ELSEVIER

Kolmogorov-Smirnov statistic and its application in library design

Dmitrii N. Rassokhin and Dimitris K. Agrafiotis

3-Dimensional Pharmaceuticals, Inc., Exton, Pennsylvania, USA

After several years of frantic development, the dream of an "ideal" library remains elusive. Traditionally, combinatorial chemistry has been used primarily for lead generation, and molecular diversity has been the method of choice for designing and prioritizing experiments. One aspect that often has been overlooked is the drug likeness of the resulting collections. Recently, there have been several attempts to quantify this concept and incorporate it directly into the design process. This article demonstrates the limitations of some conventional methodologies and proposes a new paradigm for experimental design based on the principles of multiobjective optimization. This method allows traditional design objectives such as diversity or similarity to be combined with secondary selection criteria in order to bias the selection toward more pharmacologically relevant regions of chemical space. The method is robust, general, and easily extensible, and it allows the medicinal chemist to create designs that represent the best compromise between several, often conflicting, objectives. Two types of designs are discussed (singles, arrays), and a novel criterion based on the Kolmogorov-Smirnov statistic is proposed as a means to enforce a particular distribution on key molecular properties that are related to drug likeness. The potential of this approach is illustrated in the design of an exploratory library based on the simultaneous optimization of five different parameters. These parameters are combined in an intuitive manner to produce a design that is sufficiently diverse, exhibits a molecular weight and logP profile that is consistent with the respective distributions of known drugs, requires a small number of reagents, and can be synthesized easily in array format using robotic hardware. © *2000 by Elsevier Science Inc.*

Keywords: data mining, multiobjective optimization, simulated annealing, synchronous annealing, Komogorov-Smirnov, principal component analysis, nonlinear mapping, molecular descriptor, combinatorial chemistry, combinatorial library, high-throughput screening, molecular diversity, molecular similarity

Corresponding author: Dimitris K. Agrafiotis, 3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Exton, PA 19341, USA. Tel.: 610-458-6045; fax: 610-458-8249. *E-mail address*: dimitris@3dp.com (D.K. Agrafiotis).

INTRODUCTION

In 1997, Lipinski et al.[1] published an influential article describing a simple set of heuristic rules for determining the solubility and permeability of compounds being considered for high-throughput screening and structure-activity relationship development. This analysis was based on the hypothesis that compounds with poor physicochemical properties, particularly poor solubility and permeability, usually are detected during preclinical and Phase I safety evaluation, and are not likely to advance to the considerably more expensive Phase II efficacy studies. The study was based on a data set comprised of 2,287 compounds from the World Drug Index whose annotation suggested that they were clinically exposed, were not polymers or peptides, and did not contain the fragment O═P—O. The properties of interest included the molecular weight, computed logP, and number of hydrogen bond donors and acceptors as determined by the number of OH and NH groups and the number of nitrogen and oxygen atoms, respectively. His analysis resulted in the well-known "rule of 5" which states that for compounds that are not substrates of biological transporters, poor absorption and permeation are more likely to occur when there are more than 5 H-bond donors, more than 10 H-bond acceptors, the molecular weight is >500, or logP is >5.

However, in the last decade, combinatorial chemistry and high-throughput screening have changed the physicochemical profile of the compounds that emerge as new drug leads. In the years preceding high-throughput screening, most lead compounds had already undergone considerable scrutiny before being identified as such, and they possessed physical properties that were consistent with established preferences of known orally active agents. This situation changed dramatically with the advent of high-throughput screening, which required a shift from the traditional method of drug solubilization. For a variety of reasons that are related to the presence of potentially interfering assay components and the fact that compounds are dissolved in dimethylsulfoxide, the "apparent" solubility of compounds tested in an *in vitro* high-throughput screen is significantly higher than the true aqueous thermodynamic solubility determined under equilibrating conditions. This problem has been compounded by

Journal of Molecular Graphics and Modelling 18, 368–382, 2000
© 2000 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00063-2

the increasing reliance on combinatorial chemistry as a source of compounds for mass screening. Most combinatorial chemistry efforts are aimed at producing exploratory or universal libraries, which are target independent and are designed to span a wide range of physicochemical and structural characteristics. Unfortunately, the emphasis on quantity and the poorly guided quest for molecular diversity often have resulted in the design of combinatorial libraries with ranges of properties that are expected to result in poor pharmacokinetic profiles. Consequently, most hits identified from such libraries reflect the physicochemical characteristics of their respective collections and represent suboptimal starting points for SAR development. Although many chemists believe that undesirable properties can be eliminated through classic medicinal chemistry approaches, experience suggests that correcting solubility and permeability remains the rate-limiting step in the development of a drug candidate.

Instead of using medicinal chemistry as a substitute for good experimental design, it would be far more desirable to incorporate the principles outlined earlier directly into the library design process. This can be done in two ways. The first is to apply a "hard" filter and eliminate synthons that lead to products that fall beyond some predetermined allowed property range. A straightforward implementation of the Lipinski "rule of 5" would represent an example of this strategy. However, given the probabilistic nature of the problem, perhaps a more appropriate approach would be to focus on the actual *distributions* of the properties that are related to drug likeness.

Recently, several attempts have been made to bias the design toward the physicochemical properties of orally active drugs. Martin et al.[2] presented a reagent selection algorithm based on D-optimal design, wherein the candidate reagents were assigned to categorical bins according to their properties, and successive steps of D-optimal design were performed to generate diverse substituent sets consistent with required membership quotas from each bin. This technique later was elaborated,[3,4] and a new "parallel" sampling approach was proposed in order to eliminate the order dependence of the original algorithm. Most recently, Koehler et al.[5] proposed a multipass algorithm designed to facilitate addition of compounds to an existing chemical library. This method was intended to prioritize compounds that are most similar to a specified set of favorable target molecules and, at the same time, most dissimilar to the compounds that reside in the library being augmented. The algorithm scores and ranks each compound in the external library according to the local density of similar compounds in the target and internal libraries. Density arising from the target library adds a positive contribution to the score, whereas density arising from the internal library adds a negative contribution. The highest-ranking compounds are included in the internal library and the process is repeated until the specified number of compounds is selected.

Both of these algorithms are specific to the problem at hand and suffer from lack of generality. In this article, we present a multiobjective approach to library design that circumvents many of the problems associated with conventional compound selection techniques. Our method allows traditional design objectives such as diversity or similarity to be combined with secondary selection criteria in order to bias the design toward more pharmacologically relevant regions of chemical space. The general algorithm is outlined, and a new selection criterion aimed at producing designs that obey a particular property distribution is described. The advantages of our approach are demonstrated using a two-component combinatorial library based on the reductive amination reaction, and a subset of compounds from the World Drug Index generously provided to us in electronic format by Dr. Christopher Lipinski of Pfizer, Inc.[6]

We should note that multiobjective library design is not a novel idea. This concept was introduced by our group in 1995[7] and was elaborated further in 1996,[8] in 1997,[9] and in several other subsequent publications.[10–13] These publications laid out the algorithmic foundations of this approach, demonstrated its potential as a general and extensible system for library design,* and outlined several options for search and optimization, including evolutionary programming and genetic algorithms.† This general paradigm was adopted subsequently by several groups. Two years later, Brown and Martin[14] employed a similar genetic scheme to generate libraries designed to minimize the effort required to deconvolute biological hits by mass spectroscopic techniques. Their approach was based on a fitness function that combined size, diversity, molecular weight, and formula redundancy. A year later, Gillet et al.[15] used a similar genetic approach with an objective function that included an additional term based on the root mean square (RMS) difference between the binned property distribution of the compounds in the library and a user-prescribed reference distribution. Finally, one of the reviewers brought to our attention a recent paper by Hassan and Waldman,[16] presented at the 218th ACS National Meeting, whose abstract appears related to the subject matter.

METHODS

Compound Selection

In its simplest form, the selection problem can be stated as follows: given an n-member virtual library and a number k, find the "best" set of k compounds in that population. The

*Citing from Agrafiotis[8]: "Because the search engine and the performance metric are treated as independent entities, this technique can be readily extended to perform selections based on any suitably encoded, user-defined selection criterion. Indeed, many existing diversity algorithms such as maxmin or d-optimal design can be recast in the form of an optimization problem by devising an appropriate objective function. Furthermore, other important criteria can be directly incorporated into the penalty function, such as cost of starting materials and/or compliance to existing SAR or pharmacophore models, and can be combined to perform complex multiobjective selections in advanced decision support systems."

†Citing from Agrafiotis et al.[7]: "In a preferred embodiment of the present invention, in step 608 each subset of candidate compounds is represented as a binary string which uniquely encodes the number and indices of the candidate compounds comprising the subset. A population of binary encoded subsets is then initialized by a random process, and allowed to evolve through repeated application of genetic operators, such as crossover, mutation and selection. Selection is based on the relative fitness of the subsets, as measured by their ability to satisfy requirements (1), (2), and (3) discussed above. Upon completion, the present invention yields a population of subsets, ranked according to their ability to satisfy requirements (1), (2) and (3). The highest ranking set is then processed in accordance with step 610." These requirements are described earlier in that document: "As represented by step 608, the Synthesis Protocol Generator 104 ranks the candidate compounds identified in step 606, individually or in combination, according to their predicted ability to (1) exhibit improved activity/properties, (2) test the validity of the current structure-activity models, and/or (3) discriminate between the various structure activity models. The candidate compounds may also be ranked according to their predicted three-dimensional receptor fit."

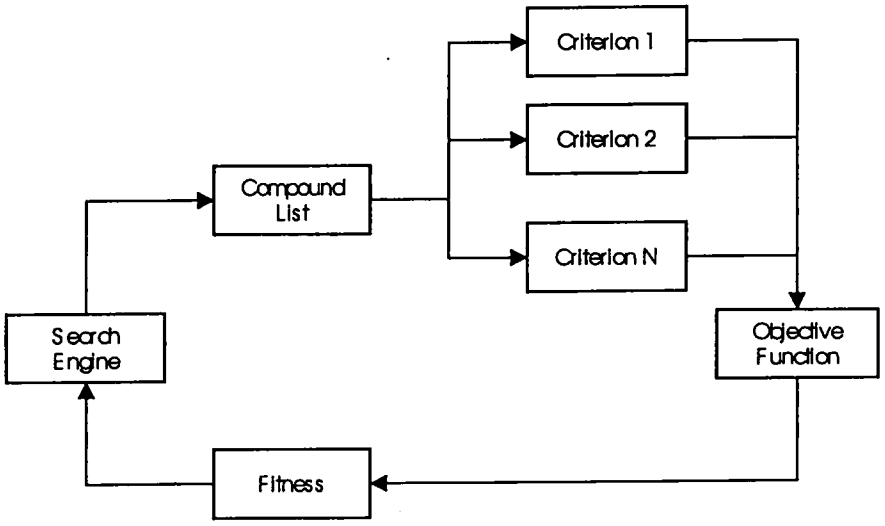

Figure 1. Process flow of the multiobjective selector.

approach taken here is to use a multiobjective fitness function to evaluate the quality of a design and use a stochastic search engine, such as simulated annealing, evolutionary programming, or a genetic algorithm, to maximize that function and identify the optimal (or a nearly optimal) set.[8,9,10,13]. The method is illustrated in Figure 1.

In its prototypical form, the search engine produces a list of k compounds from the virtual collection (also referred to as a *state*), which subsequently is evaluated against a prescribed set of fitness criteria, f_1, f_2, ..., f_C, that encode specific design objectives such as the intrinsic diversity of the compounds or their similarity to a predefined set of leads. Each criterion returns a value that measures how well a particular set of compounds satisfies the underlying objective. These individual fitness values then are combined into a unifying objective function, $f^* = f(f_1, f_2, \ldots, f_C)$, which provides the overall performance measure for that particular set. The function f* can assume any functional form. Its value is fed back to the search engine, which modifies the state in a controlled manner and produces a new list of compounds, which is, in turn, evaluated against the selection criteria in the manner described earlier. This process is repeated until no further improvement is possible, or until some predetermined convergence criterion or time limit is met.

The major advantage of this approach is that the search algorithm is completely independent of the performance measure and can be applied on a wide variety of selection criteria and fitness functions.[8,9,10,13] Unlike alternative algorithms such as maxmin,[17] cluster analysis,[18] binning,[19] D-optimal design,[20] and stepwise elimination,[21], which are tailored to a particular application, this approach is completely general, programmatically simple, and easily extensible.

In the present work, the optimization was carried out using a parallel implementation of simulated annealing. As in conventional annealing, the process starts with a random initial state and walks through the state space by a series of small stochastic steps. In the problem at hand, these steps represent a small change in the composition of the set (i.e., replacement of a small fraction of the points comprising the set). The objective function, f*, maps each state to a real value that represents its energy or fitness. Whereas downhill transitions are always accepted, uphill transitions are accepted with a probability that is inversely proportional to the energy difference between the two states. This probability is controlled by a parameter called temperature, which is adjusted in a systematic manner during the course of the simulation. In our parallel implementation (known as synchronous annealing), each execution thread is allowed to follow its own independent Monte Carlo trajectory during each temperature cycle. The threads synchronize at the end of each cycle, and the best among the last states visited by each thread is recorded and used as the starting point for the next iteration. Given sufficient simulation time, this parallel algorithm produces results that are comparable to those obtained with the traditional serial implementation.

In this work, two different kinds of designs are considered. The first is called "singles" and refers to a subset of products

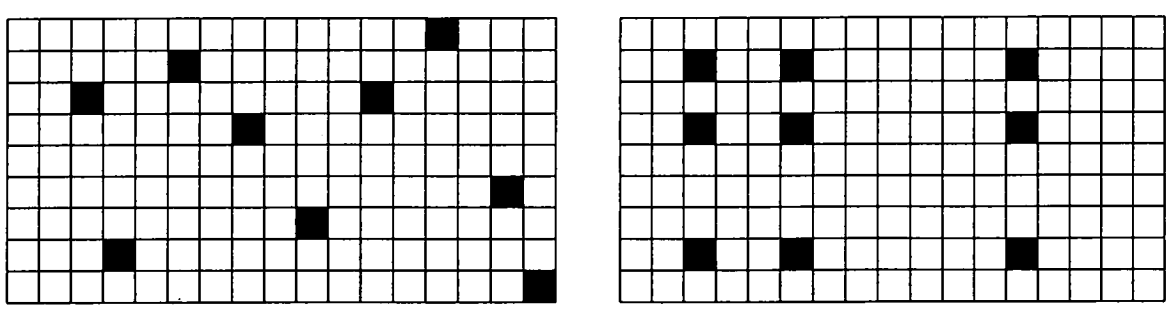

Figure 2. Selection strategies for combinatorial library design. (a) Singles; (b) arrays.

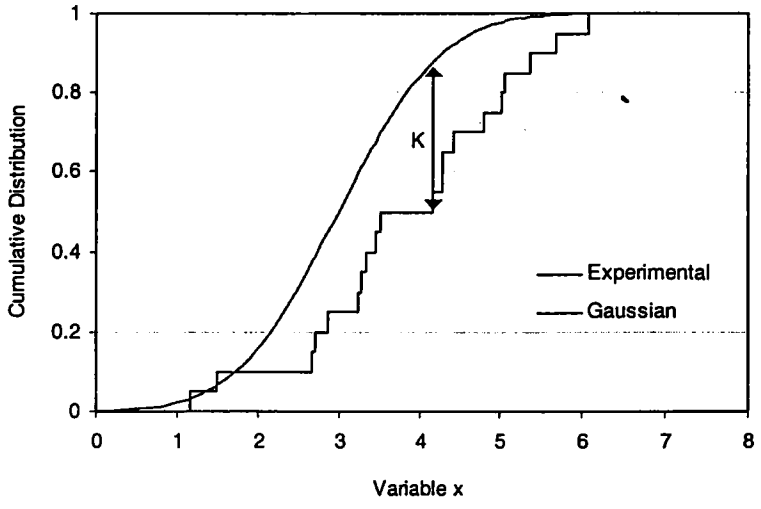

Figure 3. The Kolmogorov-Smirnov statistic K.*

that is not constrained by the number or types of reagents involved. The second is called an "array" and represents the products derived by combining a given subset of reagents in all possible combinations as prescribed by the reaction scheme. These two types of designs are illustrated in Figure 2.

The combinatorial nature of the two problems is vastly different. For singles, the number of states that one has to consider (the number of different k-subsets of an n-set) is given by the binomial:

$$C_s = \frac{n!}{(n-k)!k!}. \quad (1)$$

In contrast, the number of different $k_1 \times k_2 \times \ldots k_R$ arrays derived from an $n_1 \times n_2 \times \ldots n_R$ R-component combinatorial library is given by:

$$C_a = \prod_{i=1}^{R} \frac{n_i!}{(n_i - k_i)!k_i!}. \quad (2)$$

For a 10×10 two-component combinatorial library, there are 10^{25} different subsets of 25 compounds, and only 63,504 different 5×5 arrays. For a 100×100 library and a 100 (or 10×10) selection, those numbers increase to 10^{241} and 10^{26} for singles and arrays, respectively. In the context of simulated annealing, these two types of designs are encoded using two different internal state representations, each with its own mutation protocol. In a singles selection, a step represents the substitution of a few compounds comprising the current state, whereas in an array selection a step represents the substitution of a single reagent in the combinatorial array. Note that, in this context, the term "array" is basically equivalent to reagent selection at the product level[20] and does not refer to the physical layout and execution of the experiment. Although as we demonstrate later arrays are generally inferior in terms of meeting the design objectives, they require fewer reagents and are much easier to synthesize.

Two different selection criteria were used in this study: molecular diversity and the Kolmogorov-Smirnov statistic. These are defined as follows.

Diversity Criterion

The diversity of a set of compounds, C, is defined as the average nearest neighbor distance:

$$D(C) = \frac{1}{N}\sum_{i} \min_{j \neq i} (d_{ij}), \quad (3)$$

where N is the cardinality of C, and d_{ij} is the Euclidean distance between the i-th and j-th compounds in some molecular descriptor space. Because the value of this criterion increases with spread, diverse libraries are obtained by maximizing D. We have found that this function is smoother than the more commonly used maximum minimum dissimilarity and can discriminate more effectively between the various ensembles, particularly when used as the objective function in a Monte Carlo optimization procedure such as simulated annealing or evolutionary programming. Naively implemented, Equation 3 requires N(N−1)/2 distance computations and scales adversely with the number of compounds selected. To reduce the quadratic complexity of the problem, D is computed using the k-d tree algorithm presented in Agrafiotis and Lobanov.[13] This algorithm achieves computational efficiency by first organizing all the points in C in a k-dimensional tree, and then performing a nearest neighbor search for each point using a branch-and-bound approach. For a small number of dimensions, this algorithm exhibits N logN time complexity and scales favorably with the number of compounds selected.

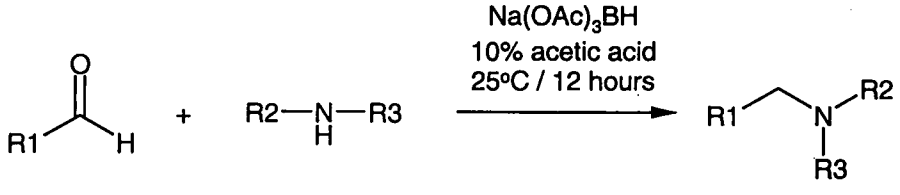

Figure 4. Synthetic sequence for the generation of the reductive amination library.

Kolmogorov-Smirnov Criterion

The Kolmogorov-Smirnov criterion measures how well an experimental distribution is approximated by a particular distribution function.[22] It is applicable to unbinned distributions that are functions of a single independent variable and is defined as the maximum value of the absolute difference between two cumulative distribution functions:

$$K^* = \max_{-\infty<x<\infty} |P(x) - P^*(x)|, \quad (4)$$

where $P(x)$ is an estimator of the cumulative distribution function of the actual probability distribution from which it is drawn, and $P^*(x)$ is a known cumulative distribution function. For a set of N points x_i, i = 1, ..., N, $P(x)$ represents the fraction of data points to the left of a

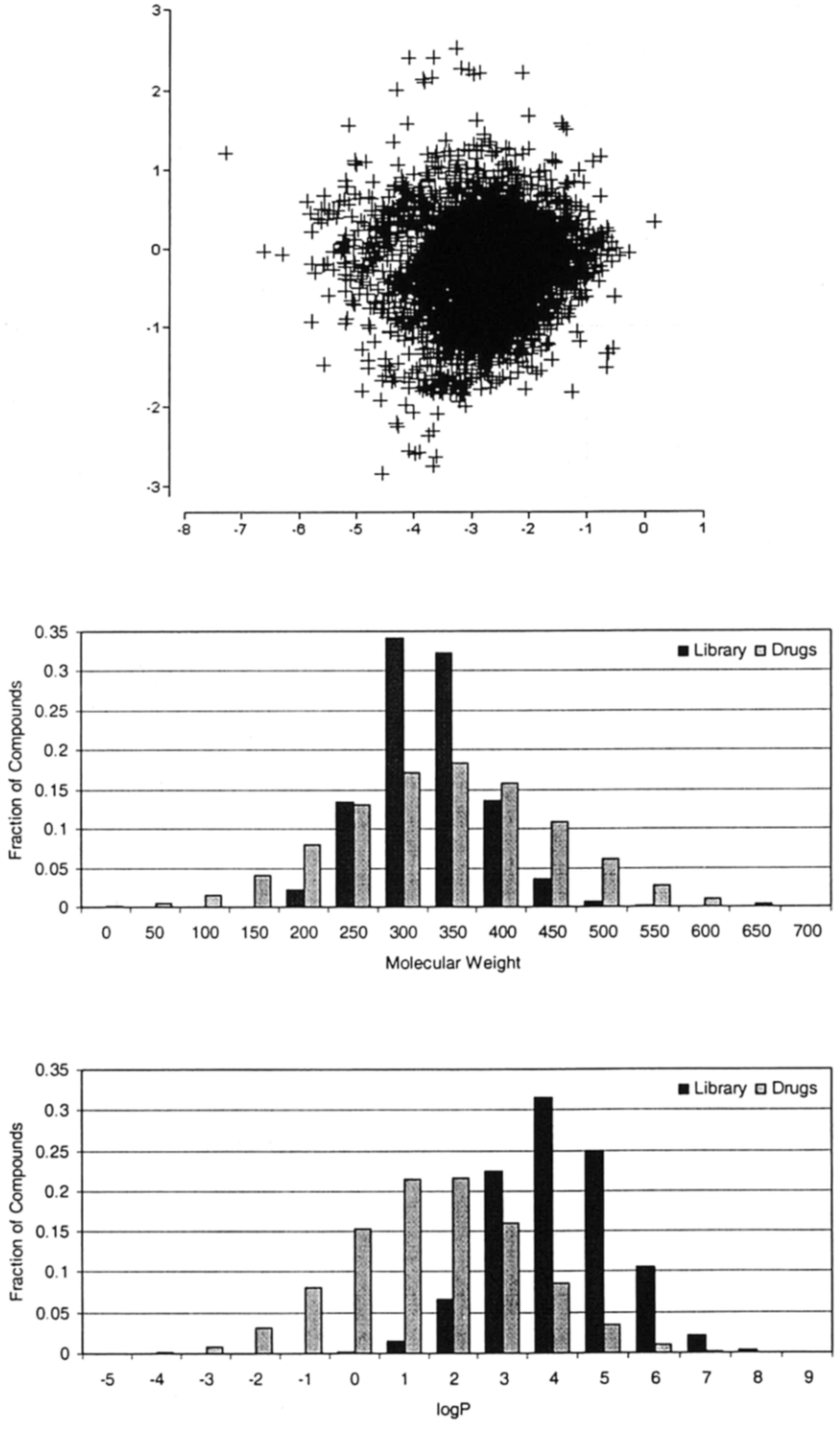

Figure 5. The reductive amination data set. (a) Nonlinear projection; (b) molecular weight distribution; (c) logP distribution.

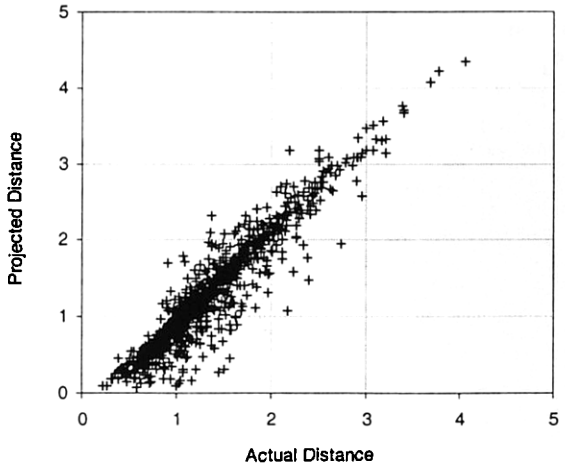

Figure 6. Actual vs *projected distances for 5,000 randomly chosen pairs from the reductive amination library. For a perfect nonlinear projection, all points should be located along the diagonal.*

given value x (inclusive). The method is illustrated in Figure 3.

Unlike the more commonly used χ^2 test, the Kolmogorov-Smirnov statistic does not require binning of the data, which is arbitrary and leads to loss of information. More importantly, the function is very fast to compute because it involves sorting the data in ascending order, followed by a linear scan to identify the maximum difference from the user-defined cumulative distribution function. Speed of computation is particularly important in the application at hand, where the fitness function needs to be evaluated tens of thousands of times in the course of the optimization.

The significance level of a particular value of K* as a disproof of the hypothesis that two distributions are the same is a function of K* and the number of data points, N. This function is relatively slow to compute, but when N is constant, it is a monotonic function of K*. Because all we want is to determine which experimental distribution is closer to the "ideal" distribution P*(x), the significance level need not be computed.

The Kolmogorov-Smirnov criterion as defined by Equation 4 is a measure of dissimilarity and takes values in the interval [0,1]. Alternatively, we can define the similarity between two probability distributions, K:

$$K = 1 - K^*. \tag{5}$$

Thus, designs that obey a particular distribution function are obtained by maximizing K.

Software

All programs were implemented in the C++ programming language and are part of the DirectedDiversity® software

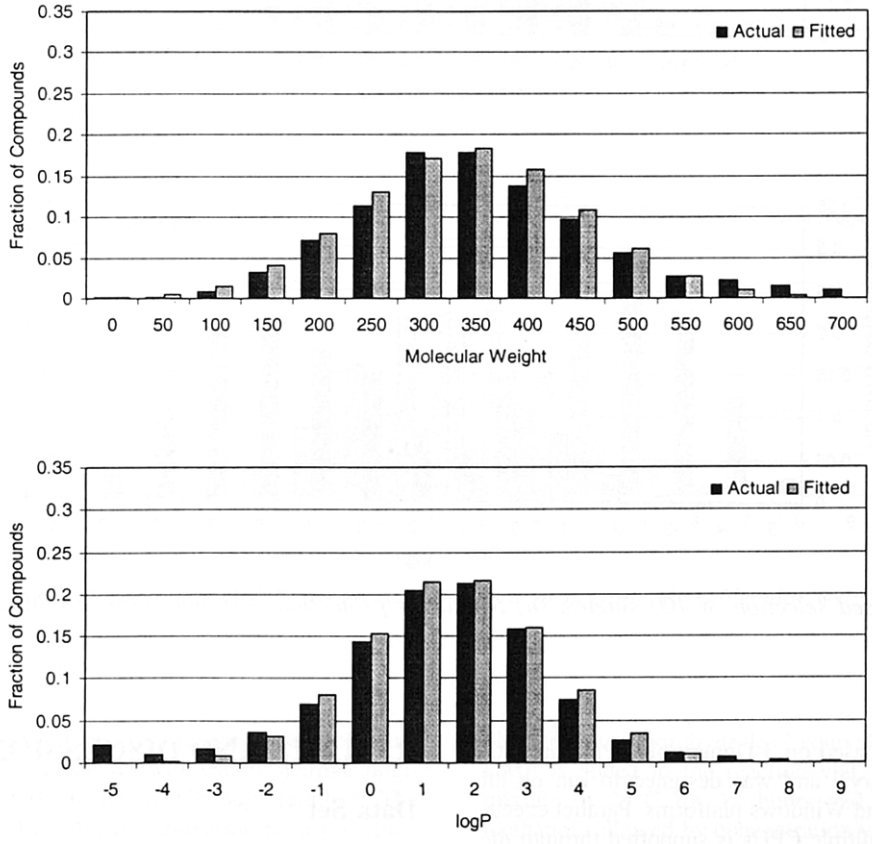

Figure 7. (a) Molecular weight and (b) logP distributions of drug-like molecules based on 7,484 marketed drugs from the World Drug Index. Two series are shown for each property. The one on the left is the true distribution; the one on the right is a normal approximation determined by a least-squares fitting procedure.

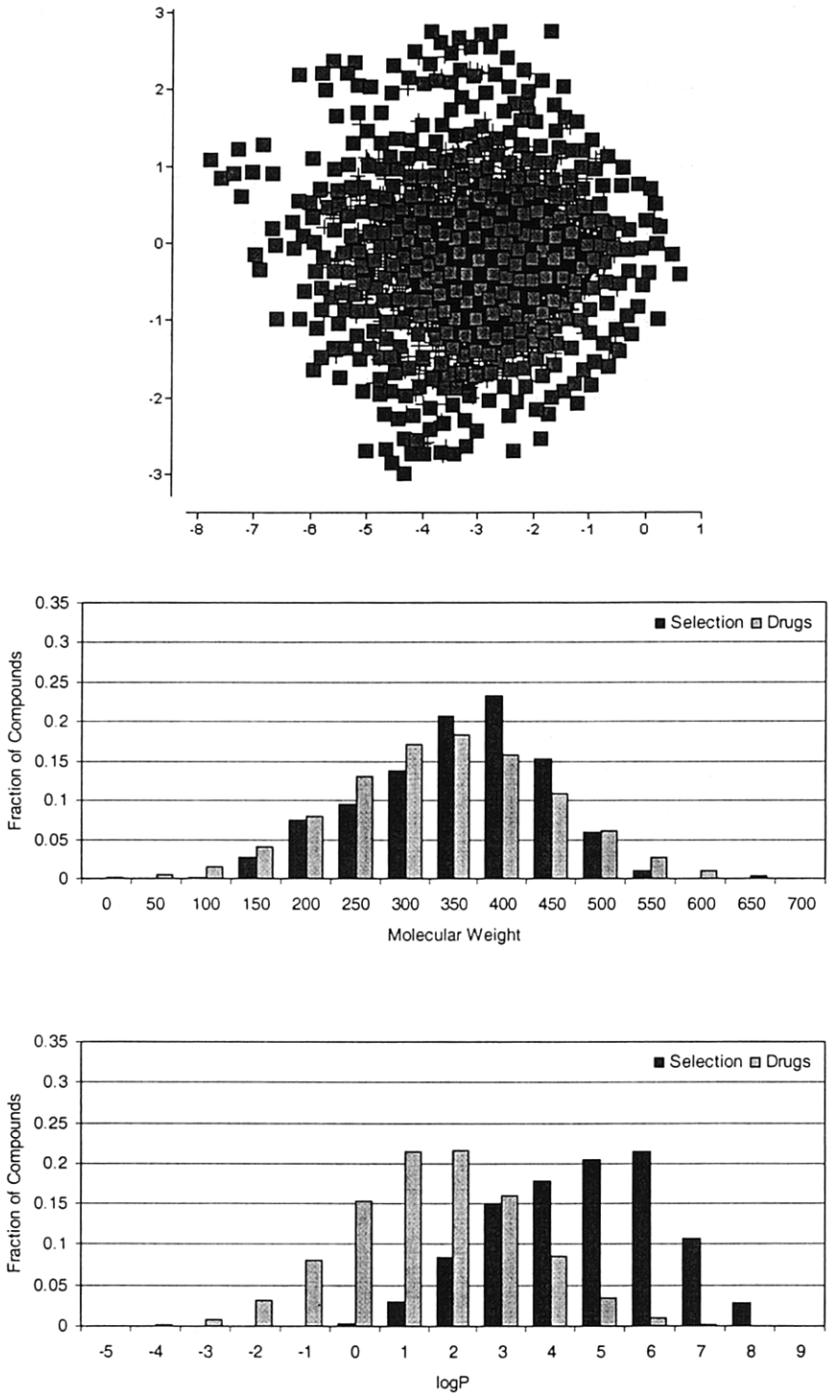

Figure 8. Diversity-based selection of 400 singles. (a) Nonlinear projection; (b) molecular weight distribution; (c) logP distribution.

suite.[7] The software is based on 3-Dimensional Pharmaceuticals' Mt++ class library[23] and was designed to run on all Posix-compliant Unix and Windows platforms. Parallel execution on systems with multiple CPUs is supported through the multithreading classes of Mt++. All calculations were carried out on a Dell Dimension workstation equipped with two 400-MHz Pentium II Intel processors running Windows NT 4.0.

RESULTS AND DISCUSSION

Data Set

The data set used in our study is a two-component combinatorial library based on the reductive amination reaction. This library is part of a synthetic strategy that exploits the pivotal

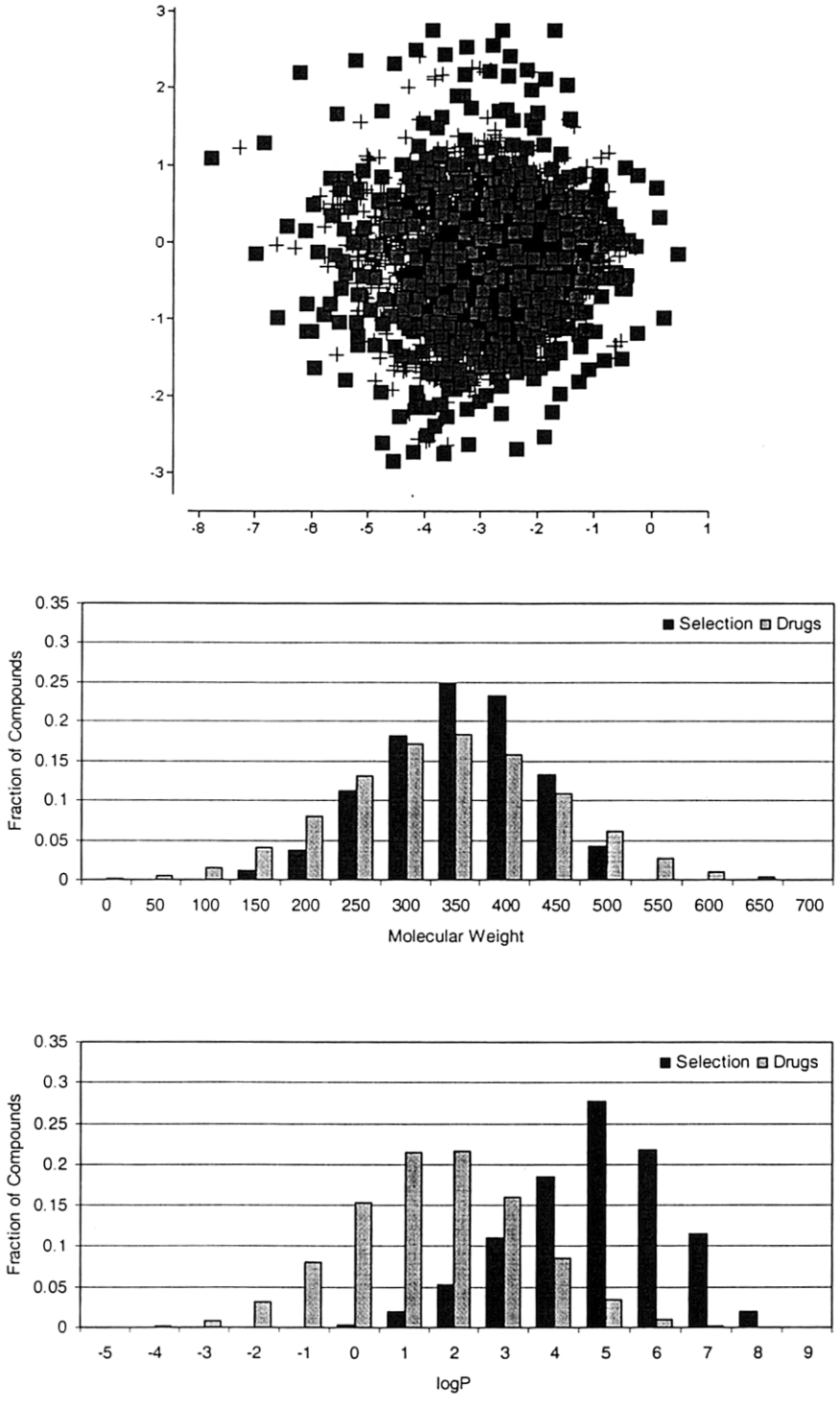

Figure 9. Diversity-based selection of a 20×20 array. (a) Nonlinear projection; (b) molecular weight distribution; (c) logP distribution.

imine intermediate and is utilized for the construction of structurally diverse drug-like molecules with useful pharmacological properties, particularly in the GPCR superfamily.[24] The synthetic protocol for the reductive amination library is illustrated in Figure 4. The reaction is carried out by adding a solution of primary or secondary amine to an equimolar ratio of aldehyde in 1,2-dichloroethane/N,N-dimethyl formamide. Sodium triacetoxyborohydride (2 equivalents) in 10% acetic acid/DMF is added to the reaction vial. Stirring of the reaction mixture at 25°C for 12 hours and subsequent addition of methanol followed by concentration yields the product in high purity.

In this work, a set of 300 primary and secondary amines with 300 aldehydes were selected at random from the Avail-

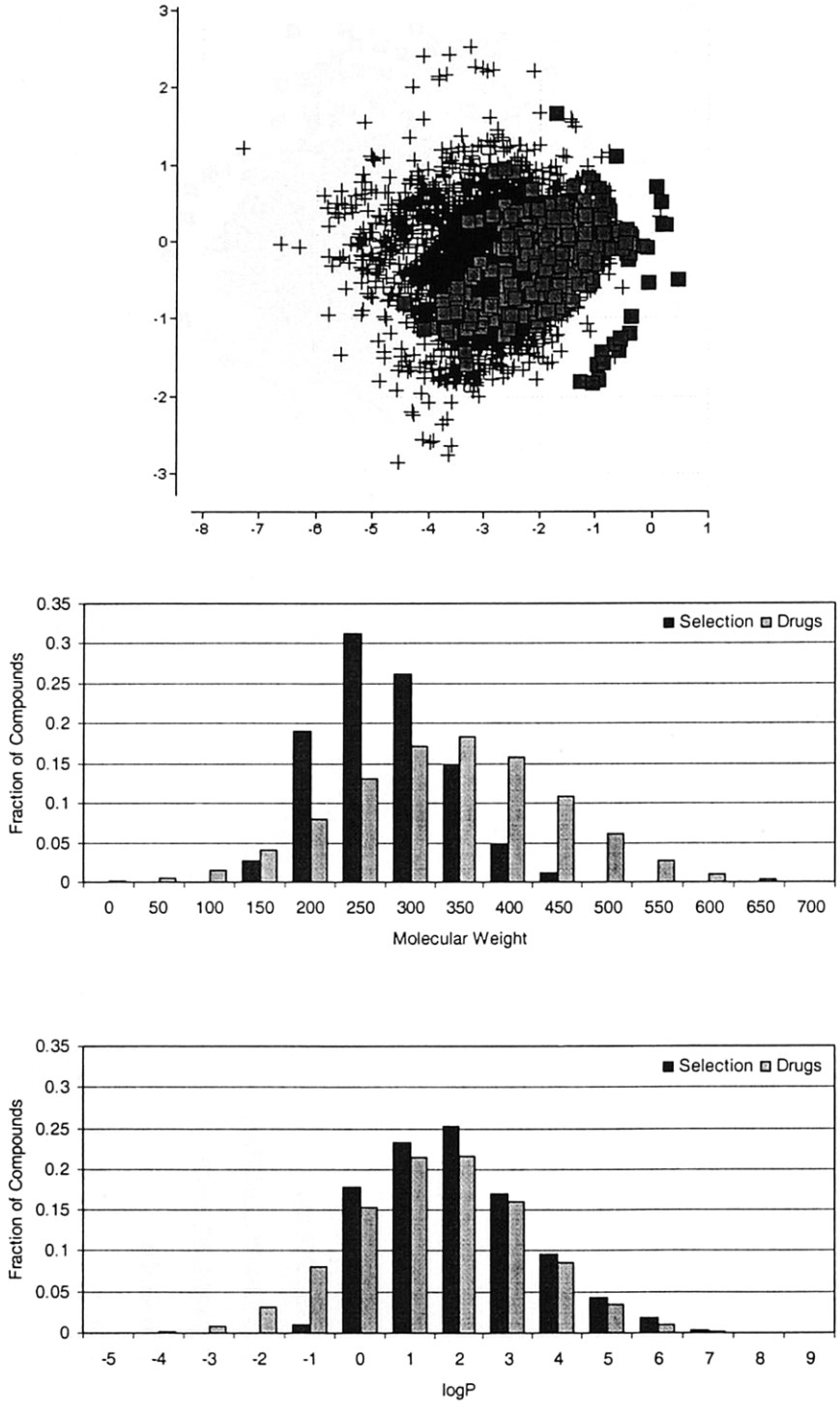

Figure 10. logP-based selection of a 20×20 array. (a) Nonlinear projection; (b) molecular weight distribution; (c) logP distribution.

able Chemicals Directory[25] and were used to generate a virtual library of 90,000 products using the library enumeration classes of the DirectedDiversity® toolkit.[23] These classes take as input lists of reagents supplied in SDF or SMILES format, and a reaction scheme written in a proprietary language that is based on SMARTS and an extension of the scripting language Tcl. All chemically feasible transformations are supported, including multiple reactive functionalities, different stoichiometries, cleavage of protecting groups, stereospecificity, and many others. The computational and storage requirements of the algorithm are minimal (even a billion-membered library can be generated in a few seconds on a personal computer) and scale linearly with the number of reagents. Although the products are encoded

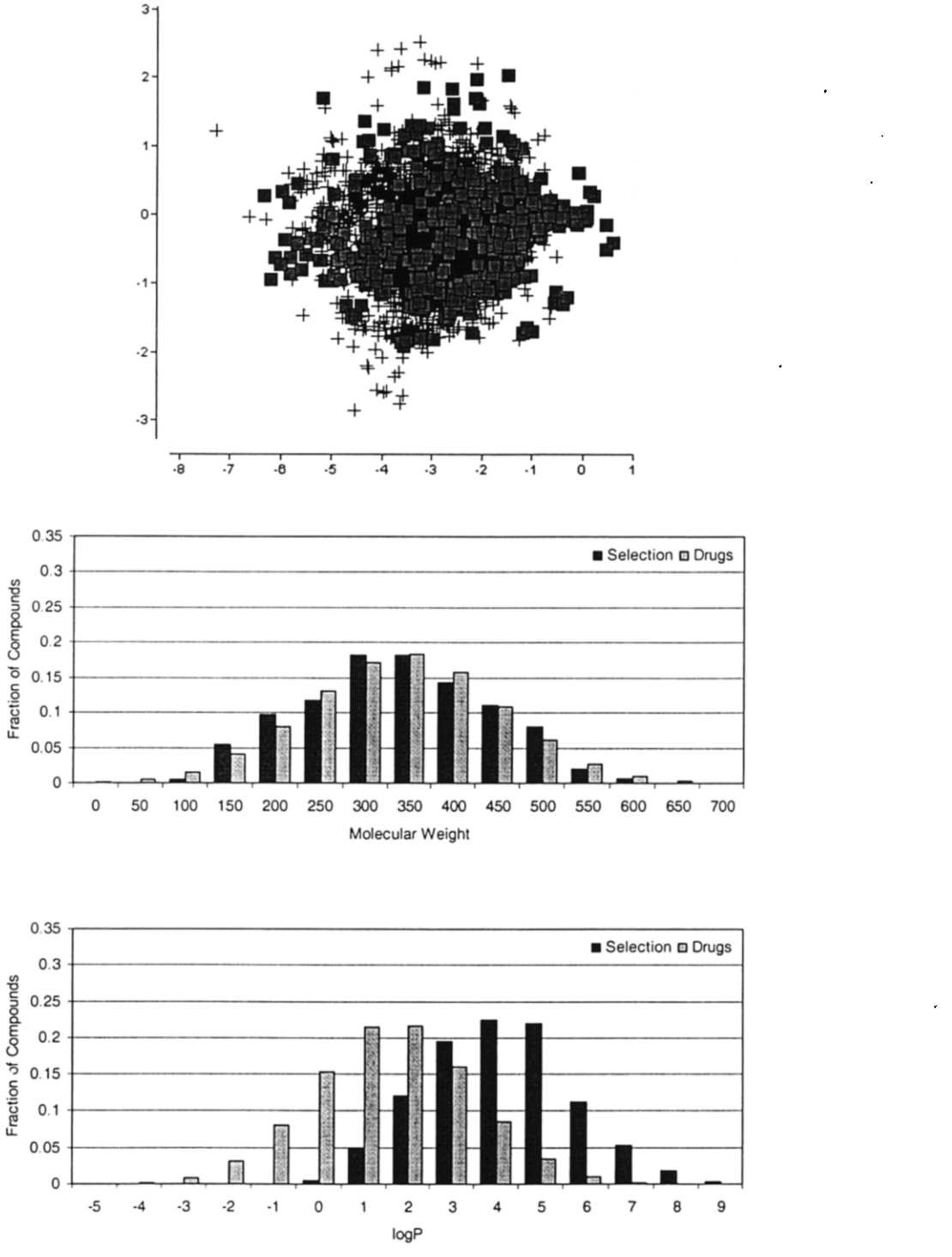

Figure 11. Molecular weight-based selection of a 20×20 array. (a) Nonlinear projection; (b) molecular weight distribution; (c) logP distribution.

implicitly, individual structures are accessible at a rate of 1,000,000 per CPU second.

Each compound in the 90,000-membered library was characterized by a standard set of 117 topological descriptors computed with the DirectedDiversity® toolkit.[23] These descriptors included an established set of topological indices with a long, successful history in structure-activity correlation such as molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev-Trinajstis indices, and topological state indices.[26,27] We previously showed that these descriptors exhibit proper "neighborhood behavior" and thus are well suited for diversity analysis and similarity searching.[28] The term "neighborhood behavior" refers to a characteristic trapezoidal distribution of the pairwise similarity

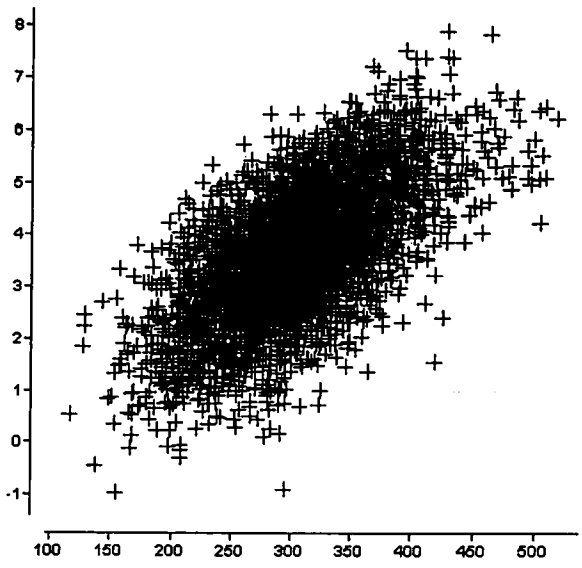

Figure 12. Scatter plot of molecular weight vs *logP for the entire reductive amination library.*

scores plotted against the respective differences in biological activities for a set of related molecules and originally was suggested by Paterson et al.[29] as a qualitative measure of the ability of a particular descriptor set to discriminate between active and inactive molecules.

These 117 molecular descriptors subsequently were normalized and decorrelated using principal component analysis. This process resulted in an orthogonal set of 23 latent variables, which accounted for 99% of the total variance in the data. To simplify the analysis and interpretation of results, this 23-dimensional data set was reduced further to two dimensions using a very fast nonlinear mapping (NLM) algorithm developed by our group.[30,31] The projection was carried out in such a way that the pairwise distances between points in the 23-dimensional principal component space were preserved as much as possible on the two-dimensional map (Figure 5a). These NLM coordinates were used for all diversity calculations described later, even though the Kruskal stress of 0.187 would indicate that the projection is somewhat distorted. The extent of this distortion is illustrated in Figure 6, which shows the true distances (dissimilarities) of 5,000 randomly chosen pairs plotted against the corresponding distances on the nonlinear map. Although some of these distances cannot be perfectly reproduced, the vast majority of points cluster tightly along the diagonal, which indicates that the internal structure and clustering of the data is nicely preserved. However, the reader should be aware that we do not generally advocate the use of two- or three-dimensional nonlinear maps for quantifying molecular diversity, unless the distortion is sufficiently small. The only reason for adopting this approach in the present work is interpretability. If diversity were assessed on a higher dimensionality, minor differences in diversity scores would not necessarily be evident on the projection, and this would compromise our ability to convey the trade-offs involved in multiobjective optimization in a graphical form.

Finally, in addition to the 117 topological descriptors, the octanol-water partition coefficient (logP) of each compound was computed independently using the Ghose-Crippen approach[32] as implemented in the DirectedDiversity® toolkit[23] and was used as the target variable along with molecular weight for all distribution-based designs. This parameter was not included in the descriptor set used for diversity assessment.

DISCUSSION

To establish the "ideal" property distributions of drug-like molecules, we used an expanded subset of 7,484 compounds from the World Drug Index, similar to the one used in Lipinski's original publication. This new set consisted of drugs that had a INN or USAN number and were approved for marketing in at least one country.[6] The distributions of molecular weights and computed logP values are shown in Figure 7a and b, respectively. Although our method for computing logP differs from that used by Lipinski, our results are consistent with his original observations. Indeed, 97% of the compounds had a logP <5 and 87% of the compounds had a molecular weight <500. Because computation of the Kolmogorov-Smirnov criterion becomes significantly faster if the target cumulative distribution function is known analytically, we decided to fit the molecular weight and logP data to a normal distribution using a least-squares fitting procedure and to use that distribution for all property-based designs. The mean and sigma of the fitted Gaussians were 314.3 and 108.3 for molecular weight, and 1.04 and 1.78 for logP, respectively. These functions are shown in Figure 5a and b, along with the original distributions.

How does a typical small-molecule combinatorial library compare to these distributions? As shown in Figure 5b and c, for the reductive amination library, the two distributions differ substantially from those of marketed drugs. In the case of molecular weight, the distribution is much sharper, but the vast majority of compounds still lie within acceptable bounds. However, this is not the case for logP, where there is an upward shift by more than three units compared to the WDI set. If we assume that the shape of the distribution is an indication of the likelihood of a compound being a drug, this result would imply that a significant fraction of these compounds are likely to exhibit adverse pharmacokinetic properties and prove difficult candidates for SAR refinement.

What one would expect from a random design is shown in Figure 5b anc c. However, most combinatorial libraries are not designed at random but usually in a way that maximizes their molecular diversity. Diversity typically is expressed as the degree of "coverage" of chemical space, defined either directly using a set of structural, physical, chemical, or biological properties of the compounds, or indirectly through the use of some pairwise measure of molecular similarity.[11,12] To assess whether diversity has any significant effect on the two distributions, we carried out a selection of 400 diverse structures using the annealing algorithm described earlier and the diversity function in Equation 3. These compounds were selected as singles, i.e., without regard to the number or types of reagents involved. The selection is shown in Figure 8a, and the corresponding distributions in Figure 8b and c, respectively. Whereas the molecular weight distribution matches more closely that of the reference set, the logP distribution becomes wider but remains shifted toward the higher logP range. The same is true when the selection is carried out in the form of an array. Figure 9 shows a diverse set of 400 compounds selected as a 20×20 array (i.e., 20 amines combined with 20 aldehydes in all possible combinations). Although the array is noticeably

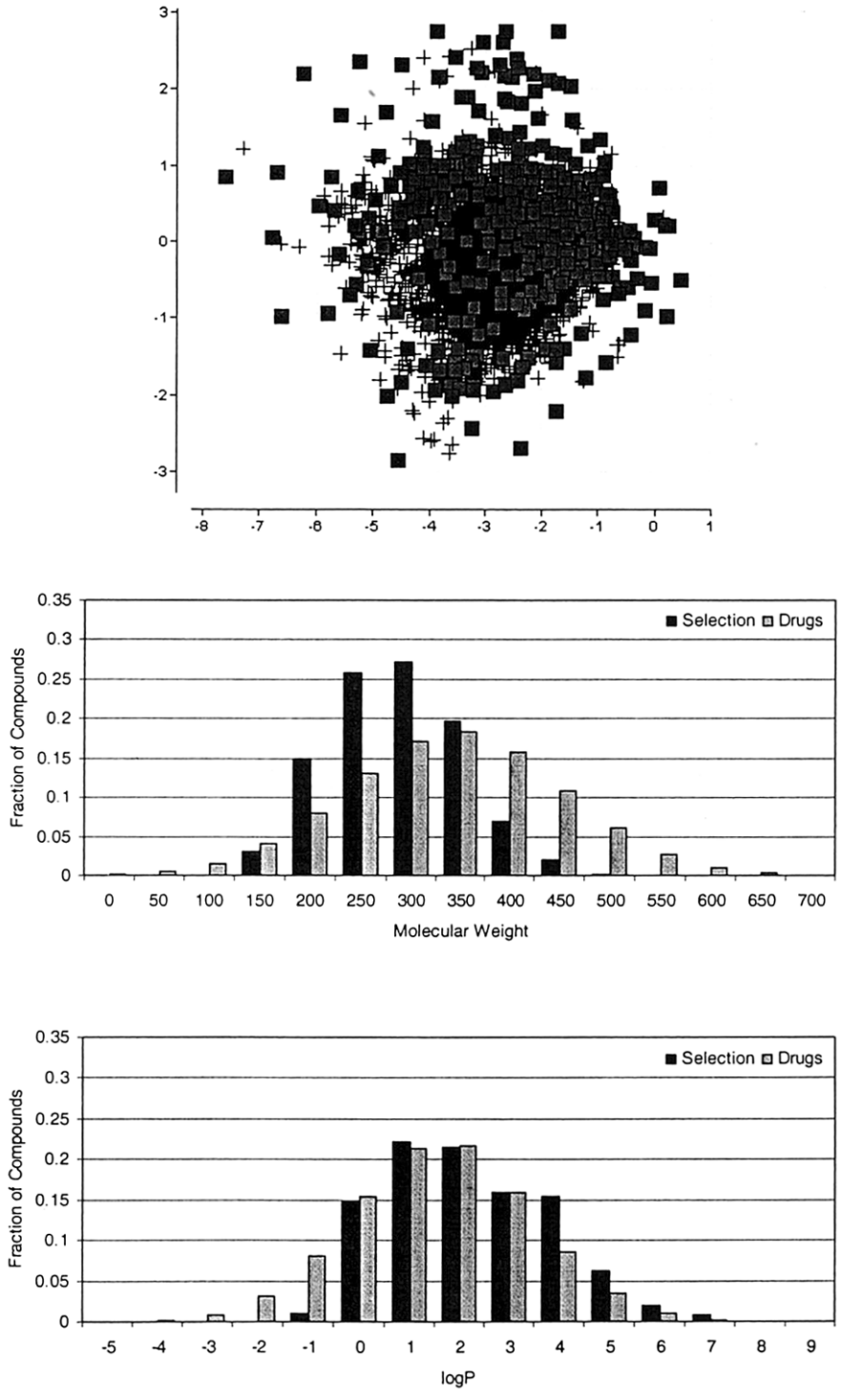

Figure 13. Multiobjective selection of a 20×20 array based on diversity and logP (Equation 6). (a) Nonlinear projection; (b) molecular weight distribution; (c) logP distribution.

less diverse in terms of spread, it requires only 40 reagents as compared to 316 reagents (154 amines and 162 aldehydes) required by the singles. For this reason, we decided to use 20×20 arrays for all the remaining selections.

Thus, diversity by itself is not a sufficient condition and leads to designs that are severely compromised in terms of their logP characteristics. The opposite also is true, that is, selections based on molecular weight or logP considerations alone lead to an undesired behavior in terms of the other two design parameters. If the selection is based exclusively on logP (i.e., if it is carried out in a way that maximizes the value of the Kolmogorov-Smirnov criterion applied to the logP distribution), the diversity of the resulting array is reduced (Figure 10a) and the molecular weight is shifted downward by approximately 60 D. This is partly due to the

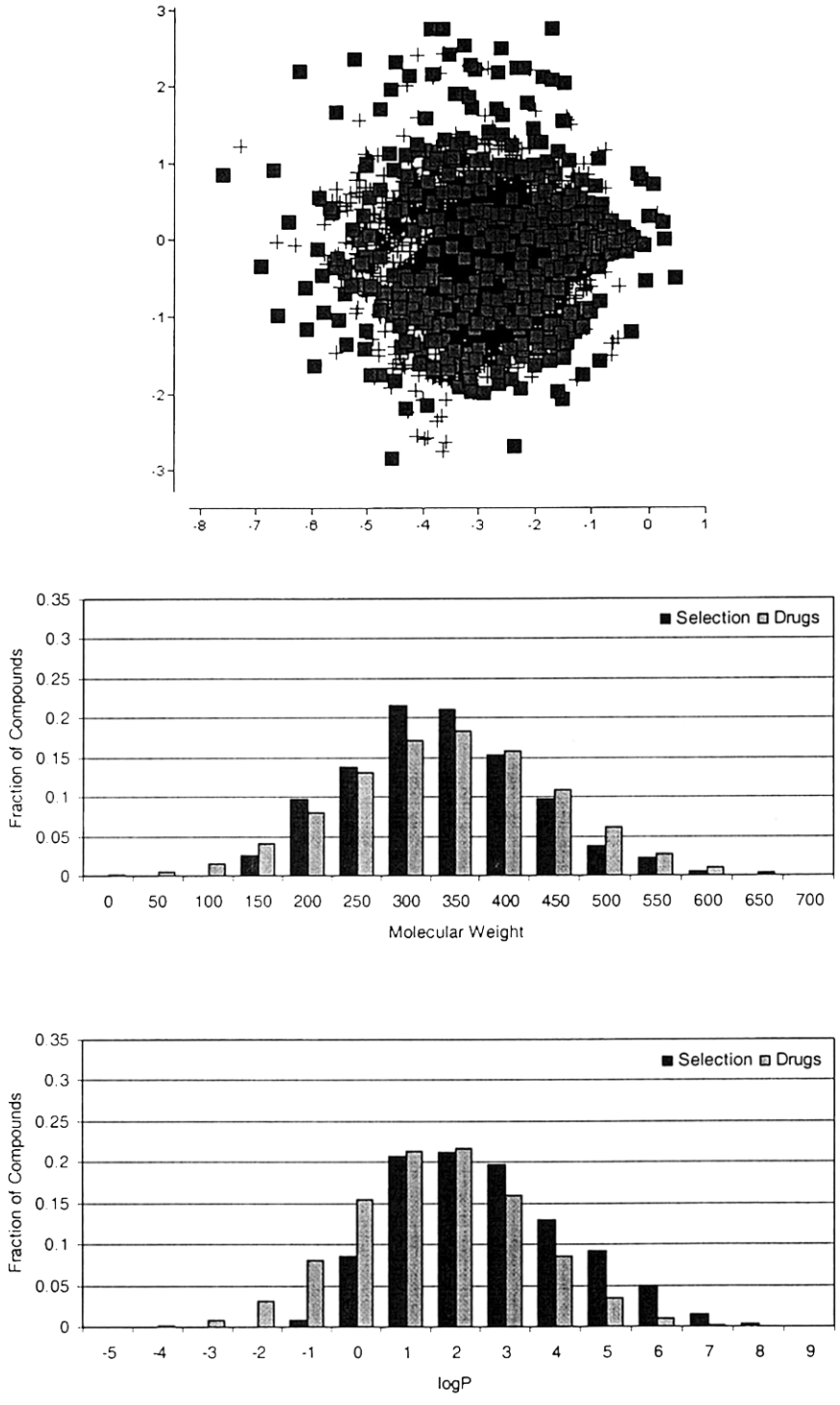

Figure 14. Multiobjective selection of a 20×20 array based on diversity, molecular weight and logP (Equation 7). (a) Nonlinear projection; (b) molecular weight distribution; (c) logP distribution.

fact that the two variables are correlated (the correlation coefficient is 0.72) (See Figure 11). As suggested by Figure 12, in the absence of a stabilizing factor, a downward shift in molecular weight will be accompanied by a similar shift in logP and vice versa. These two parameters also play a dominant role on the shape of the nonlinear map, which is organized in a way that clusters compounds from left to right in decreasing order of size.

In a similar fashion, selections based exclusively on molecular weight have very little effect on the logP distribution, but they do improve the diversity of the set compared to the logP-based design (a consequence of the fact that the Kolmogorov-Smirnov criterion causes the molecular weight distribution to widen and, as a result, the spread on the nonlinear map to increase).

These examples make it abundantly clear that selection

algorithms based on a single objective are limited in the types of designs they can produce and cannot accomplish complex tasks that involve the simultaneous optimization of several, often conflicting, parameters. As we demonstrate in the last two examples, this can be accomplished easily using multiobjective fitness functions that encode all the desired selection criteria and the influence that each should have on the final design. Consider the following objective function:

$$f = D + 0.2 \cdot K(\log P), \quad (6)$$

where D represents the diversity of the ensemble computed by Equation 3, and K(logP) is the value of the Kolmogorov-Smirnov criterion for the logP distribution computed by Equation 5. As with most problems of this type, the most difficult task is to assign a meaningful set of coefficients used to weight the individual objectives. In our case, the coefficients were determined empirically based on the maximum value that each criterion could assume independently. These values represent the "upper bounds," i.e. the best diversity and logP distribution that one could expect from a 20×20 array. In the case at hand, the mean nearest neighbor distance, D, in the diversity-based selection in Figure 9 was 0.18, whereas the value of K for the logP-based selection in Figure 10 was 0.88. These values suggested that, for the two criteria to be placed on an equal footing, the value of K had to be scaled down by approximately a factor of 5. In our experience, this method of assigning weights is simple and extremely effective in practice. In the pathological cases where the energy landscapes (i.e., the distributions of scores) of the individual criteria are very different, alternative, more complex objective functions can be devised. In fact, this represents one of the greatest advantages of our multiobjective approach, which offers the user full control over the design.

As is evident from Figure 13, Equation 6 accomplishes a goal that no criterion could achieve independently. The logP distribution is nearly perfect, and at the same time the diversity of the set is greatly enhanced compared to the logP-based design (Figure 10a). Clearly, some diversity is sacrificed, but this is an inevitable compromise that one must be prepared to accept in order to achieve all the desired goals. However, the selection has still one imperfection: for reasons described earlier, the downward shift in the logP distribution caused a concomitant shift in molecular weight. Clearly, to achieve the "perfect" design, we need an objective function that takes into account all three aforementioned objectives. Using the same type of reasoning on assigning coefficients, we performed a final selection using the objective function:

$$f = D + 0.2 \cdot K(\log P) + 0.2 \cdot K(mw) \quad (7)$$

where D and K(logP) have the same significance as in Equation 6, and K(mw) represents the value of the Kolmogorov-Smirnov criterion for the molecular weight distribution. The results are illustrated in Figure 14. The new objective function removes the imperfection in the molecular weight distribution without seriously affecting diversity and logP, and it leads to a design that is not only diverse but also fully consistent with historical evidence of what constitutes drug likeness.

Although it is not immediately obvious, there are two additional design objectives that are satisfied in all the previous selections. These objectives relate to the number of building blocks and ease of synthesis, which are implicitly accommodated by the fact that the selection is carried out as an array. Unlike diversity and property distribution, these criteria are incorporated directly into the search engine and cannot be compromised in favor of another objective. Finally, we should stress that the examples used in this work represent only a small fraction of possible selection criteria that the medicinal chemists may wish to employ in library design. A more thorough discussion of the potential applications of this method will be presented elsewhere.

ACKNOWLEDGMENT

The authors are grateful to Dr. Raymond F. Salemme of 3-Dimensional Pharmaceuticals, Inc., for his insightful comments and support of this work.

REFERENCES

1 Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeny, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 1997, **23**, 3–25
2 Martin, E.J., Spellmeyer, D.C., Critchlow, R.E., and Blaney, J.M. Does combinatorial chemistry obviate comptuer-aided drug design? In: *Reviews in computational chemistry, Volume 10*, Lipkowitz, K.B., and Boyd, D.B., Eds., VCH, New York, 1997, pp. 75–100
3 Martin, E.J., and Critchlow, R.E. Beyond mere diversity: Tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.*, 1999, **1**, 32–45
4 Martin, E.J., and and Wong, A. Sensitivity analysis and other improvements to tailored combinatorial library design. *J. Chem. Info. Comput. Sci.*, 2000, **40**, 215–220
5 Koehler, R.T., Dixon, S.L., and Villar, O.H. LASSOO: A generalized directed diversity approach to the design and enrichment of chemical libraries. *J. Med. Chem.*, 1999, **42**, 4695–4704
6 Lipinski, C.A. Personal communication
7 Agrafiotis, D.K., Bone, R.F., Salemme, F.R., and Soll, R.M. United States Patents 5,463,564, 1995; 5,574,656, 1996; 5,684,711, 1997; and 5,901,069, 1999
8 Agrafiotis, D.K. Stochastic algorithms for maximizing molecular diversity. Third Electronic Computational Chemistry Conference, *http://hackberry.chem.niu. edu/ECCC3/paper48*, 1996
9 Agrafiotis, D.K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Info. Comput. Sci.*, 1997, **37**, 841–851
10 Agrafiotis, D.K. On the use of information theory for assessing molecular diversity. *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 576–580
11 Agrafiotis, D. K. The diversity of chemical libraries. In: *The encyclopedia of computational chemistry*, Schleyer, P.v.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer, H.F. III, and Schreiner, P.R., Eds., John Wiley & Sons, Chichester, 1998, pp. 742–761
12 Agrafiotis, D.K., Myslik, J.C., and Salemme, F.R. Advance in diversity profiling and combinatorial series design. *Mol. Diversity*, 1999, **4**, 1–22
13 Agrafiotis, D.K., and Lobanov, V.S. An efficient implementatin of distance-based diversity metrics based on k-d trees. *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 51–58
14 Brown, R.D., and Martin, Y.C. Designing combinatorial

library mixtures using genetic algorithms. *J. Med. Chem.*, 1997, **40**, 2304–2313
15 Gillet, V.J., Willet, P., Bradshaw, J., and Green, D.V.S. Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Info. Comput. Sci.*, 1999, **39**, 169–177
16 Hassan, M., and Waldman, M. Penalty biased diversity: Design of diverse drug-like libraries. *Book of Abstracts*, 218 ACS National Meeting, New Orleans, Louisiana, August 22–26, 1999
17 Lajiness, M.S. The evaluation of the performance of dissimilarity selection. In: *QSAR: Rational approaches to the design of bioactive compounds*, Silipo, C., and Vittoria, A., Eds., Elsevier, Amsterdam, 1991, pp. 201–204
18 Downs, G.M., and Willett, P. Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1094–1102
19 Pearlman, R.S., and Smith, R.S. Novel software tools for chemical diversity. *Perspect. Drug Discovery Design* 1998, **9**, 339–353
20 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.*, 1995, **38**, 1431–1436
21 Taylor, R. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 59–67
22 von Mises, R. *Mathematical theory of probability and statistics*. Academic Press, New York, 1997
23 The Mt Toolkit: An Object-Oriented C++ Class Library for Molecular Simulations. Copyright 3-Dimensional Pharmaceuticals, Inc., 1994–2000
24 Dhanoa, D.S., Gupta, V., Sapienza, A., and Soll, R.M. Poster 26, American Chemical Society National Meeting, Anaheim, California, 1999
25 Available Chemicals Directory is marketed by MDL Information Systems, Inc., 140 Catalina Street, San Leandro, CA 94577
26 Hall L.H., and Kier, L.B. The molecular connectivity chi indexes and kappa shape indexes in structure-property relations. In: *Reviews of computational chemistry*, Boyd, D.B., and Lipkowitz, K.B., Eds., VCH Publishers, New York 1991, pp. 367–422
27 Bonchev, D., and Trinajstic, N. Information theory, distance matrix and molecular branching. *J. Chem. Phys.* 1977, **67**, 4517–4533
28 Lobanov, V.S., and Agrafiotis, D.K. Stochastic similarity selection from large combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 2000, **40**, 460–470
29 Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., and Weinberger, L.E. Neighborhood behaviour: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 1996, **39**, 3049–3059
30 Agrafiotis, D.K. A new method for analyzing protein sequence relationships based on Sammon maps. *Prot. Sci.* 1997, **6**, 287–293
31 Agrafiotis, D.K., Lobanov, S.V., and Salemme, F.R. Patents pending
32 Ghose, A.K., Viswanadhan, V.N., and Wendoloski, J.J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of alogp and clogp methods. *J. Phys. Chem. A*, 1998, **102**, 3762–3772