Active Learning and Feature Selection in the Drug Discovery Process

Manfred K. Warmuth

October 21, 2002

Abstract

Non-technical:
In collaboration with the computational chemists at Telik, we will develop and apply novel approaches of Machine Learning to the characterization and classification of organic molecules with respect to their potential as pharmaceutical agents. In preliminary research we have already shown that our methods greatly improve the efficiency of the drug discovery cycle. In particular, we will develop search methods that identify small sets of chemical features of the compounds that are likely to be responsible for the relevant pharmaceutical properties.

Technical:
We propose to use modern Machine Learning techniques to help speed up the drug discovery cycle. Candidate compounds are represented as high-dimensional descriptor vectors. The algorithms are to decide which batch of compounds should be tested next and which features are responsible for the activity of the compounds. We use the maximum margin hyperplane separating the labeled compounds for selecting the next batch of unlabeled compounds. An alternate method based on the Voted Perceptron is more suitable for high-dimensional data. We also determine small sets of relevant features using the Maximum Entropy principle.

*Computer Science Dept., University of California, Santa Cruz, CA 94065, USA
1 Specific Aims

We focus on the following data mining problem from computer-aided drug design: Given a collection of compounds for which some pharmacologically relevant properties have been measured and translated into labels, construct a computational model that can predict the labels of untested compounds based on features of the molecules. This process involves both the construction, selection and testing of important features, the design of a computational model that can predict labels based on these features, and the application of data mining algorithms for selecting unlabeled compounds for testing.

Our goals for the one-year duration of the grant are as follows:

1. Optimize compound selection strategies for lead discovery: We will thoroughly test our selection strategies on data sets provided by Telik, Inc. These will include compounds that have been assayed in Telik's internal drug discovery efforts, compounds from the literature, and known drugs. The goal will be to identify which search methodologies are best suited to each type of compound classification problem.

2. Develop lead optimization methodologies: Lead optimization is the stage of the drug discovery process that follows the identification of a promising molecule (i.e., the drug “lead”). Compounds that are highly similar to the lead compound are synthesized and tested. The resultant data set has molecules that have very similar features and it is likely that different methods will be needed for this part of the discovery process. The goal will be to create a novel and effective strategy to guide synthesis of compounds that are similar to the lead compound.

3. Molecular descriptor validation: Molecular descriptors (the “features”) are the representations of the molecules in silico. There is a large number of different descriptor methodologies available, which we will test using machine learning techniques. The goal is to identify which molecular descriptors are best suited for the different problems in drug discovery.

4. Feature selection: Once the features/molecular descriptors have been constructed, we will attempt to identify minimal subsets of features needed to accurately model the data for particular drug discovery projects and to establish a scientific rationale for feature selection.
The goal is to design computational models that are interpretable by chemists.

5. Prospective studies: Most of the studies will be retrospective analyses of already labeled data sets. Once we have established which methodologies are best suited to lead discovery and optimization, we will apply them to active projects. The goal will be to demonstrate the efficiency of our methods in the medicinal chemistry laboratory.

2 Relevance

The pharmaceutical industry, which has a significant presence in California and plays a key role in the state’s economic health, continues to optimize the process of discovering new therapeutics for unmet medical needs.

Computational methods associated with the field of Machine Learning are playing an increasing role in the discovery of new drugs. Through this research collaboration the company will be exposed to the most recent Machine Learning techniques developed at our universities, and our students will be able to test their methodologies on industrial data sets and learn how such techniques are applied to the important problems of drug discovery.
3 Background, Significance and Preliminary Studies

We focus on the following data mining problem from computer-aided drug design: Given a collection of compounds for which some pharmacologically relevant property has been measured (the "labels"), construct a computational model based on other properties of the molecules (the "features") that can predict the labels for untested compounds.

Typically, the labels are obtained by assaying a set of small organic molecules against a therapeutic target of interest (e.g., a protein responsible for a disease) and the goal of the process is to identify which molecules should be assayed next to increase the chances of finding molecules that are active against the target.

The process is iterative, and the goal is to find compounds that bind to a target molecule in as few iterations of biochemical testing as possible. In each iteration a comparatively small batch of compounds is screened for binding activity towards this target. The labels of all compounds tested so far are used to determine a model of activity. This model is then used to choose the most promising next batch of compounds among the available pool of unlabeled compounds. The screening experiment provides a labeling of the compounds in the batch as "active" or "inactive".

The chemical compounds are represented as high-dimensional descriptor vectors. The compounds may come from different sources such as vendor catalogs, corporate collections, or combinatorial chemistry. In fact, the compounds need to exist only virtually, being defined in terms of their descriptor vectors (cf. Section 3.2.1).

![Figure 1: The Drug Discovery Cycle](image)

In the drug discovery cycle (cf. Figure 1) [MGST97] one typically starts with some initial set of already tested compounds. Then the chemists iter-
ately design/select batches of compounds for testing. Note that it is more
efficient to test multiple compounds in parallel. However, often only a small
number of chemical classes can be pursued in parallel. The idea is to refine
the model of activity in each step, based on all tested compounds at hand
and to choose the most promising compounds for the next batch. The cycle
is repeated until the ultimate goal is achieved, i.e., active compounds with
good enough properties for a clinical trial is found.

In this research we attempt to do the selection step in the cycle by
the aid of a Machine Learning algorithm. At any stage of the process,
three types of compounds can be distinguished: (a) A very small fraction
of compounds that already have been identified as active, (b) a much larger
fraction of compounds that already have been identified as inactive and
(c) by far the largest fraction of compounds that have not yet been tested (the unlabeled compounds). This situation is illustrated in Figure 2, where for the sake of simplicity the descriptors have only two com-
ponents (so we obtain a two-dimensional plot). We use Machine Learning techniques
for selecting successive batches. In prelimi-
nary work we have tested a large number of selection strategies on rather limited data
sets provided by DuPont Pharmaceuticals
[WRMaCL02, WLR+02]. We could show
that a number of our selection strategies
clearly outperform simpler strategies such as random selection and nearest
neighbor based selection.

It is important to note that our Machine Learning algorithm does its
selection based on all previous test batches. Such learning approaches are
collectively called active learning techniques [Ang88, CGJ95, SS95, CCS00].
We were able to show experimentally [WLR+02] that if we restrict the algo-
rithm not to make use of the cumulative information from previous batches,
i.e. only consider a static setup, then the performance degrades dramatically.

3.1 Selection Strategies

Probably the simplest selection strategy is to choose the next batch at ran-
dom from the unlabeled compounds. This strategy does not make use of
the labels obtained in previous iterations. The number of active hits grows
only linearly with the number of iterations. Since the number of actives
is usually quite small, the performance of the random selection strategy is poor.

Another straightforward selection strategy is to pick unlabeled compounds that are closest to previously known actives [DV98]. Different distance measures on binary descriptors are possible here. We used the total number of bits differing in the two vectors as a distance measure. An unlabeled compound receives the (negative) distance to its closest active compound as a score. The strategy then is to pick those unlabeled compounds with the highest scores (i.e. smallest distance to another active). Note that this strategy takes actives into account that are found in previous iterations. However, it searches only locally and will not find actives that are remote from the previously known actives. In our experiments (see below) the closest to previously known actives selection strategy is inferior to the more powerful methods we will describe next.

These methods are based on a linear model of activity. For this we assume that the know active and inactive compounds can be separated by a hyperplane in descriptor space. This is a mild assumption because of the high-dimensional descriptors. In the one-dimensional case (illustrated in Figure 3), this means that going from right to left there is a sequence of actives until the leftmost active is reached. Going further to the left we run into the rightmost inactive followed by all inactives to the left. In order to determine the boundary between “active” and “inactive”, it is most effective to test unlabeled compounds that are near the boundary and of distance less than the margin away from the boundary. Independent from the result of such a test, i.e. the activity of the respective compound, the area of uncertainty will be reduced by almost a factor of two. The strategy is similar to a binary search and suggests exponential convergence to the

![Figure 3: Binary search in one-dimensional case: Select unlabeled points closest to plane.](image-url)
optimal classifier.

This type of argument can be generalized to arbitrary dimensions. Among all possible hyperplanes that separate the inactive from the actives (c.f. Figure 4) we chose the one whose distance from the closest active and inactive is maximized. This is called the maximum margin hyperplane (c.f. Figure 5) and algorithms that generate this hyperplane are called support vector machines. The margin of a separating hyperplane is the minimum distance of any labeled data point to the hyperplane. The score of an unlabeled compound is the signed distance to the maximum margin hyperplane generated from all previously labeled compounds.

Given the model for activity and a scoring method, the obvious selection strategy is to select the compounds with largest positive score, since they are most likely to be active. We shall see that this is a good strategy to find many compounds in a few iterations, which is clearly one of the primary goals in drug design.

If, however, the goal is to understand the structure-activity relationship, then it is most important to rapidly improve the model. We will show that in this case the best strategy is to select examples near the decision boundary.

Figure 3 can be used to illustrate\(^1\) the difference between the largest positive and near boundary selection strategies. One of the four unlabeled compounds within the margin is the rightmost positive. It takes at most three tests (i.e. \(\geq (\log_2 4) + 1\)) to determine the leftmost positive using the near boundary selection. Hence, the model of activity is determined quickly.

\(^{1}\)It is dangerous to generalize from a one-dimensional picture to high-dimension. See [WRMaCL02] for more thorough arguments.
with the *near boundary* selection strategy. The *largest positive* strategy tests from right to left. It would take up to seven tests (in the worst case) to determine the leftmost positive instead of three; however, this procedure uncovers many active compounds already along the way.

These observations suggest that the iterative refinement of the model is an essential part of any effective selection strategy. It is also empirically supported in our experiments [WLR⁺02]. In the earlier study [WRMaCL02] we investigate a number of additional selection strategies based on other Machine Learning techniques (such as the Voted Perceptron and the Bayes Point Machine).

### 3.2 Preliminary Experiments

#### 3.2.1 Data Sets

Our experiments are based on a data set provided by Dupont Pharmaceuticals for which Thrombin was the target. This data set was also used for a recent competition, the Knowledge Discovery and Data mining Cup 2001 (cf. [http://www.cs.wisc.edu/~dpage/kddcup2001](http://www.cs.wisc.edu/~dpage/kddcup2001)). We extensively tested our algorithms in a second much larger internal data set with CDK2 as the target (provided by Dupont Pharmaceuticals). The results were similar (not shown).

The Thrombin data consists of two *rounds* of data. Round₀ is the result of an initial screen against CombiChem’s *Universal Informer Library™* (UIL) [SMB⁺97]. This is a diverse collection of compounds routinely used for target validation and initial screening. The entire UIL has been reduced here to the subset of compounds that contain a positive charge which is a known predominant feature in Thrombin actives. Additionally, a number of literature active compounds have been included in round₀. Round₁ is the result of an informative library design around five templates, based on the medicinal chemistry insight gained from the round₀ data. Thus, round₁ is already a highly enriched data set.

After removing 593 compounds that only had zero entries in all descriptor components, Round₀ consists of 1,316 compounds with 40 nominated actives. Round₁ has 634 compounds with a total of 150 actives. Each descriptor vector has 139,351 binary components. The average number of non-zero bits is 1,378 in round₀, 7,613 in round₁. The descriptors were produced by internal software tools developed at DuPont Pharmaceuticals for shape-based comparison and alignment of compounds (see [Lem00] and [PLBG02]).
3.2.2 Comparison of Selection Strategies

A selection procedure is specified by three parameters: initialization, batch size, and selection strategy. In practice it is not cost effective to test single examples at a time. In this short summary we typically chose the batch size as 5% of the total number of unlabeled compounds in the data set, which appears reasonable in comparison with typical experimental constraints. Moreover, one obtains only negligible more active hits when testing single examples in each round instead of 5% batches (result not shown, cf. [WRMaCL02]). We initialize by choosing initial 5% batches at random until at least one positive and one negative example are found. Typically this was achieved already with the first 5% batch. All subsequent batches are then chosen using the selection strategy.

In Figure 6 we plot the total fraction of hits/positives (in the test batches) for all four methods: random, closest to an active, SVM largest positive and SVM near boundary. We use log scale at X-axis because the performances in the first few batches are more important than in later batches. To provide an upper bound we also plot the number of hits of the unrealistic optimal selection strategy which chooses purely active compounds in the test batches until all active compounds have been selected. The fraction of hits of the random selection strategy grows linearly with the fraction of examples tested. Note that since we use log scale at the X-axis, the curve for random selection strategy is not a line. Closest to an active is inferior because it does a local search. SVM largest positive is closest to the optimum selection. SVM near boundary performs not as well as the SVM largest positive strategy, but is not much worse.

The results shown in Figure 6 are averages of ten runs. Each run is initialized with a different random batch. For all SVM-based selection strategies reported in this paper, we always normalize the descriptor vectors by their two-norm. This normalization consistently improves the performance (not shown).

3.2.3 Exploration vs. Exploitation

In this subsection, we compare near boundary strategy and largest positive strategy more. We plot their generalization performance and hit performance in Figure 7 when the base algorithm is SVM. We can see that the near boundary strategy is better at “exploration” (i.e., giving better generalization on the entire data set) while the largest positive strategy is better at “exploitation” (i.e., higher number of total hits). One might actually
switch between strategies at different stages of a project. In the lead discovery phase one needs to find actives quickly, whereas in lead optimization, a more refined model of activity is required. At the latter stage, chemists know already how to make actives but need to understand in detail which are the important factors for binding to the target.

We gave a simple one-dimensional motivation of our selection strategies using Figure 3. However the dimension of our data is around 139,351. Using a trick we can obtain a one-dimensional snapshot of our partially labeled data by projecting each example onto the normal direction of the current maximum margin hyperplane. Thus each example maps to a signed distance to the hyperplane. In Figure 8 we visualize the location of all examples after each 5% test batch and use different colors for the already selected and unselected examples of each label. To show the density of each type of examples along the normal direction we scatter the points within a thin stripe. Note that each stripe corresponds to a differently oriented hyperplane in the descriptor space and the hyperplane crosses each stripe at the zero-point.

The “minimum margin” can be seen as the margin of the selected examples (black in the plot) that are the closest to the center (called “support
vectors”). In Figure 8, the left plot shows the progress of the near boundary selection strategy and the right plot shows the progress of the largest positive selection strategy. During the initial batches, the minimum margin shrinks quickly and then stabilizes. The minimum margin of the left plot shrinks a little bit faster because the closest selection strategy stresses exploration. As soon as the “window” between the support vectors is cleaned (at around 50%), the label of most examples is predicted correctly. As we see in Figure 7 (left), after 50% of the examples are selected, 93% of the actives and almost all of the inactives are predicted correctly.

Figure 7: Exploitation versus Exploration: (left) Total hits performance (exploit) and true and false positives performance on the whole set (explore) and (right) ROC plots of the classifiers after the selecting the 2nd, 4th, 6th and 8th batch. X-axis is in log scale. The dashed line shows the performance of the largest positive strategy and the solid line the performance of the near boundary method. SVM is the base algorithm. We used Thrombin round1 data and a batch size of 2% (13 compounds).

Again we want to point out that we only report results on the Thrombin data set here. However, our algorithms achieved similar performance on the much larger CDK2 data set.
Figure 8: Scatter plot of the signed distance of examples to the hyperplane: (Left) near boundary selection strategy, (right) largest positive selection strategy. Each stripe shows location of points after an additional 5% batch has been labeled (using SVM). Selected examples are black, unselected actives are blue, unselected inactives are red. (Thrombin round 1)

4 Research Design and Methods

We will begin by doing retrospective studies on labeled data sets provided by Telik Inc. That is, initially we hide all labels from the selection strategies. Whenever a strategy requests a test batch, we uncover the labels for the requested compounds and proceed iteratively. As a long-term goal, we want to move away from retrospective studies and try our best methods in current drug design projects. In more detail, we plan to do the following:

1. We will thoroughly test our selection strategies on much larger data sets provided by Telik, Inc. Some of the methods will become computationally infeasible and others will have to be optimized. For example, SVM based methods are prohibitively expensive when the dimension is very high. Fortunately, we developed a simple selection strategy based on the Voted Perceptron [FS98] that requires only a couple of passes over the data to form its hypothesis whose performance is at least as good as SVM.

As discussed before, for lead discovery problems, we will pick the unlabeled compounds with the largest positive score to get more hits, for lead optimization problems, we will pick the unlabeled compounds that are near the decision boundary to quickly explore uncertain region. Since in lead optimization problems, a compound has multiple labels (e.g., activity, toxicity, solubility and permeability), we will extend our methods to these multi-label problems by using loss measures
for multiple labels, e.g. Hamming loss and ranking loss [?].

2. We have observed in our preliminary study [WRMaCL02] that the relative performance of the selection strategies varies greatly depending on the data set and the stage of the process. We are currently investigating how to combine selection strategies. Our goal is to find a master strategy whose performance is always at least as good as any of the base strategies.

In preliminary experiments we found out that the following simple master strategy works well. Assume we are to select the next batch of $s$ unlabeled compounds from a pool of a large number of unlabeled compounds. Then each base strategy gives one vote to the $s$ unlabeled compounds with the highest score. The master then choose the $s$ compounds which receive the most votes from the base strategies. We are planning to do a thoroughly comparative study of this and other master strategies.

3. There is a large number of different descriptor methodologies available and Telik Inc. has developed a very concise descriptor methodology using affinity scores of a small set (around 16) of target proteins [DV98, BVWM02] We will carefully test which type of descriptors are most helpful to the Machine Learning algorithms for finding the actives quickly. Some descriptors might complement each other while others are redundant.

In particular, if the affinity scores for one of the protein can be reliably predicted from the scores of the others by any Machine Learning algorithm, then the affinity score w.r.t. to this protein becomes redundant [DV98, BVWM02]. The goal is to find a small set of non-redundant proteins that covers the entire “chemistry space” as completely as possible.

4. Even though our selection strategies find the actives quickly, the model of activity used by our algorithms is not easily interpretable by the chemist. For examples linear function in a ($>10^6$) dimensional space are not helpful. Affinity scores based on 16 proteins are low-dimensional. However they don’t translate to specific chemical and physical properties of the target compound. Thus our most important goal is to design selection strategies that rely on few descriptor components. This way the chemist will get continuous feedback about which features are likely to be responsible for the activity. Previous work (in
the off-line setting) has shown that for the Thrombin data set, about 40 descriptor components are relevant for the discrimination of actives from inactives [WPCB+02]. Thus it is desirable to have a method at hand that simultaneously improves the classifier for the purpose of iteratively selecting good test batches and at the same time does this based on a small number of descriptor components.

In a preliminary study we use the following novel feature selection technique. We maintain a distribution on the labeled examples and iteratively choose a single feature (out of the $10^6$) which has the smallest expected error on the labeled examples. We then update our distribution on the examples such that all previously selected features have expected error exactly 50% and otherwise the entropy of the weighting distribution is maximized. This method is related to Boosting [KW99, FS96]. We stop after selecting about 50 features and then use the resulting weighting on the features as to score the unlabeled examples for the purpose of selection.

The resulting selection strategies work at least as well as the SVM based methods for finding actives quickly and we are eager to work with a team of chemists to check whether the features our method is finding are useful for exploring chemistry space.

5. So far we only have done retrospective analysis of already labeled data sets. However, our long-range goal is to provide software that will be used by computational chemists in the laboratories in the actual search for new drugs.
5 Timetable and Milestones

The grant is for only one year. Clearly we will begin by applying all of our algorithms to the new in-house data sets provided by Telik Inc. and then proceed to the more ambitious goals. We have preliminary results for all the proposed research including the feature selection component and we are already actively collaborating the researcher from Telik Inc. The timetable and milestones are as follows:

- 02/01/2003: Initiation of research plan
- 03/01/2003: Comparative study of all selection strategies we investigated on the new data sets provide by Telik Inc. Study of combining the selection algorithms in a master strategy. Developing selection strategies for very large data sets such as those based on the Voting Perceptron Algorithm.
- 04/01/2003: Identification of effective feature selection algorithm based on the Maximum Entropy Principle
- 05/01/2003: Identification of novel, effective molecular descriptor for lead discovery
- 07/01/2003: Completion of in-silico lead optimization tool
- 08/01/2003: Milestone 1: Manuscript #1 submitted for publication
- 09/01/2003: Initiation of prospective compound selection and synthesis in an active therapeutic project
- 02/01/2004: Milestone 2: Manuscript #2 submitted for publication
6 Literature cited

References


