspecific inhibition, despite considerable evidence linking the activity with compound aggregation. Using algorithms designed for data mining in high dimensions, we developed a computational model that correctly classified > 80% of promiscuous inhibitors, while maintaining a misclassification rate of < 20%. This consensus model should be a valuable tool for flagging suspicious or undesirable molecules during such tasks as chemical library selection and the analysis of HTS screening hits.

An investigation of the model's descriptors revealed how certain physical properties, such as high logP, are associated with promiscuous inhibition. However, the presence of inhibitors that were not correctly classified by any computational model underscored our limited understanding of this phenomenon. The interesting covariates identified in this study should provide the starting point for further investigation into the mechanisms behind this complex biophysical property.

Chapter IV. Assay Reporter: An Integrated Informatics Environment for Identifying "Good" Hits from HTS Data

In the previous three chapters, we described the development and application of machine learning methods to predict interesting molecular properties. In Chapter 1, a naïve Bayes algorithm was trained to identify ΔF508 CFTR potentiators; the model's interpretation formed the foundation of pharmacophore-based hypotheses which guided the search for novel, more potent compounds. In Chapters 2 and 3, we detailed the construction and performance of a consensus model for detecting promiscuous inhibitors—compounds that nonspecifically perturb the function of biological macromolecules. This predictor can be used to flag likely false-positives in HTS data, thereby reducing the amount of time and resources expended for costly re-screening and orthogonal assays. Such data mining techniques accelerate lead discovery by facilitating the analysis of primary screening data.

Unfortunately, when applied in isolation, the gains from these computational tools are modest. Translating an HTS hit into a *bona fide* lead scaffold with high potency and selectivity, good bioavailability, and little toxicity is a formidable multivariate optimization problem. The relationships between important properties are often non-linear, and attempts to improve one molecular characteristic might be detrimental to another. In order to make the best decisions on how to proceed, pertinent information must be readily accessible, appropriately indexed, and examined *in the context of other chemical and biological knowledge*. In this chapter, we describe the Assay Reporter, a computational framework for integrating all available data in an HTS experiment. Our methods provide a work flow that helps to identify "good" hits—molecules that are more likely to become quality leads—from the assay. Examples of results from screens

conducted at the Bay Area Screening Center are presented. We conclude with comments on the future direction of this research.

4.1 Introduction to HTS Analysis

Traditionally, drug leads were derived from natural products with a proven history of *in vivo* activity (Lipinski, 2004). As a result, such molecules had already been filtered for minimal levels of bioavailability and cytotoxicity. The paradigm of drug discovery changed in the 1980s with advances in robotics, molecular biology, and combinatorial chemistry. Synthesizing or screening compounds were no longer rate-limiting steps: hundreds of thousands of compounds could be assayed directly against a biomolecular target using high-throughput methods. But despite the large number of initial hits generated from this technology, the number of New Chemical Entities (NCE) reaching the market has remained constant (Bleicher, 2003). Seminal work by Lipinski et al demonstrated that the quality of hits emerging from HTS was poor (Lipinski, 2000). A number of factors, from the techniques employed by combinatorial chemistry to the physical methods of detecting active molecules, biased hits towards high lipophilicity, high molecular weight, and low solubility. Such compounds proved to be unworkable and failed to advance to subsequent rounds of development.

The unfulfilled promise of combinatorial chemistry and HTS spawned two recent movements: early ADMET (absorption, distribution, metabolism, excretion, and toxicity) assessment and robust methods for HTS data analysis. The notion of applying ADMET criteria earlier in drug discovery projects stems from the fact that any orally administered drug must be absorbed, distributed to the site of action, and then excreted either

unchanged or as a metabolite. It follows then that only compounds possessing such properties, or able to be modified accordingly, should be screened. Consequently, a considerable amount of research has been directed at describing "drug-likeness," the characteristics that differentiate drugs from other organic compounds[2]. Some of the computational methods now employed to predict these qualities include simple counting schemes such as Lipinski's "rule of five" (Lipinski, 2001) and similar work (Ghose, 1999; Oprea, 2000), knowledge-based metrics (Andrews, 1984; Muegge, 2002) gleaned from medicinal chemists, substructure filtering techniques such as REOS (Walters, 1998), chemical space formulations (Oprea, 2002), and machine learning models (Ajay, 1998).

In parallel, researchers have developed new statistical techniques for assessing the quality of HTS assays. High-throughput methods, due to the reliance on automation, parallelization, and miniaturization, are often prone to significant systematic error. Moreover, screening compounds with behaviors such as auto-fluorescence and promiscuous inhibition can exacerbate the problem. In response, quality control procedures were developed to help detect specious results. For example, Zhang et al proposed Z-prime and Z-factor, a pair of metrics that capture aspects of the performance of entire plates in a single value (Zhang, 1999). Z-prime measures the separation between positive and negative controls; Z-factor assesses how well the positive control is distinguished from the screening compounds, which are presumed to be mostly inactive. Irregular plates can be quickly identified by surveying a scatter plot of both metrics. Alternatively, Brideau et al reported a method to reduce systematic error in HTS that was independent of the controls (Brideau, 2005). Their "B-scores" correction employs

[2] For an excellent review, see Walters, 2002.

techniques including median polishing and temporal smoothing to normalize the assay response of the screening compounds, thereby providing a more accurate measure of activity.

Thus, a thorough investigation of HTS data must include an assessment of both quality control (QC) and ADMET parameters. For example, a biologist might define an interesting "hit" as one the most active molecules in a set that satisfies some statistically significant threshold (e.g., minimum Z score) for activity. On the other hand, a chemist might prefer a less active molecule that corresponds better to her notion of "drug-likeness." The definition of a "good" hit, then, takes into account confidence in the biological information derived from the assay, and any relevant chemical knowledge about the molecule. Indeed, the occurrence of such behaviors as promiscuous inhibition demonstrates that the concepts of QC and ADMET are intimately related; successful quality control requires knowledge about the chemical properties of the molecules in the HTS.

The goal of the Assay Reporter is to bring these disparate streams of information together in a single place where biologists and chemists can jointly assess the quality of a screening hit. For each HTS project, our system employs a relational database to store both the raw and metadata from biological assays and the chemical properties of the screening library. The "drug-likeness" of molecules is predicted using computational ADMET models such as the ones described earlier in this chapter. In parallel, QC measures of the reliability of the biological data are calculated and suspicious results are automatically flagged for in-depth investigation. Our algorithms then query the database

to produce an html hierarchy that reports the original HTS data and all derivative information in a readily accessible and intuitive manner.

The output from our program mimics a natural work flow. The root page contains visual elements such as heat maps and scatter plots that facilitate the detection of systematic errors present in the HTS. Once assay results are reviewed and verified, investigators can follow a link to a separate page containing an analysis of the top hits (defined by a user-defined cut-off). Here, the hits are ranked according to a user-defined rubric that takes into account activity, the confidence in the assay result, chemical similarity to known bioactive molecules, commercial availability, and ADMET properties. Each molecular record contains URLs to additional pertinent information—a table detailing the activity of the compound in prior assays and a preliminary SAR based on automated similarity and substructure searches.

By integrating chemical and biological data, Assay Reporter allows investigators to weigh the assay results of HTS hits in the context of other important data, thereby providing a more complete view of the suitability of the compound for further development.

4.2 Methods

All programs are executed within the Pipeline Pilot 4.5.2 environment using native PilotScript and Perl 5.8.1. Scatter plots and statistical metrics were calculated via external calls to the R program (version 2.0.1). Data pertaining to the screening library, chemical properties, and assay results was stored in mySQL databases (version 4.1.7).

4.3 Database Structure

The entity relationship diagram (ERD) for the Assay Reporter relational database is presented in Figure 4-1. The three colored squares group collections of related data: location and source information about the screening library (blue), chemical properties of the screening compounds (green), and biological assay results and associated meta-data (red). Each text box represents a table in the database; the rows in each box define the name and type of data stored. Relationships between records from different tables are identified by connecting lines. For example, Rel_01 in the green box indicates that a single record in the *compound* table corresponds to a single record in the *molprops* table. On the other hand, every record in the *protocol* table is associated with one or more records in the *assay* table (red box).

According to the principles of relational database design, each record in a table must be designated with a unique value, or primary key (identified by the key icon in Figure 4-1). More than one column can be aggregated to form this element. Furthermore, redundant information is eliminated by employing foreign keys (denoted by the FK in Figure 4-1), data columns whose values are derived from another table. These constraints save storage space and preserve data integrity by avoiding the need to simultaneously update different tables sharing the same information.

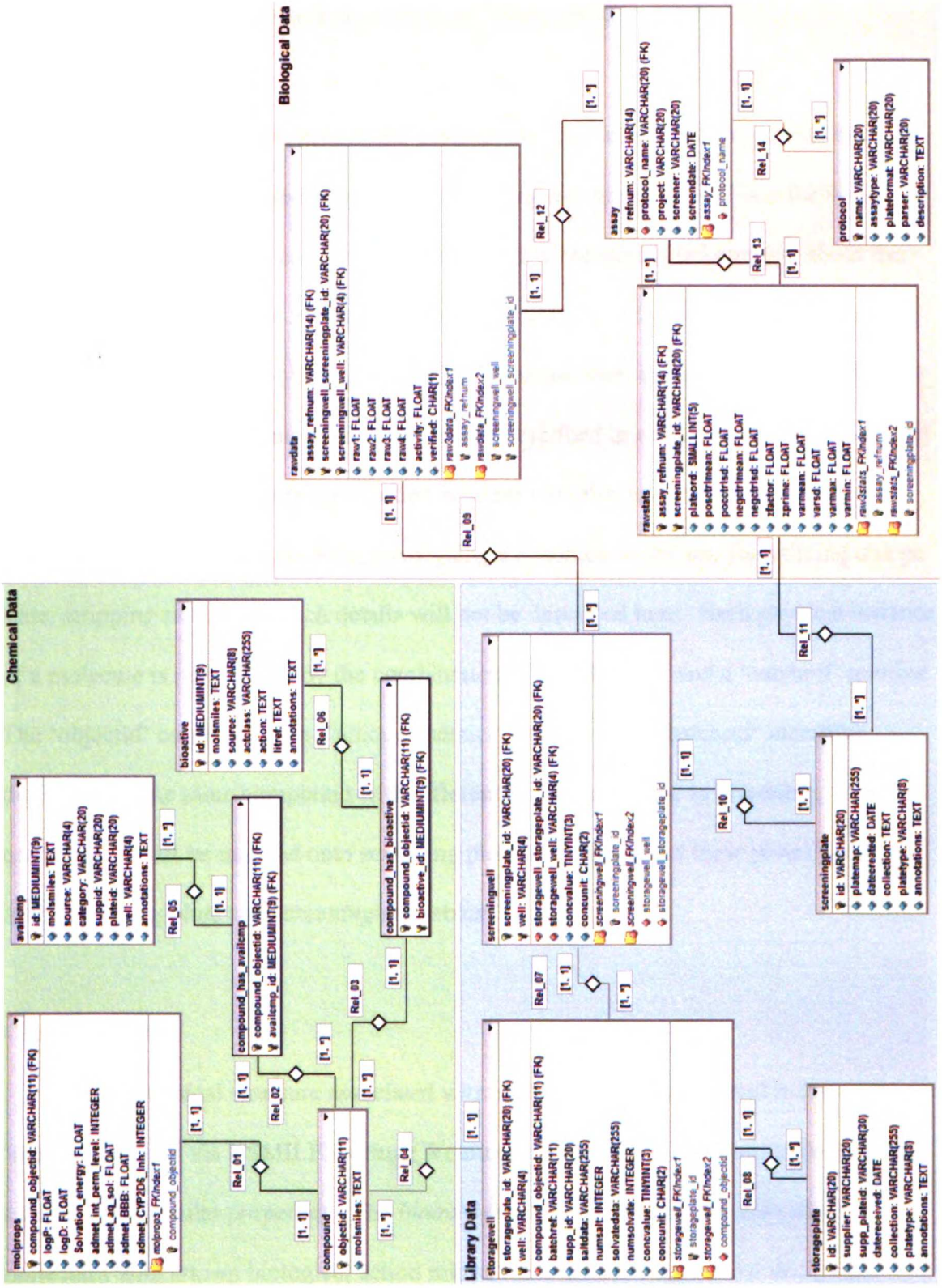


Figure 4-1. Entity Relationship Diagram (ERD) for the Assay Explorer relational database.

Details of the three data groups are described below:

Library Data

These tables contain source and plate information about the compounds in the screening library. Purchased or synthesized compounds are registered into the system as 96 or 384 well plates with an accompanying data file containing information about the contents of each well. The *storageplate* table keeps track of the plate IDs for these incoming items as well as other annotation information such as the date received and the supplier name. The contents of these plates are described in *storagewell.* The process of extracting chemical structure information is rather complex and involves topological assessments and manipulations (e.g., assigning a canonical tautomer, formalizing charge state, stripping salts, etc.); such details will not be described here. Each physical instance of a molecule is represented by the combination of an 'objectid' and a 'batchref' number. The 'objectid' corresponds to a unique chemical topology; the 'batchref' identifies duplicates of the same compound with different salt forms. Prior to screening, compounds must be mapped onto screening plates. Details about these plates are stored in the *screeningplate* and *screeningwell* tables.

Chemical Data

The chemical structure associated with each 'objectid' is reported in the 'molsmiles' field via a SMILES string (Weininger, 1988). The *molprops* table contains calculated molecular properties. The *bioactives* table contains information about molecules with known biological action mined from various commercial and public databases. Likewise, the *availcmps* table provides information about commercially

available compounds. Using the Tanimoto similarity metric, compounds in these two tables can be related to compounds in the screening library by a many-to-many relationship; this similarity matrix is stored in the *compound_has_bioactive* and *compound_has_availcmp* tables.

Each table in the Chemical Data collection must be updated every time a new chemical entity enters the system. Additionally, the *compound_has_bioactive* and *compound_has_availcmps* tables must be modified to reflect changes in their parent tables. However, the computational cost of this method is significantly cheaper than calculating these properties for each Assay Reporter run.

Biological Data

The *assay* and *protocol* tables in this collection contain metadata about each HTS experiment, such as the name of the screener, the date screened, and a description of the protocol employed. Each assay is an instance of a protocol, and is related to the *rawdata* and *rawstats* tables via a one-to-many relationship. The *rawdata* table parses the output of the assay detection devices: all data associated with a plate and well location, and the calculated activity relative to the control, is stored here. Statistical measures for each plate are captured in the *rawstats* table.

4.4 Assay Reporter Output

The Assay Reporter generates an html hierarchy for each HTS project—a set of assays conducted under a single protocol for a single target. Rather than providing a cursory description of all aspects of the output, this section will focus on specific

examples of how the application has helped identify systematic errors, suspicious results, and "good" HTS hits using real-world data from screens conducted at the Bay Area Screening Center.

Diagnostics Page

The first page encountered by investigators assesses the quality of the HTS using data visualization tools and summary statistics. Scatter plots of activity vs. compounds in the order screened can be a powerful tool for discerning anomalies in the experiment. Figures 4-2 through 4-4 depict common HTS problems.

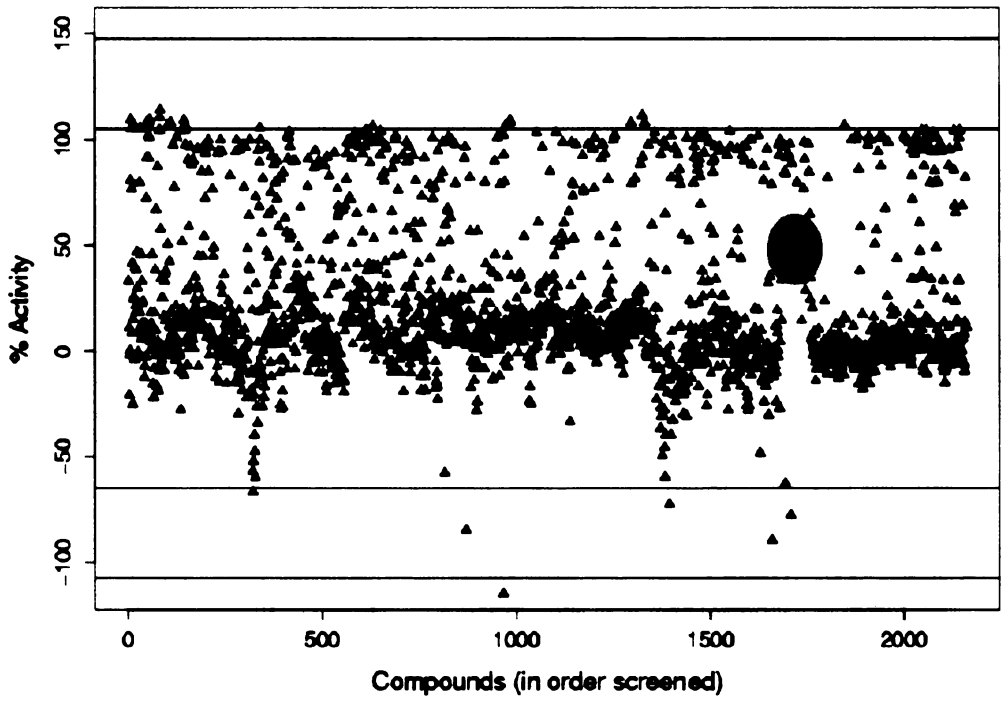


Figure 4-2. Detecting assay errors using the activity scatter plot I. The mean activity of the plate containing the compounds marked by the red sphere is shifted (red and green lines correspond to Z score equal to 1.96 and 2.5). This could be indicative of a malfunction in the controls, aberrant reaction conditions, or a peculiarity of the compounds on the plate.

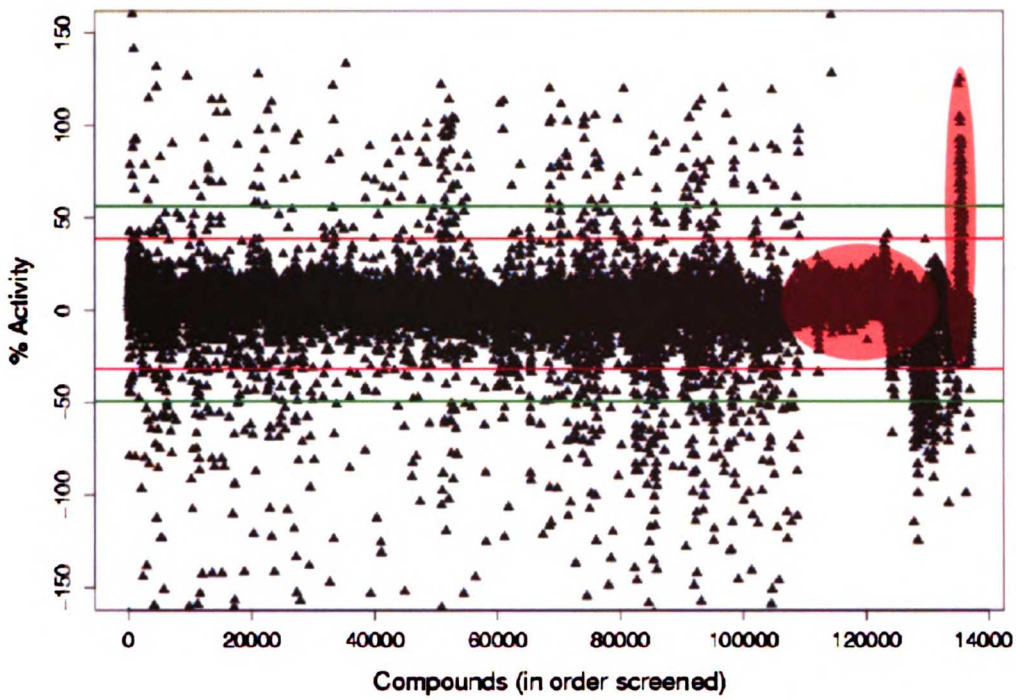


Figure 4-3. Detecting assay errors using the activity scatter plot II. Areas of extremely low and high variance are marked by red ellipses (red and green lines correspond to Z score equal to 1.96 and 2.5). This characteristic is highly usual for the type of assay employed (fluorescence polarization).

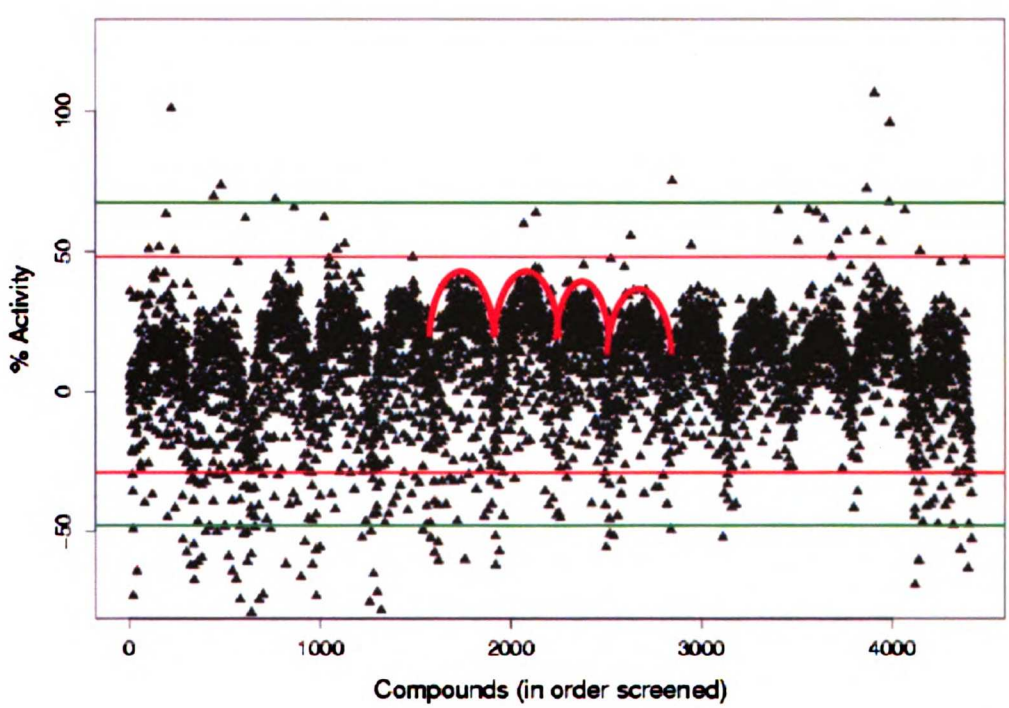


Figure 4-4. Detecting assay errors using the activity scatter plot III. The thick red curves describe periodicity in the assay with frequency on the order of one plate (red and green lines correspond to Z score equal to 1.96 and 2.5), indicative of positional effects.

The scatter plot for an optimal assay would normally be tightly clustered around zero activity, as the vast majority of randomly selected compounds would be inactive. The quality of the HTS in Figure 4-2 is generally good, except for a single plate wherein the mean activity is shifted higher. Such behavior might indicate a malfunction in the controls on this plate, or aberrant reaction conditions for the screening compounds. Alternatively, the activity results could be accurate if derived from a cell-based screen using compounds known to be biologically active (e.g., cytotoxins). In Figure 4-3, the activity scatter plot suggests that the assay failed at the end of the screen. This HTS employed a fluorescence polarization detection method known to produce high scatter in both positive and negative directions of the activity axis (Gribbon, 2003); the unusually low and high variance after 110,000 compounds merits further investigation. Finally, the periodicity in the screen in Figure 4-4 is indicative of positional effects, whereby the upper left-hand and lower right-hand corners of the plate have intrinsically lower activities due to systematic errors (see well analysis section below). Thus, an entire HTS can be quickly surveyed for spurious results using activity scatter plots.

The behavior of individual plates can be further assessed using scatter plots of the Z-prime and Z-factor metrics described earlier in this chapter. Assay Reporter automatically flags poor performing outliers for both statistics using a user-defined Z score cutoff; these plates are reported in a table which provides a link to the plate heat map. An example of this process is detailed in Figure 4-5. Tracing back to the suspicious plate revealed discrepancies in the negative control wells. The top and bottom wells had unusually high and low activities respectively, suggesting an error in the handling of these controls.

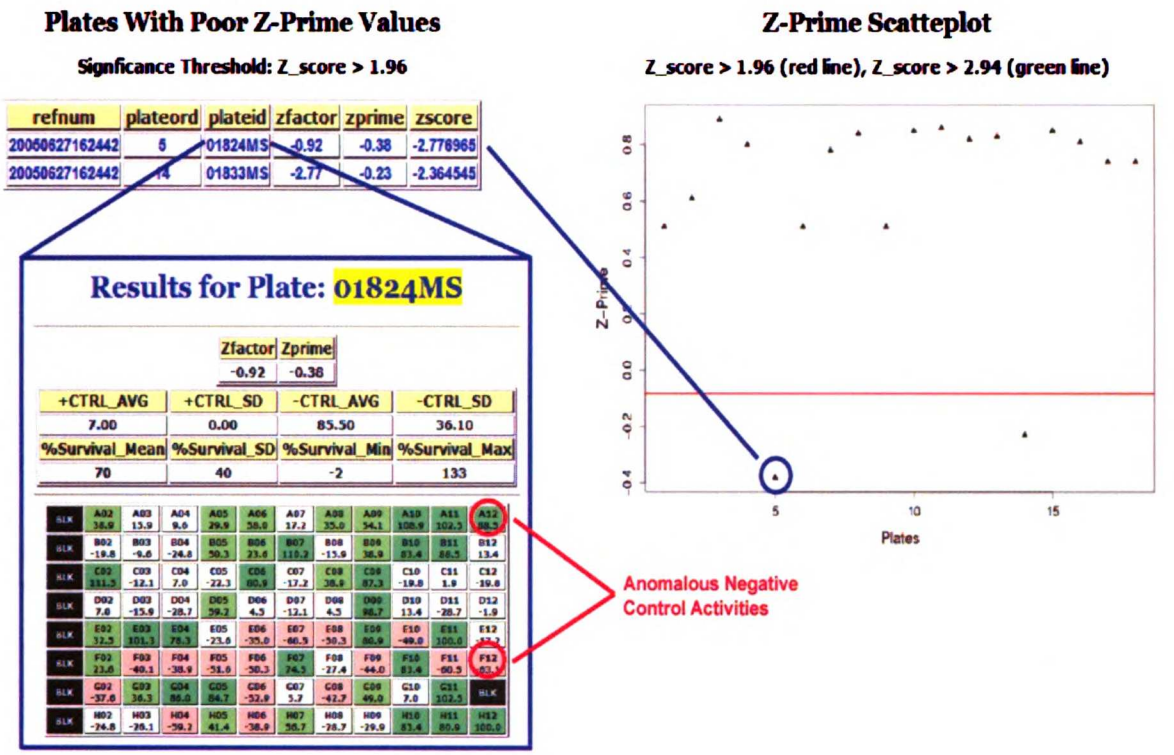


Figure 4-5. Outlier detection in the Z-prime scatter plots reveals errors in the negative control wells for the plate.

Additional examples of the utility of this type of analysis are reported in Figure 4-6, which presents the heat maps for two plates flagged for poor Z-factor. The grid-like pattern of activity probably stems from a failure to properly deliver reagents to the wells because of liquid handling errors. Remarkably, both plates would have escaped detection if only Z-prime was assessed.

Zfactor	Zprime
-3.16	0.91

+CTRL_AVG	+CTRL_SD	-CTRL_AVG	-CTRL_SD
72.00	0.70	111.80	0.40
Plate_Ratio_Mean	**Plate_Ratio_SD**	**Plate_Ratio_Min**	**Plate_Ratio_Max**
126.80	75.30	97.00	496.00
Intensity_Mean	**Intensity_SD**	**Intensity_Min**	**Intensity_Max**
21959400	8270220	401653	138332000

A01 100.0	A02 102.5	A03 14.5	A04 4.4	A05 -0.6	A06 1.9	A07 11.9	A08 4.4	A09 9.4	A10 -0.6	A11 4.4	A12 9.4	A13 1.9	A14 -5.7	A15 6.9	A16 6.9	A17 1.9	A18 11.9	A19 1.9	A20 1.9	A21 -10.7	A22 1.9	BLK	BLK
B01 100.0	B02 97.5	B02 -923.9	B04 -913.8	B05 -944.0	B06 -750.3	B07 -3.1	B08 -0.6	B09 -0.6	B10 4.4	B11 4.4	B12 9.4	B13 -3.1	B14 11.9	B15 -0.6	B16 9.4	B17 37.1	B18 1.9	B19 -0.6	B20 6.9	B21 14.5	B22 6.9	BLK	BLK
C01 -0.6	C02 -0.6	C03 -0.6	C04 -3.1	C05 -0.6	C06 1.9	[illegible]	C08 4.4	C09 1.9	C10 6.9	C11 -3.1	C12 -5.7	C13 4.4	C14 6.9	C15 6.9	C16 -0.6	C17 11.9	C18 6.9	C19 6.9	C20 4.4	C21 -0.6	C22 6.9	BLK	BLK
D01 1.9	D02 -0.6	D03 -931.3	D04 -964.1	D05 -988.7	D06 4.4	D07 1.9	D08 -3.1	D09 -8.2	D10 -0.6	D11 -0.6	D12 1.9	D13 -5.7	D14 6.9	D15 11.9	D16 -0.6	D17 -5.7	D18 -30.8	D19 -8.2	D20 1.9	D21 4.4	D22 1.9	BLK	BLK
BLK	BLK	E03 4.4	E04 -0.6	E05 4.4	E06 6.9	E07 -5.7	E08 6.9	E09 -0.6	E10 -3.1	E11 -0.6	E12 1.9	E13 -3.1	E14 -0.6	E15 -8.2	E16 -0.6	E17 1.9	E18 6.9	E19 -5.7	E20 -8.2	E21 -10.7	E22 -10.7	BLK	BLK
BLK	BLK	F03 -8.2	F04 4.4	F05 6.9	F06 14.5	F07 -3.1	F08 4.4	F09 -3.1	F10 -3.1	F11 -0.6	F12 9.4	F13 4.4	F14 -0.6	F15 14.5	F16 6.9	F17 -3.1	F18 6.9	F19 4.4	F20 6.9	F21 1.9	F22 4.4	BLK	BLK
BLK	BLK	G03 4.4	G04 9.4	G05 1.9	G06 4.4	G07 -0.6	G08 1.9	G09 4.4	G10 6.9	G11 6.9	G12 1.9	G13 -10.7	G14 -0.6	G15 6.9	G16 6.9	G17 4.4	G18 1.9	G19 -0.6	G20 -25.8	G21 4.4	G22 6.9	BLK	BLK
BLK	BLK	H03 9.4	H04 4.4	H05 6.9	H06 -3.1	H07 4.4	H08 -8.2	H09 6.9	H10 -13.2	H11 1.9	H12 1.9	H13 -3.1	H14 -0.6	H15 -3.1	H16 1.9	H17 -0.6	H18 9.4	H19 1.9	H20 4.4	H21 4.4	H22 1.9	BLK	BLK
BLK	BLK	I03 4.4	I04 4.4	I05 4.4	I06 4.4	I07 4.4	I08 6.9	I09 -0.6	I10 -8.2	I11 -0.6	I12 4.4	I13 -3.1	I14 -0.6	I15 1.9	I16 1.9	I17 -3.1	I18 -0.6	I19 11.9	I20 4.4	I21 -0.6	I22 14.5	BLK	BLK
BLK	BLK	J03 -3.1	J04 -3.1	J05 1.9	J06 -5.7	J07 1.9	J08 -0.6	J09 1.9	J10 1.9	J11 1.9	J12 11.9	J13 -15.7	J14 4.4	J15 11.9	J16 -0.6	J17 1.9	J18 6.9	J19 -3.1	J20 19.5	J21 -0.6	J22 14.5	BLK	BLK
BLK	BLK	K03 6.9	K04 -0.6	K05 11.9	K06 4.4	K07 -3.1	K08 1.9	K09 1.9	K10 -3.1	K11 -10.7	K12 4.4	K13 4.4	K14 -0.6	K15 6.9	K16 -5.7	K17 9.4	K18 11.9	K19 11.9	K20 6.9	K21 1.9	K22 11.9	BLK	BLK
BLK	BLK	L03 -931.3	L04 -8.2	L05 -5.7	L06 -5.7	L07 4.4	L08 -3.1	L09 4.4	L10 6.9	L11 1.9	L12 1.9	L13 11.9	L14 -3.1	L15 11.9	L16 -8.2	L17 1.9	L18 6.9	L19 -0.6	L20 9.4	L21 11.9	L22 9.4	BLK	BLK
BLK	BLK	M03 9.4	M04 4.4	M05 -15.7	M06 -3.1	M07 1.9	M08 11.9	M09 1.9	M10 1.9	M11 1.9	M12 19.5	M13 -0.6	M14 -3.1	M15 9.4	M16 6.9	M17 6.9	M18 -3.1	M19 17.0	M20 1.9	M21 6.9	M22 11.9	BLK	BLK
BLK	BLK	N03 -923.9	N04 -886.7	N05 -18.2	N06 14.5	N07 -0.6	N08 6.9	N09 -0.6	N10 -0.6	N11 6.9	N12 4.4	N13 6.9	N14 -13.2	N15 4.4	N16 1.9	N17 1.9	N18 1.9	N19 17.0	N20 9.4	N21 11.9	N22 11.9	BLK	BLK
BLK	BLK	O03 1.9	O04 1.9	O05 6.9	O06 14.5	O07 17.0	O08 9.4	O09 1.9	O10 -0.6	O11 6.9	O12 9.4	O13 1.9	O14 9.4	O15 -0.6	O16 11.9	O17 17.0	O18 9.4	O19 6.9	O20 17.0	O21 17.0	O22 4.4	BLK	BLK
BLK	BLK	P03 -961.6	P04 -926.4	P05 -998.2	P06 -898.7	P07 14.5	P08 6.9	P09 11.9	P10 17.0	P11 -0.6	P12 22.0	P13 9.4	P14 4.4	P15 11.9	P16 11.9	P17 4.4	P18 11.9	P19 9.4	P20 24.5	P21 17.0	P22 17.0	BLK	BLK

Zfactor	Zprime
-39.74	0.20

+CTRL_AVG	+CTRL_SD	-CTRL_AVG	-CTRL_SD
517.00	5.50	726.10	50.40
Plate_Ratio_Mean	**Plate_FRET_SD**	**Plate_FRET_Min**	**Plate_Ratio_Max**
549.00	428.50	112.00	6476.00
Intensity_Mean	**Intensity_SD**	**Intensity_Min**	**Intensity_Max**
45187200	66780700	6381330	340363000

A01 2.9	A02 -1.4	A03 2.9	A04 -5.7	A05 3.4	A06 0.0	A07 1.0	A08 -8.1	A09 0.5	A10 -35.8	A11 -4.3	A12 -1.4	A13 -5.2	A14 -4.3	A15 -2.4	A16 -3.3	A17 1.0	A18 24.4	A19 -11.9	A20 -6.2	A21 -2710.3	A22 -6.7	A23 100.0	A24 97.6
B01 -6.7	B02 -4.7	B03 138.3	B04 285.6	B05 30.2	B06 169.4	B07 117.2	B08 287.5	B09 285.1	[illegible]	B11 287.5	B12 24.4	B13 286.6	B14 -5.7	B15 281.8	B16 -9.0	B17 280.8	B18 0.5	B19 280.8	B20 -6.2	B21 99.1	B22 18.7	B23 99.1	B24 98.6
C01 -11.0	C02 -6.7	C03 -0.4	C04 -7.6	C05 0.0	C06 -6.2	C07 0.0	C08 -1.4	C09 -3.3	C10 -0.4	C11 -7.1	C12 -11.9	C13 0.5	C14 -6.2	C15 -2.4	C16 1.0	C17 -8.6	C18 74.7	C19 -15.3	C20 -8.6	C21 -25.8	[illegible]	C23 98.1	C24 101.5
D01 16.8	D02 -3.8	D03 83.8	D04 284.2	D05 285.8	D06 285.6	D07 183.2	D08 241.1	D09 288.9	D10 285.1	D11 285.1	D12 -5.2	D13 258.3	D14 -4.7	D15 286.6	D16 -3.8	D17 284.2	D18 -4.7	D19 270.8	D20 -7.6	D21 244.0	D22 -4.3	D23 98.6	D24 98.1
E01 -3.8	E02 -4.7	E03 -3.8	E04 -2.4	E05 6.3	E06 -0.4	E07 -5.2	E08 -4.3	E09 -3.8	E10 -3.8	E11 -1.4	E12 -0.4	E13 -8.6	E14 -1.9	E15 2.4	E16 -1.9	E17 -0.9	E18 -4.7	E19 -21.5	E20 5.3	E21 -6.2	E22 271.3	E23 101.0	E24 102.4
F01 -3.8	F02 -8.1	F03 137.3	F04 283.7	F05 281.3	F06 287.0	F07 273.2	F08 278.9	F09 284.2	F10 283.7	F11 277.5	F12 6.7	F13 287.5	F14 -4.3	F15 284.2	F16 -3.3	F17 215.8	F18 1.0	F19 246.4	F20 -7.6	F21 283.7	F22 -3.8	F23 98.6	F24 100.5
G01 -7.1	G02 -8.6	G03 -8.1	G04 -7.1	G05 0.0	G06 0.5	G07 49.8	G08 -4.3	G09 -3.8	G10 0.0	G11 -1.9	G12 -3.3	G13 -5.2	G14 -11.0	G15 -0.4	G16 -6.2	G17 -3.3	G18 107.2	G19 -4.7	G20 -6.2	G21 -2.4	G22 163.6	G23 102.9	G24 101.0
H01 -4.3	H02 -2.4	H03 109.1	H04 289.9	H05 287.5	H06 286.6	H07 249.7	H08 281.8	H09 287.0	H10 286.1	H11 285.6	H12 9.1	H13 287.0	H14 -0.4	H15 284.2	H16 2.0	H17 199.5	H18 278.9	H19 279.4	H20 0.0	H21 278.4	H22 0.5	H23 100.0	H24 106.2
I01 131.1	I02 2.0	I03 -2.8	I04 3.4	I05 -4.7	I06 8.2	I07 0.5	I08 1.0	I09 1.5	I10 3.9	I11 4.3	I12 0.5	I13 6.3	I14 -5.7	I15 0.5	I16 -3.8	I17 6.3	I18 17.7	I19 -3.3	I20 3.9	I21 -7.6	I22 271.3	I23 97.1	I24 99.1
J01 -3.3	J02 -7.6	J03 286.6	J04 293.7	J05 286.6	J06 287.0	J07 188.5	J08 282.7	J09 287.5	J10 -5.2	J11 287.0	J12 2.0	J13 288.5	J14 1.5	J15 270.6	J16 -3.3	J17 267.9	J18 -4.3	J19 282.3	J20 2.9	J21 284.6	J22 -7.6	J23 103.8	J24 101.9
K01 1.5	K02 -5.7	K03 -4.3	K04 -9.0	K05 -9.0	K06 -3.3	K07 -8.1	K08 -13.8	K09 -3.8	K10 -3.8	K11 -1.9	K12 -11.0	K13 -3.8	K14 -19.6	K15 -5.2	K16 -8.1	K17 -1.4	K18 -6.2	K19 24.0	K20 -14.8	K21 -13.3	K22 212.9	K23 93.3	K24 98.1
L01 -0.4	L02 -0.4	L03 286.6	L04 285.1	L05 287.0	L06 288.0	L07 98.1	[illegible]	L09 286.1	L10 2.9	L11 127.8	L12 -7.1	L13 283.2	L14 -5.2	L15 289.9	L16 -4.3	L17 282.7	L18 -8.6	L19 280.3	L20 -8.6	L21 272.2	L22 -11.0	L23 98.1	L24 100.0
M01 -6.7	M02 -6.7	M03 -2.8	M04 -2.8	M05 -6.7	M06 -7.1	M07 2.0	M08 -6.7	M09 -2.4	M10 -8.6	M11 -9.0	M12 -5.2	M13 -5.7	M14 -6.2	M15 -2.8	M16 17.7	M17 -8.6	M18 -4.3	M19 83.8	M20 -4.7	M21 -11.9	M22 70.4	M23 98.1	M24 98.6
N01 -5.2	N02 -10.5	N03 285.6	N04 266.0	N05 278.9	N06 215.8	N07 277.0	N08 236.3	N09 287.5	N10 -1.9	N11 286.1	N12 12.0	N13 281.3	N14 -6.2	N15 262.2	N16 -4.3	N17 282.3	N18 -5.2	N19 288.9	N20 -3.8	N21 287.0	N22 -12.4	N23 99.1	N24 97.6
O01 -9.0	O02 -7.1	O03 -7.1	O04 -12.9	O05 -3.3	O06 -4.3	O07 -2.8	O08 230.6	O09 -3.3	O10 -6.7	O11 -4.3	O12 -4.7	O13 -2.8	O14 -0.9	O15 0.5	O16 -3.8	O17 0.0	O18 -4.3	O19 18.7	O20 -5.2	O21 -3.3	O22 153.6	O23 99.5	O24 101.9
P01 -4.3	P02 -10.5	P03 284.6	P04 276.5	P05 281.8	P06 264.6	P07 287.0	P08 284.2	P09 287.5	P10 -19.1	P11 286.1	P12 -2.8	P13 285.1	P14 3.9	P15 283.2	P16 -1.9	P17 280.8	P18 1.0	P19 281.8	P20 -9.0	P21 284.2	P22 0.0	P23 105.3	P24 104.3

Figure 4-6. Z-factor analysis reveals two examples of liquid handling errors. Note the grid-like pattern of activity.

An analysis of the time-dependence of the Z-prime and Z-factor metrics can also be informative. Despite best efforts to manage the work flow, plates may experience significantly different incubation times during the course of an HTS experiment. Plotting a quality control statistic for a plate as a function of the time taken to be measured can reveal trends in the performance of the assay. Though not as ideal, the order in which the plate was screened, as determined by its sequence in the raw data, can approximate time effects. For example, as demonstrated in Figure 4-7, the standard deviation of Z-factors for plates screened at later times tends to be larger, perhaps indicative of reagent degradation

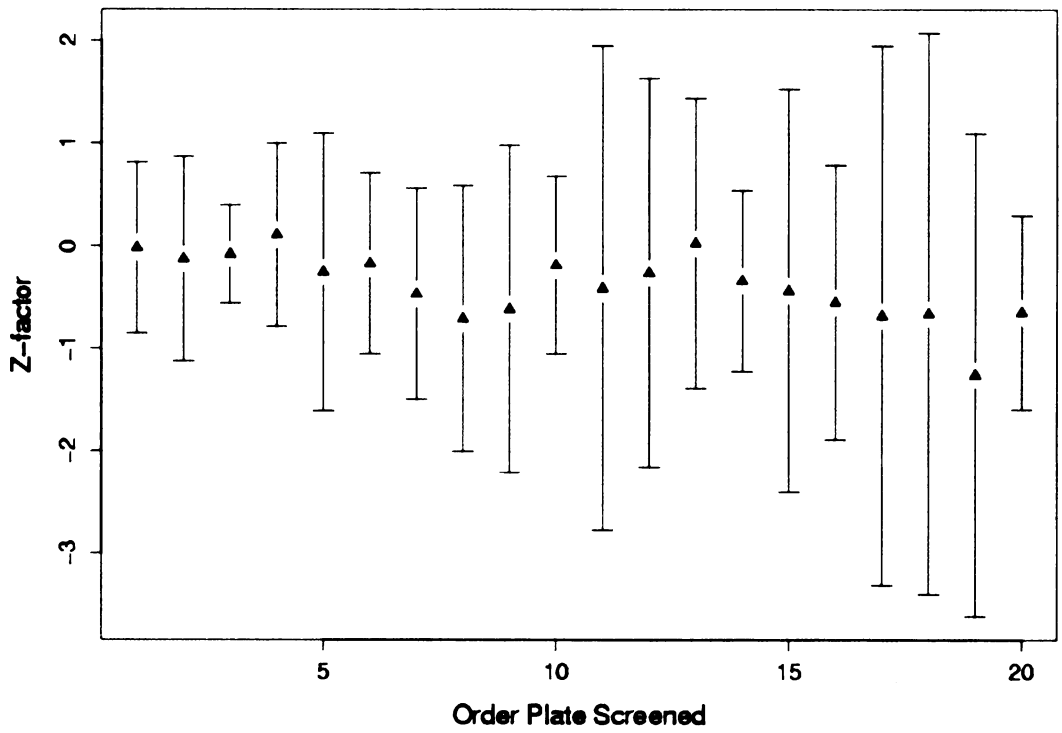


Figure 4-7. The graph of time dependent Z-factors reveals an increase in variance for plates sitting longer prior to measurement (points include 95% confidence interval).

Finally, a statistical analysis of the assay values across an entire screen for each well position can be useful for determining edge effects. Under optimal conditions, no

privileged position on the plate exists—the average variance and mean of the activity should be roughly equal at any location. However, the outer edges of a screening plate might experience different temperatures, or be located farther away from the detector, relative to central positions. Alternatively, flaws in the liquid handling device might produce deviations in the reagent conditions depending on plate geography. The signal for such errors might be masked by the random noise of the screening compounds; therefore, assay values for each well must be examined across multiple plates. Figure 4-8 demonstrates this type of well analysis. A clear positional effect is apparent in the mean well activity heat map (bottom).

Well Analysis Across Project

Each well is color coded by its variance Z-score (in Master Reference Frame)

BLK	BLK	-0.44	0.11	-0.79	0.80	0.52	-0.40	-0.36	-0.32	-0.87	0.06	1.59	0.02	-0.30	1.19	0.33	-0.50	-0.33	1.22	1.22	2.15	BLK	BLK
BLK	BLK	-0.88	-0.95	-0.41	-1.02	0.47	0.70	-0.91	2.70	-0.26	-0.45	-0.41	-0.64	-0.53	-0.28	-0.20	-0.16	0.23	0.28	1.05	2.43	BLK	BLK
BLK	BLK	0.65	-0.17	-0.63	-0.23	-0.16	-0.89	0.28	1.24	-0.78	0.03	-0.17	-0.71	-0.36	-0.87	-0.51	-0.87	-0.26	-0.54	0.57	2.19	BLK	BLK
BLK	BLK	-1.36	-0.50	-0.74	0.68	-0.69	-0.96	-0.79	-0.99	-0.23	-0.73	-0.78	-1.10	-0.95	-1.24	-0.34	-0.52	-0.66	-0.28	0.64	1.38	BLK	BLK
BLK	BLK	-0.85	0.63	-0.91	-0.53	-0.46	-0.22	-0.37	1.40	-0.33	-0.16	0.94	0.19	-0.45	0.53	-0.11	-0.88	0.48	-0.28	0.02	1.04	BLK	BLK
BLK	BLK	-1.15	-0.16	-0.91	0.81	-0.99	-0.12	-0.92	-0.67	-0.77	-1.07	-0.69	-0.94	-0.23	-1.06	-1.08	-1.08	0.24	-0.24	0.06	3.08	BLK	BLK
BLK	BLK	-1.12	-0.44	0.45	-0.59	-0.45	-0.44	1.01	2.72	-0.67	0.41	-0.29	-0.66	-0.15	0.04	-0.68	-0.61	-0.11	-0.10	0.43	1.29	BLK	BLK
BLK	BLK	-1.09	-1.06	-0.38	0.96	-0.02	-1.09	-0.63	-1.02	0.35	-0.77	-0.61	-1.00	-1.17	-1.01	-0.75	0.25	0.92	-0.20	1.93	0.62	BLK	BLK
BLK	BLK	1.08	-1.01	-0.24	-0.87	-0.75	-0.48	-0.74	-0.24	-0.77	0.11	-0.47	-0.67	-0.09	-0.72	-0.58	-0.56	-0.51	-0.32	0.44	1.72	BLK	BLK
BLK	BLK	-1.39	-0.72	-1.22	0.62	-1.16	-0.05	-1.32	-0.84	-1.08	-0.36	-1.19	-0.85	0.68	-0.84	-1.34	0.09	-0.66	0.45	1.34	2.25	BLK	BLK
BLK	BLK	-0.71	-0.59	-0.90	0.93	-0.34	-0.45	-0.59	2.25	0.24	0.66	0.30	-0.22	0.86	-0.20	-0.06	-0.55	0.54	0.18	0.76	6.45	BLK	BLK
BLK	BLK	-0.40	-0.41	-0.86	-0.30	0.71	-0.92	-0.95	-0.64	0.27	-0.45	-0.94	-0.97	-0.20	-0.86	-0.20	-0.43	0.18	0.61	2.34	2.79	BLK	BLK
BLK	BLK	-1.10	-0.87	-1.00	-0.90	-0.63	-0.66	-0.68	0.02	0.46	-0.03	-0.80	-0.73	-0.56	-0.53	-0.11	0.07	-0.12	0.32	1.29	2.73	BLK	BLK
BLK	BLK	-0.69	-1.10	-1.27	0.24	2.04	0.39	0.07	-0.53	-0.66	-0.59	-0.61	-0.71	-0.32	-0.40	-0.07	-0.16	0.47	0.86	1.25	2.94	BLK	BLK
BLK	BLK	-0.17	-0.24	-0.35	-0.37	0.20	0.01	1.01	0.01	0.91	0.40	0.44	0.46	2.58	0.38	0.55	0.28	1.25	0.89	1.54	2.77	BLK	BLK
BLK	BLK	0.52	-0.29	-0.12	2.34	0.19	1.54	2.10	0.23	0.09	-0.16	0.33	0.37	1.26	-0.45	1.12	1.06	0.72	1.23	1.83	3.10	BLK	BLK

Each well is color coded by its mean Z-score (in Master Reference Frame)

BLK	BLK	-0.73	-0.75	-0.52	-0.52	-0.18	-0.54	-0.45	-0.67	-0.45	-0.54	-0.61	-1.08	-1.10	-1.16	-1.37	-2.09	-1.79	-2.72	-2.56	-4.13	BLK	BLK
BLK	BLK	-0.23	-0.02	-0.07	-0.02	0.06	0.13	-0.06	0.20	-0.18	0.00	-0.06	-0.17	-0.47	-0.38	-0.58	-0.91	-1.11	-1.41	-1.93	-3.48	BLK	BLK
BLK	BLK	-0.21	0.45	0.39	0.60	0.50	0.49	0.68	0.62	0.65	0.33	0.45	0.21	0.07	-0.19	-0.26	-0.75	-0.77	-1.31	-1.64	-2.52	BLK	BLK
BLK	BLK	0.04	0.34	0.47	0.66	0.46	0.54	0.48	0.64	0.59	0.60	0.42	0.42	0.09	0.11	-0.22	-0.37	-0.60	-0.96	-1.46	-2.20	BLK	BLK
BLK	BLK	0.62	0.79	0.89	1.01	1.09	0.81	1.06	1.08	0.93	0.87	0.69	0.50	0.60	-0.08	0.03	-0.32	0.07	-0.96	-1.06	-2.23	BLK	BLK
BLK	BLK	0.27	0.74	0.59	1.02	0.65	0.75	0.70	0.86	0.66	0.70	0.61	0.73	0.40	0.14	-0.00	-0.12	-0.26	-0.62	-1.10	-1.87	BLK	BLK
BLK	BLK	0.82	0.86	1.40	1.22	1.26	1.19	1.39	1.51	1.22	1.29	1.14	0.84	0.67	0.50	0.34	-0.07	0.03	-0.62	-0.75	-1.87	BLK	BLK
BLK	BLK	0.54	0.83	0.94	1.05	0.89	1.11	0.92	0.98	0.99	0.85	0.75	0.66	0.59	0.44	0.27	0.32	0.08	-0.47	-1.02	-1.53	BLK	BLK
BLK	BLK	0.95	0.76	1.28	1.03	1.19	1.03	1.15	1.15	1.10	1.07	1.18	0.70	0.69	0.37	0.28	-0.13	-0.16	-0.74	-0.87	-2.06	BLK	BLK
BLK	BLK	0.50	0.77	0.80	1.04	0.82	0.91	0.82	0.92	0.90	0.83	0.63	0.71	0.80	0.38	0.11	0.07	-0.31	-0.52	-0.95	-2.07	BLK	BLK
BLK	BLK	0.85	0.78	1.17	1.25	1.14	1.21	1.28	1.37	1.09	1.10	1.06	0.86	0.96	0.31	0.41	-0.06	-0.19	-0.77	-1.09	-1.60	BLK	BLK
BLK	BLK	0.60	0.87	0.75	1.03	1.07	0.99	0.94	0.98	0.83	0.95	0.74	0.76	0.40	0.35	0.02	-0.12	-0.13	-0.55	-0.83	-1.81	BLK	BLK
BLK	BLK	0.41	0.29	0.78	0.54	0.80	0.75	0.92	0.97	0.83	0.49	0.71	0.44	0.36	-0.19	-0.10	-0.55	-0.74	-1.32	-1.44	-2.72	BLK	BLK
BLK	BLK	-0.11	0.09	0.17	0.54	0.50	0.39	0.20	0.39	0.39	0.37	0.09	0.12	-0.04	-0.14	-0.36	-0.58	-0.83	-1.35	-1.57	-2.62	BLK	BLK
BLK	BLK	0.02	-0.09	0.29	0.15	0.63	0.47	0.60	0.33	0.75	0.18	0.23	0.04	-0.25	-0.32	-0.45	-1.21	-0.73	-1.54	-2.02	-3.13	BLK	BLK
BLK	BLK	-0.85	-0.67	-0.81	-0.19	-0.59	-0.64	-0.75	-0.38	-0.63	-0.53	-1.10	-0.91	-1.21	-0.94	-1.21	-1.39	-1.58	-2.04	-2.84	-2.48	BLK	BLK

Figure 4-8. Uncovering positional effects in the HTS using well analysis.

Molecular Profile and the Identification of "Good" Hits

Examination of the Diagnostics page is the first step in the analysis of HTS data. Suspicious plates or anomalous results should be reviewed, and corrections or normalization to the raw data made accordingly. Upon assurance of satisfactory quality control, an investigator can proceed to a URL containing the "Molecular Profile" of each compound that satisfies a user-defined threshold for activity (i.e., a minimum Z score). This profile provides links to all available information, including a plate heat map and the raw data, the performance history from different HTS experiments, a preliminary SAR, a summary of similar compounds with known bioactivity (including toxicity), a summary of commercially available compounds, and ADMET model predictions. Moreover, the molecules on this page are ranked according to a user-defined rubric, such that the most suitable candidates for further development appear at the top. Specifically, compounds are first ordered by increasing number of 'PFlags', or violations of QC and "drug-likeness" criteria, and then by decreasing activity. Compounds originating from plates with poor Z-prime or Z-factor are flagged; additional criteria, such as maximum allowable activity or aberrant fluorescence behavior, can be added at the discretion of the investigator. Furthermore, the user defines a "drug-likeness" flag by setting a minimal threshold value for the 'DScore,' a metric that quantifies similarity to known bioactives and toxic molecules (see relational database section for details), commercial availability, and ADMET model performance.

Thus, the PFlag rubric is meant to emphasize the "goodness" of an HTS hit, as determined by the confidence in the assay results and "drug-likeness." The definition of "good" is intentionally left imprecise and adjustable, as the criteria for a lead candidate

might vary according to the goals of a particular drug design project. Nevertheless, the Assay Reporter provides a framework for the qualitative assessment of important molecular properties in addition to HTS activity.

Figure 4-9 below shows an example of how molecules with higher assay values were ranked lower due to PFlags. Figure 4-10 details the four components of the 'DScore:' the BioScore (a measure of the similarity to known bioactives), the ToxScore (a measure of the similarity to compounds with cytotoxic or genotoxic activity), the AvailScore (an assessment of the commercial availability of a compounds), and the ADMEScore (predictions from ADMET models). For each of these components, the Assay Reporter links to a web page that describes the results from database searches or model predictions.

Molecular Profiler (v1.2) Summary

Profile Name:
Profile Date: 17:6 11.9.2005
Molecule Count: 210
Click HERE for the BASC Reorder List
Click HERE for the Vendor Reorder List

Please contact Anang Shelat (ashelat@itsa.ucsf.edu) if you have problems with this site.

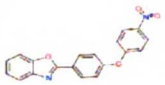
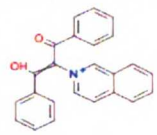
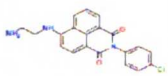
Probe	ID	Activity	RawData	PFlag	Bioactive	Toxicity	Available	ADME
 16 AB-00054474,BATCH-01	IDNM: AB-00054474 Batch: BATCH-01 LOC: 01443CD,I11 Click HERE For History	%Activ: 154.26 (2.75) ZPrime: 0.79 ZFactor: -1.16 Click HERE For SAR	FRAT: 88 FINT: 8.51783e+06 INTZ: -0.07 Click HERE For Heatmap	Failed 0 of 5 Flags	DrugScore: 10.00 NumDrugHits: 11 Click HERE For Hits	ToxScore: 3.00 NumToxHits: 3 Click HERE For Hits	AvailScore: 49.00 NumAvailHits: 100 Click HERE For Hits	ADMEScore: -0.60 ADMEViols: 2 Click HERE For Hits
 19 AB-00129026,BATCH-01	IDNM: AB-00129026 Batch: BATCH-01 LOC: 01691SF,J07 Click HERE For History	%Activ: 134.19 (2.40) ZPrime: 0.85 ZFactor: -1.27 Click HERE For SAR	FRAT: 90 FINT: 1.71553e+07 INTZ: 1.09 Click HERE For Heatmap	Failed 0 of 5 Flags	DrugScore: 0.00 NumDrugHits: 0 Click HERE For Hits	ToxScore: 0.00 NumToxHits: 0 Click HERE For Hits	AvailScore: 5.00 NumAvailHits: 7 Click HERE For Hits	ADMEScore: -0.50 ADMEViols: 1 Click HERE For Hits
	IDNM: AB-00054364	%Activ: 117.39 (2.10)	FRAT: 96		DrugScore:	ToxScore:	AvailScore:	ADMEScore:
 11 AB-00127958,BATCH-01	LOC: 01687SF,B04 Click HERE For History	ZFactor: -0.79 Click HERE For SAR	INTZ: 5.19 Click HERE For Heatmap	Int[illegible]ity Faile[illegible] of 5 Fl[illegible]	3 Click HERE For Hits	0 Click HERE For Hits	48 Click HERE For Hits	1 Click HERE For Hits
 1 AB-00053909,BATCH-01	IDNM: AB-00053909 Batch: BATCH-01 LOC: 01441CD,G07 Click HERE For History	%Activ: 337.50 (5.98) ZPrime: 0.85 ZFactor: -1.98 Click HERE For SAR	FRAT: 39 FINT: 1.78076e+09 INTZ: 17.06 Click HERE For Heatmap	Low Plate ZFactor Abnormal Activity Abnormal Intensity Failed 3 of 5 Flags	DrugScore: 10.00 NumDrugHits: 10 Click HERE For Hits	ToxScore: 0.00 NumToxHits: 0 Click HERE For Hits	AvailScore: 29.00 NumAvailHits: 57 Click HERE For Hits	ADMEScore: -0.60 ADMEViols: 2 Click HERE For Hits
 2 AB-00053907,BATCH-01	IDNM: AB-00053907 Batch: BATCH-01 LOC: 01441CD,G03 Click HERE For History	%Activ: 291.35 (5.17) ZPrime: 0.85 ZFactor: -1.98 Click HERE For SAR	FRAT: 51 FINT: 4.02605e+08 INTZ: 3.72 Click HERE For Heatmap	Low Plate ZFactor Abnormal Activity Abnormal Intensity Failed 3 of 5 Flags	DrugScore: 0.00 NumDrugHits: 0 Click HERE For Hits	ToxScore: 0.00 NumToxHits: 0 Click HERE For Hits	AvailScore: 6.00 NumAvailHits: 11 Click HERE For Hits	ADMEScore: -0.35 ADMEViols: 2 Click HERE For Hits

Figure 4-9. The PFlag rubric ranks "good" compounds higher on the molecular profile page.

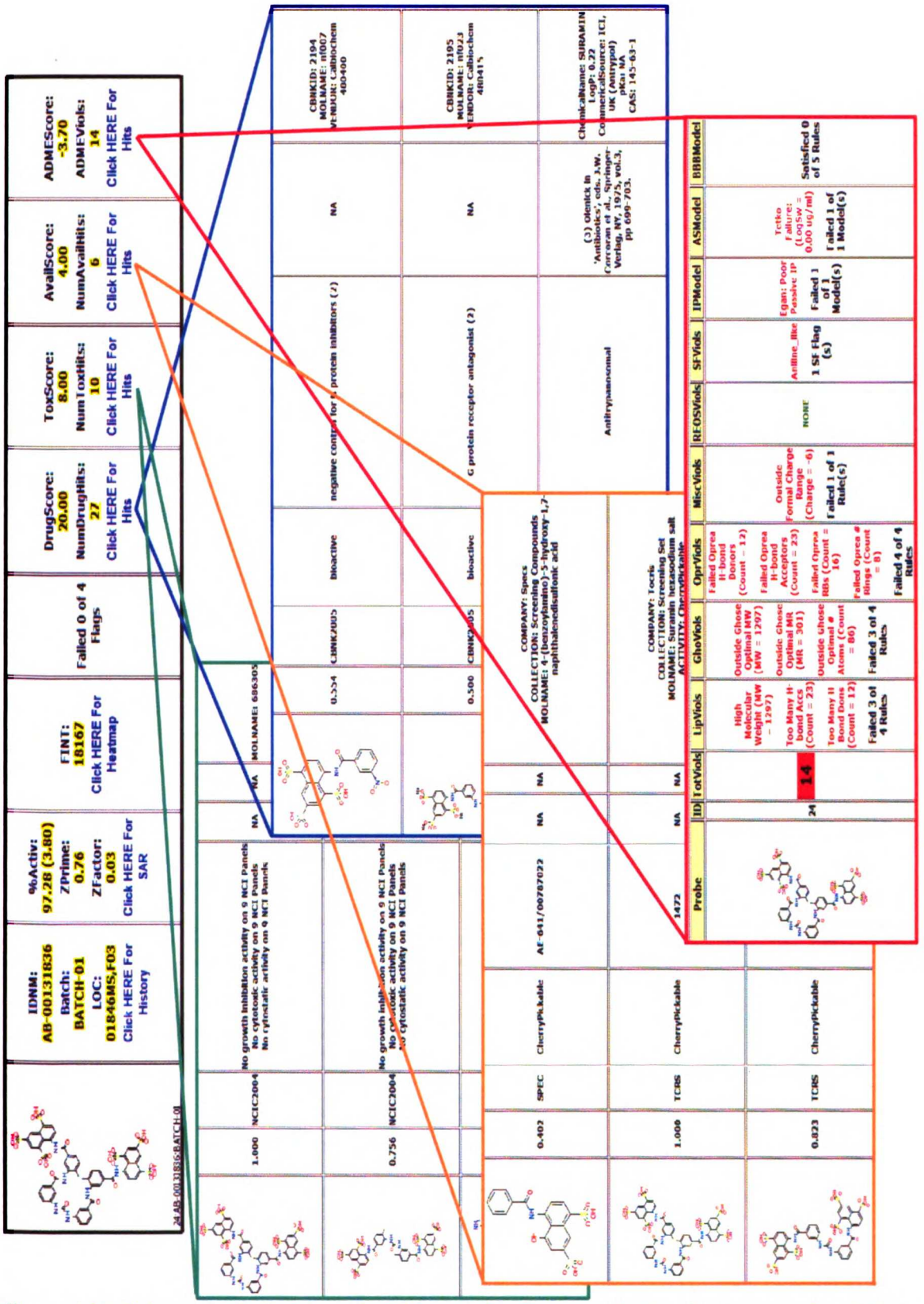
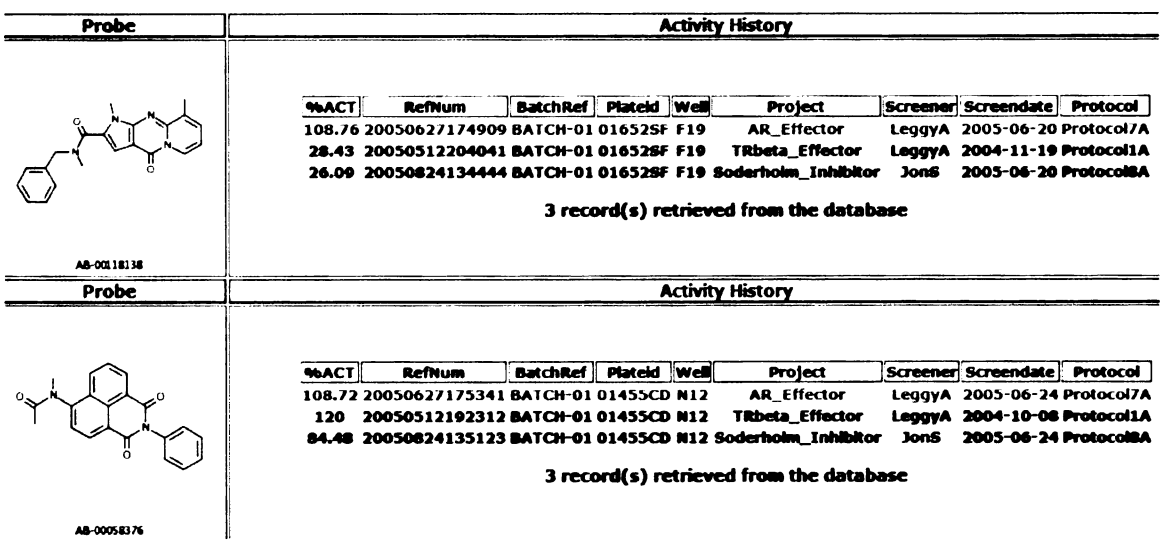


Figure 4-10. Information linked from the Molecular Profile page. Each molecular record contains links describing the similarity to known drugs (green), similarity to toxic molecules (blue), commercially availability (orange), and ADMET predictions (red)

The molecular profile page includes two additional links that are also useful for measuring "lead-like" qualities: a history of the performance of the molecule in previous HTS experiments and a preliminary SAR. These assessments were left out of the PFlag rubric because they require a more sophisticated level of interpretation that is less amenable to quantification.

For example, consider the histories presented for two molecules in Figure 4-11 (the table entry written in blue refers to the current HTS). The top molecule is selective for the Androgen Receptor (AR_Effector), whereas the bottom compound appears to be a "frequent hitter" that is active against two other unrelated targets: Thyroid Receptor Beta (TRbeta_Effector) and a GTPase (Soderholm_Inhibitor). All other things being equal, the top molecule would be considered the better hit.



Probe	Activity History

AB-00118138

%ACT	RefNum	BatchRef	PlateId	Well	Project	Screener	Screendate	Protocol
108.76	20050627174909	BATCH-01	01652SF	F19	AR_Effector	LeggyA	2005-06-20	Protocol7A
28.43	20050512204041	BATCH-01	01652SF	F19	TRbeta_Effector	LeggyA	2004-11-19	Protocol1A
26.09	20050824134444	BATCH-01	01652SF	F19	Soderholm_Inhibitor	JonS	2005-06-20	Protocol8A

3 record(s) retrieved from the database

Probe	Activity History

AB-00058376

%ACT	RefNum	BatchRef	PlateId	Well	Project	Screener	Screendate	Protocol
108.72	20050627175341	BATCH-01	01455CD	N12	AR_Effector	LeggyA	2005-06-24	Protocol7A
120	20050512192312	BATCH-01	01455CD	N12	TRbeta_Effector	LeggyA	2004-10-08	Protocol1A
84.48	20050824135123	BATCH-01	01455CD	N12	Soderholm_Inhibitor	JonS	2005-06-24	Protocol8A

3 record(s) retrieved from the database

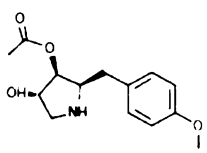
Figure 4-11. Activity histories for specific (top) and nonspecific ligands (below).

On the other hand, Figure 4-12 depicts a molecule that is active in three species of parasites: Trypansome Brucei (McKerrow_Inhibitor1), Plasmodium Falciparum W2

(Anti_Malaria1, Protocol 6A), and Plasmodium falciparum 3D7 (Anti_Malaria11, Protocol 6A). Here, the notion of a broad spectrum anti-parasitic is interesting and worth pursuing. Thus, the activity history of a molecule is another useful criterion for identifying "good" HTS hits.

Probe

Chiral



AB 00131387

Activity History

%ACT	RefNum	BatchRef	PlateId	Well	Project	Screener	Screendate	Protocol
92.85	20050630212908	BATCH-01	01839MS	C08	McKerrow_Inhibitor1	ZachM	2005-06-25	Protocol3B
107.57	20050908124157	BATCH-01	01839MS	C08	Anti_Malaria11	AllyL	2005-08-30	Protocol6B
101.01	20050908123438	BATCH-01	01839MS	C08	Anti_Malaria11	AllyL	2005-08-30	Protocol6A
99.74	20050516120310	BATCH-01	01839MS	C08	McKerrow_Inhibitor1	ZachM	2005-05-04	Protocol3A
96.21	20050630213709	BATCH-01	01839MS	C08	McKerrow_Inhibitor1	ZachM	2005-06-25	Protocol3C
12.45	20050824133237	BATCH-01	01871MS	F16	Soderholm_Inhibitor	JonS	2005-06-14	Protocol8A
0.97	20050513082440	BATCH-01	01871MS	F16	TRbeta_Effector	LeggyA	2004-11-22	Protocol1A

7 record(s) retrieved from the database

Figure 4-12. Activity history for a potential pan-parasitic.

The preliminary SAR page can be helpful for recognizing "singletons," or compounds that do not show gradual change in activity with structural variation. In general, medicinal chemists will avoid such molecules because optimization might prove difficult. In addition, the presence of only one active molecule in a series of related compounds casts doubt on the reliability of the biological data, as the assay result might be due to impurities or degradation.

However, the definition of a "singleton" is not rigorous and depends on the perception of how well the chemical structure has been explored. To aid this process, the Assay Reporter collects similar molecules for each hit on the Molecular Profile page using the Tanimoto metric and substructure searching. The latter employs the Murcko fragment, whereby the side chains of a molecule are stripped off to expose the core scaffold. The program then attempts to orient the molecules in the same direction in

space. Investigators can use the resulting congeneric series to help determine whether the HTS hit is a singleton.

Figures 4-13 and 4-14 present an example of a good SAR and a potential singleton, respectively. The colored borders around the molecules in the series correspond to activity (green=high activity, yellow=moderate activity, red=inactive).

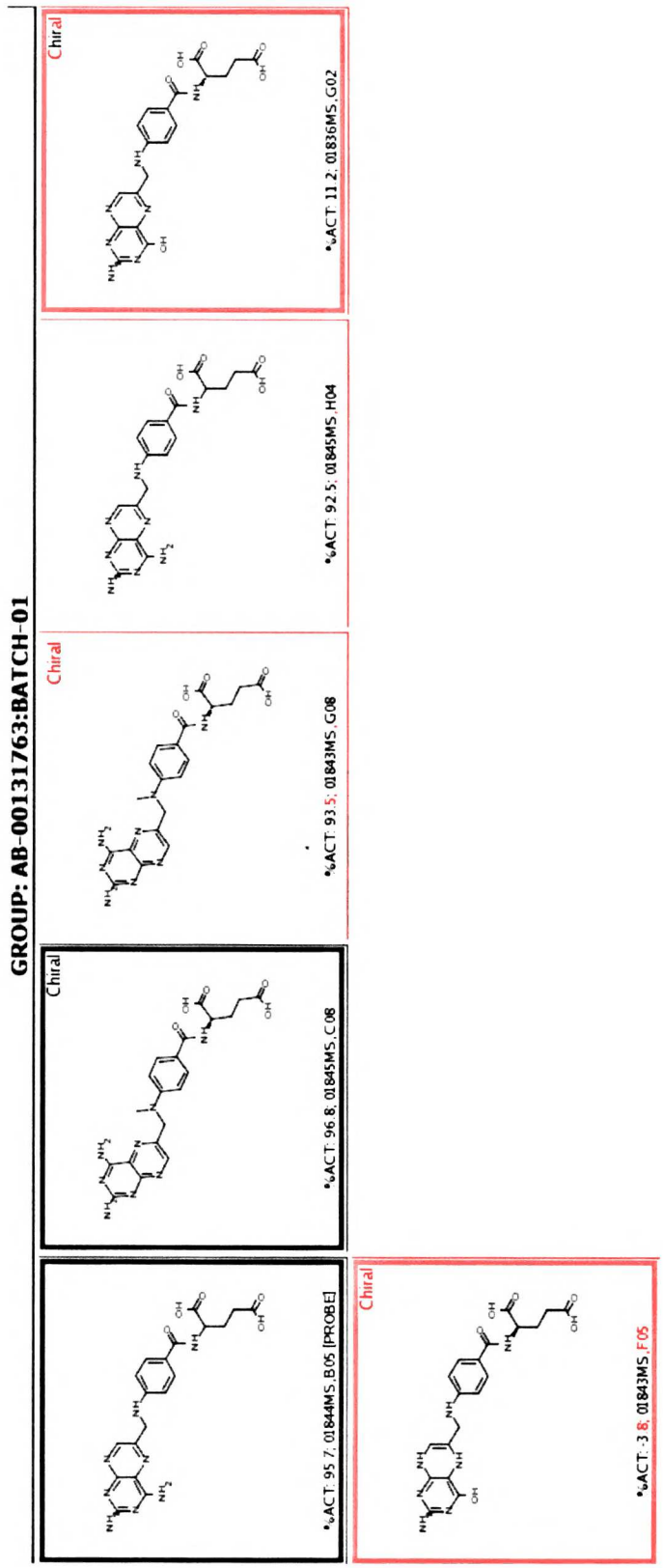


Figure 4-13. Example of a good SAR derived from the Preliminary SAR algorithm in Assay Reporter.

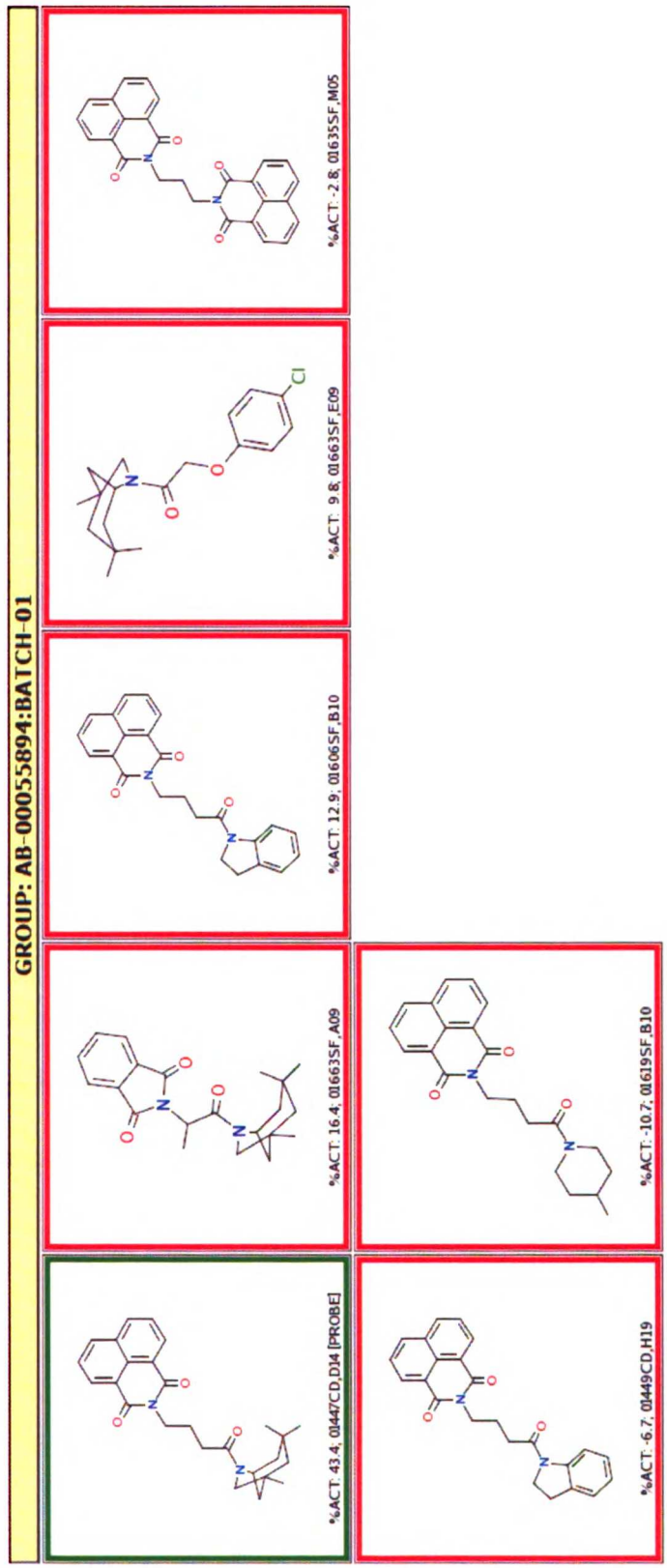


Figure 4-14. Example of a potential singleton derived from the Preliminary SAR algorithm in Assay Reporter.

4.5 Conclusion

In this chapter, we provided examples of how "good" HTS hits can be identified by combining quality control techniques and the knowledge of "drug-likeness." Though useful in the analysis of a number of projects (Arnold, 2005; Mackey, *in preparation*; Weisman, *in preparation*), the Assay Reporter is only the first step in the construction of a complete informatics system for handling HTS data. For example, the relational database must be modified to include secondary analysis data and experimentally derived molecular properties, and to track corrections and modifications to the raw data during QC assessment. More robust statistical techniques, such as replacing the mean and standard deviation with the median and median absolute deviation, should be explored. Median polishing and Fourier analysis could be useful for correcting positional effects and other frequency dependent systematic errors. Additional models, such as promiscuous inhibition, should be incorporated into the PFlags rubric of the Molecular Profile.

In the next chapter, we conclude by discussing how the Assay Reporter and the computational learning models described earlier fit together in the larger context of data mining HTS data to accelerate lead discovery.

Chapter V. Future Directions

What is the best way to discover new drugs? The most prolific drug-maker, Mother Nature, uses natural selection over thousands of years to optimize compounds. Indeed, some of the most important small molecule therapeutics used today, such as antibiotics, statins, and anti-neoplatic agents, are either modifications of natural products derived from extracts of bacteria, fungi, and other organisms or chemical analogs of human metabolites.

However, technological advances within the last two decades promised new avenues for the detection of novel chemotherapeutics. Combinatorial chemistry and HTS enabled the rapid synthesis and screening of thousands to hundreds of thousands of molecules for biological activity. Cloning and other tools from molecular biology facilitated the isolation and characterization of interesting biomolecular targets. High-speed computation and chemical informatics systems paved the way for virtual screening and structure-based drug design.

But despite the dramatic rise in the number of active molecules generated from these technologies, the rate with which new chemical entities were successfully launched during remained constant (Bleicher, 2003). A study of drug failures during the 1990s revealed that the focus on selectivity and specificity, the primary qualities emphasized during lead evaluation, needed to be balanced by a thorough assessment of ADMET properties (Kennedy, 1997). The pharmaceutical community came to the realization that drug discovery was a multi-dimensional problem requiring holistic strategies for optimization, and that translating HTS hits into *quality* lead candidates was critical to success.

This work represents an initial effort to improve the decision making process for identifying molecules from early screening hits that may potentially be good leads. We present the Assay Reporter as a framework from which to assess HTS results in the context of all available data, including predictions from computational models such as those described in earlier chapters. The PFlags rubric attempts to integrate this information into a single quantity that represents the suitability of a compound for further development.

This scoring function will be the primary focus of future research. As more chemical and biological data becomes available, we hope to refine our ability to recognize "good" lead compounds using the computational learning algorithms explored in Chapters 1 through 3. One can imagine building models of important ADMET properties using GBM and SVM technology, and using their predictions as inputs; indeed, those algorithms might eventually serve as the structure for a "meta-model" that identifies "lead-like" molecules.

Furthermore, the interpretative nature of the naïve Bayes algorithm could be harnessed to guide better optimization strategies. For example, knowledge of the favorable and unfavorable structural elements in a molecule with respect to property A will help identify flexible positions on the scaffold; these positions may be safely modified to affect other properties without disturbing property A. This technique will allow investigators to explore the tradeoffs between different properties in our scoring function, and will simplify the multivariate optimization problem by constraining the search space.

Our system will not only guide present day decisions, but will also be a source of information for retrospective analysis and an inspiration for new hypotheses. As we follow the trajectories of compounds through lead development, we can challenge the assumptions about the characteristics of quality leads. We can ask questions about the relative importance of molecular properties, when compound development should be terminated due to intractability, and how the nature of the disease for which a treatment is sought (e.g., a chronic condition requiring long-term therapy vs. an acute, short term infectious disease) changes our definition of "lead-like." With such knowledge, we can revise our scoring function and propose a new set of experiments for testing and improving our selection criteria.

By accelerating the identification of quality leads, we hope this work and developments in the future will remove some of the obstacles to modern drug discovery and uncover new possibilities for therapeutic intervention.

Works Cited

1. Ajay A, Walters WP, Murcko MA. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J Med Chem.* 1998 Aug 27;41(18):3314-24.
2. Andrews PR, Craik DJ, Martin JL. Functional group contributions to drug-receptor interactions. *J Med Chem.* 1984 Dec;27(12):1648-57.
3. Arnold LA, Estebanez-Perpina E, Togashi M, Jouravel N, Shelat A, McReynolds AC, et al. Discovery of small molecule inhibitors of the interaction of thyroid hormone receptor with transcriptional coregulators. *J Biol Chem.* 2005 Oct 31.
4. Bleicher KH, Bohm HJ, Muller K, Alanine AI. Hit and lead generation: Beyond high-throughput screening. *Nat Rev Drug Discov.* 2003 May;2(5):369-78.
5. Bobadilla JL, Macek M,Jr, Fine JP, Farrell PM. Cystic fibrosis: A worldwide analysis of CFTR mutations--correlation with incidence data and application to screening. *Hum Mutat.* 2002 Jun;19(6):575-606.
6. Brideau C, Gunter B, Pikounis B, Liaw A. Improved statistical methods for hit selection in high-throughput screening. *J Biomol Screen.* 2003 Dec;8(6):634-47.
7. Efron B, Hastie T, Johnstone I, Tibshirani R. *Least angle regression.* Stanford University: Stanford University; 2003.
8. Feng B, Shelat A, Doman T, Guy RK, Shoichet BK. High-throughput assays for promiscuous inhibitors. *Nat Chem Bio.* 2005;1(3):146.
9. Friedman JH. *Greedy function approximation: A gradient boosting machine.* Stanford University: Stanford University; 2001.

10. Galietta LJ, Haggie PM, Verkman AS. Green fluorescent protein-based halide indicators with improved chloride and iodide affinities. *FEBS Lett.* 2001 Jun 22;499(3):220-4.

11. Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J Comb Chem.* 1999 Jan;1(1):55-68.

12. Gribbon P, Sewing A. Fluorescence readouts in HTS: No gain without pain? *Drug Discov Today.* 2003 Nov 15;8(22):1035-43.

13. Hand D, Mannila H, Smyth P. *Principles of Data Mining.* Cambridge, MA: MIT Press; 2001.

14. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* New York, NY: Springer; 2001

15. Kennedy T. Managing the drug discovery/development interface. *DDT.* 1997;2(10):436.

16. Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. *Nature.* 2004 Dec 16;432(7019):855-61.

17. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods.* 2000 Jul-Aug;44(1):235-49.

18. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* 2001 Mar 1;46(1-3):3-26.

19. Loo TW, Clarke DM. Correction of defective protein kinesis of human P-glycoprotein mutants by substrates and modulators. *J Biol Chem.* 1997 Jan 10;272(2):709-12.

20. Ma T, Thiagarajah JR, Yang H, Sonawane ND, Folli C, Galietta LJ, et al. Thiazolidinone CFTR inhibitor identified by high-throughput screening blocks cholera toxin-induced intestinal fluid secretion. *J Clin Invest.* 2002 Dec;110(11):1651-8.

21. Mackey ZB, Baca AM, Fujii N, Apsel B, Shelat AA, Hansell EJ, et al. Discovery of trypanocidal compounds by whole cell HTS of live trypanosoma brucei. *In preparation.*

22. McGovern SL, Caselli E, Grigorieff N, Shoichet BK. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J Med Chem.* 2002 Apr 11;45(8):1712-22.

23. (a) McGovern SL, Helfand BT, Feng B, Shoichet BK. A specific mechanism of nonspecific inhibition. *J Med Chem.* 2003 Sep 25;46(20):4265-72. (b) McGovern SL, Shoichet BK. Kinase inhibitors: Not just for kinases anymore. *J Med Chem.* 2003 Apr 10;46(8):1478-83.

24. Muegge I. Pharmacophore features of potential drugs. *Chemistry.* 2002 May 3;8(9):1976-81.

25. O'Brien SE, de Groot MJ. Greater than the sum of its parts: Combining models for useful ADMET prediction. *J Med Chem.* 2005 Feb 24;48(4):1287-91.

26. Oprea TI. Chemical space navigation in lead discovery. *Curr Opin Chem Biol.* 2002 Jun;6(3):384-9.

27. Oprea TI. Property distribution of drug-related chemical databases. *J Comput Aided Mol Des*. 2000 Mar;14(3):251-64.

28. Ridgeway G. *Generalized boosted models: A guide to the gbm package*. ; 2005.

29. Rishton GM. Reactive compounds and in vitro false positives in HTS. *DDT*. 1997;2(9):382-4.

30. Roche O, Schneider P, Zuegge J, Guba W, Kansy M, Alanine A, et al. Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J Med Chem*. 2002 Jan 3;45(1):137-42.

31. Schultz BD, Singh AK, Devor DC, Bridges RJ. Pharmacology of CFTR chloride channel activity. *Physiol Rev*. 1999 Jan;79(1 Suppl):S109-44.

32. Seidler J, McGovern SL, Doman TN, Shoichet BK. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J Med Chem*. 2003 Oct 9;46(21):4477-86.

33. Walters WP, Murcko MA. Prediction of 'drug-likeness'. *Adv Drug Deliv Rev*. 2002 Mar 31;54(3):255-71.

34. Walters WP, Stahl MT, Murcko MA. Virtual screening -- an overview. *DDT*. 1998;3(4):161-78.

35. Wang F, Zeltwanger S, Hu S, Hwang TC. Deletion of phenylalanine 508 causes attenuated phosphorylation-dependent activation of CFTR chloride channels. *J Physiol*. 2000 May 1;524 Pt 3:637-48.

36. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *JCICS*. 1988;28:31-6.

37. Weisman JL, Liou AP, Shelat AA, Cohen FE, Guy RK, DeRisi JL. Searching for new antimalarial therapeutics amongst known drugs. *In preparation.*

38. Wermuth CG. *The Practice of Medicinal Chemistry*. 2nd ed.; 2003.

39. Yang H, Shelat AA, Guy RK, Gopinath VS, Ma T, Du K, et al. Nanomolar affinity small molecule correctors of defective delta F508-CFTR chloride channel gating. *J Biol Chem*. 2003 Sep 12;278(37):35079-85.

40. Zhang JH, Chung TD, Oldenburg KR. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen*. 1999;4(2):67-73.