

LASSOO: A Generalized Directed Diversity Approach to the Design and Enrichment of Chemical Libraries

Ryan T. Koehler, Steven L. Dixon, and Hugo O. Villar*

Telik, Inc., 750 Gateway Boulevard, South San Francisco, California 94080

Received June 17, 1999

Pharmaceutical discovery relies on the screening of chemical libraries that are as diverse as possible yet constrained in favor of compounds possessing attributes that are normally associated with successful drug candidates. We describe a new algorithm for simultaneously addressing both objectives, providing an effective means to increase structural diversity in a chemical library while maintaining a bias toward compounds that retain the desirable properties of drugs. The LASSOO algorithm exploits differences in descriptor distributions to identify novel compounds that are most dissimilar to the members of an existing screening library and most similar to members of a target library with desirable characteristics. We illustrate the LASSOO technique using publicly available compound databases and bit string descriptors. The architecture of the algorithm is general enough to allow any set of descriptors or similarity measures to be employed, and it is easily adaptable to other means of directing diversity, such as the avoidance of toxicity and/or poor pharmacokinetic properties.

Introduction

High-throughput screening capabilities, combined with advances in combinatorial chemistry, have made possible routine screening of increasingly large chemical libraries in the search for bioactive lead compounds.¹ In principle, the chances of uncovering active compounds grow with library size, thus providing impetus for screening very large libraries. However, a brute force approach of blindly increasing library size to improve the odds of uncovering active leads is neither efficient nor sufficient. Rapid, random expansion of a library could result in a significant chemical redundancy; hence the amount of additional information provided is frequently far offset by the consumption of resources. Furthermore, even huge libraries are insufficient if they are devoid of compounds possessing some essential pharmacophore. To address both concerns, a number of diversity-based methods have been devised for selection of nonredundant structures which cover the greatest possible fraction of “chemical space” as measured by any particular set of descriptors.^{2,3}

While chemical diversity and coverage are clearly desirable attributes of a screening library, the *quality* of compounds comprising such libraries is equally important.⁴ Compounds screened against biological targets should have attributes consistent with their intended end use as pharmaceuticals.⁵ All else being equal, a library enriched with “drug-like” compounds is considered superior to a library without such a bias, simply because any active leads discovered are less likely to be eliminated downstream during lead optimization and clinical evaluation.

Empirical rules for evaluating compounds as potential pharmaceuticals have been used by medicinal chemists for some time. For example, constraints on molecular

weight, log *P*, the number of heteroatoms, or the presence or absence of certain chemical functionalities have been described.^{6,7} More recently, computer-aided quality evaluation methods based on pattern recognition have been implemented, facilitating automatic assessment of how drug-like a compound appears to be.^{8,9} Rules and classification methods are useful for evaluating existing or even virtual compounds, but how such methods should direct selection of compounds for construction or augmentation of a screening library is not clear. Simply filtering out unacceptable structures or selecting some subset that appears most drug-like does not guard against the addition of numerous closely related analogues. Moreover, because such approaches ignore the composition of any existing library, it is entirely possible that compounds will be selected which are highly similar to those that one already has. One potential solution is to utilize diversity-based selection criteria either before or after the filtering steps. However, this simple “intersection” of drug-like properties and chemical diversity can lead to the acquisition of compounds that are far from optimal. What is needed is a means of directly coupling the two driving forces, so that compounds with favorable attributes are selected while the overall diversity of the library is simultaneously improved.

Here we describe a new algorithm, LASSOO (Library Acquisition with Simultaneous Scoring to Optimize Ordering), that facilitates selection of compounds for addition to an existing chemical library. The algorithm is intended to prioritize compounds that are most similar to a specified set of favorable target molecules, and, at the same time, most different from compounds that reside in the library which is being augmented. Our goal is to strike a balance between simply “filling holes” for the sake of increasing diversity per se and restricting interest to only those compounds conforming to some rigid predefined notion of quality.

* Corresponding author: e-mail, hugo@telik.com; fax, (650) 244-9388.

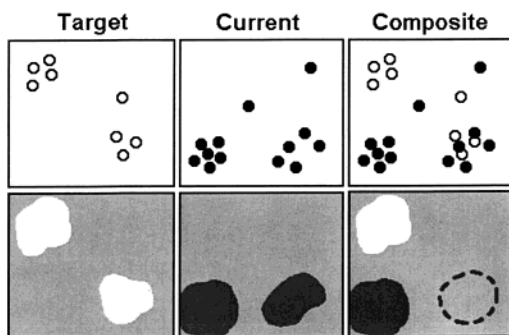


Figure 1. Schematic depiction of compounds (top row) and corresponding local densities (bottom row) in a two-dimensional descriptor space. Target and internal libraries are illustrated, together with the composite of both libraries. Top row panels depict individual compounds; open circles represent target pool records; filled circles represent internal pool records (see text for details). Bottom row panels depict local densities corresponding to records shown in the panels immediately above them. Regions of descriptor space that are (relatively) overpopulated with target records are light, and regions overpopulated with internal records are dark. For the situation illustrated, LASSOO favors external records corresponding to light regions and disfavors records in dark regions.

Overview of the Strategy

Figure 1 illustrates the general premise behind LASSOO. Here, libraries are represented in a hypothetical two-dimensional descriptor space. Individual compounds are denoted by circles in the upper set of panels, while localized regions of elevated compound density are shown in the lower panels. The *target* library is comprised of compounds which have desirable characteristics, for example, known drugs, and the *internal* library contains compounds that have already been acquired. In the *composite* library, compound densities from the target and internal libraries are combined to define regions of descriptor space that are favorable (light) and unfavorable (dark). This shading reflects the numerical score that a compound from an *external* library would receive if it were a candidate for addition to the internal library. For example, the upper left corner of space is heavily populated by desirable target compounds yet devoid of any density from the internal library, so an external candidate compound located in this region would receive a favorable score. Conversely, a candidate compound located in the lower left corner would receive an unfavorable score because there are no target compounds in this region, and the existing internal library already exhibits a high density of compounds there. Note that the lower right corner (circled) is neither distinctly favorable nor distinctly unfavorable because this region is populated in both the target and internal libraries. Regions where density is *lacking* in both the target and internal libraries, such as the upper right area, also receive a neutral score.

The LASSOO algorithm treats each library as a *pool of records*. A record is a unique point in descriptor space that is occupied by one or more compounds in a library, with a many-to-one mapping occurring whenever the descriptors are unable to distinguish certain compounds from each other. A pool is simply the collection of records generated by a library. Each record in the external library is evaluated or scored according to its position in the descriptor space and the corresponding local

density of records arising from the target and internal libraries. Local density is based on probing the region around a point in space and counting the number of records in the pool that are within some threshold distance of that point. Density arising from the target library adds a positive or favorable contribution to the score, while density from the internal library makes a negative or unfavorable contribution. Records with the highest scores thus provide the best candidates for inclusion in the internal library.

We note that while records are scored here using only target and internal pool densities, it is possible to employ any number of pools encoding desirable or undesirable characteristics. For example, one may wish to avoid compounds that are likely to be toxic or which may have poor pharmacokinetic properties. Accordingly, a library containing these sorts of compounds could be constructed and used to provide additional negative contributions to the scores of external compounds.

To assess the utility and parameters of LASSOO, a fraction of the target library was removed and used to "spike" the external library. Performance is then measured by the algorithm's ability to locate these spiking compounds while maintaining diversity in the internal library. For comparison, baselines of performance are established by "turning off" either the target density contribution or the internal density contribution to the score. When target density is turned off, the algorithm simply tries to fill holes in the internal library, with no particular attention being paid to adding compounds with desirable properties. Conversely, when internal density is ignored, the entire focus is on acquiring compounds that are most similar to the target collection, irrespective how much redundancy results.

Methods

Compound Libraries. Testing was conducted on libraries of compounds taken from the Available Chemicals Directory (ACD version 98.2), National Cancer Institute (NCI3D version 94.1a), and Comprehensive Medicinal Chemistry (CMC3D version 98.1) databases.¹⁰ After performing a series of filtering steps (vide infra), the remaining compounds were used to define the various libraries required by LASSOO. The ACD database served purely as a source of compounds for the external library, while the NCI database provided a starting point for the internal library. Because of its high content of pharmaceuticals and clinically evaluated compounds, the CMC database was used to supply drug-like compounds for the target library and for the purpose of spiking the external library.

No filtering was applied to the NCI database, and thus all 126 554 compounds contained therein were used to provide a pool of records for internal libraries. The ACD collection was filtered to remove compounds that did not contain at least one carbon atom, which left 249 867 entries. Filtering the CMC was more complex, as a number of nondrug or unclassified compounds appear in this database. First, compounds with missing, ambiguous, or otherwise unusable structures (non-standard elements such as X, R; amino acid codes; etc.) were discarded. Next, compounds of unspecified therapeutic class as well as compounds which have no real medicinal value, for example, adhesives, disinfectants, herbicides, lubricants, propellants, solvents, etc., were removed following a literature procedure.¹¹ This left 5 753 CMC compounds with genuine therapeutic applications to serve as the target library.

Structural Descriptors. Compounds were characterized using the MDL keys, which were accessed using the MOLSKY feature of the ISIS program.¹² These binary descriptors

are based on 166 predefined substructure queries which encode the presence or absence of numerous types of two-dimensional fragments and functionalities. Although they do not rely on any sort of exhaustive enumeration of substructures, several studies have shown the MDL keys to be effective descriptors for classification and clustering of biologically active molecules.^{13,14}

Dissimilarities between pairs of compounds were measured using city-block distances, which, in the special case of binary descriptors, is simply the number of positions at which the two bit strings differ, i.e., the logical XOR summation. This is perhaps the simplest means of measuring distance between binary representations, and it is less influenced by compound size than the well-known Tanimoto coefficient.^{15,16}

MDL keys were exported for each filtered database, and compounds with identical bit string representations were collapsed to a single record. This procedure reduced the counts of unique records to 105 219 for the NCI, 165 858 for the ACD, and 5 487 for the CMC. From this point on, records and pools replaced the more familiar notions of compounds and libraries.

Scoring Function. External record scores are a function of the local densities due to target and internal pool records, as observed at the coordinates of the external record. If we let $N_{T_k}(d)$ and $N_{I_k}(d)$ represent the number of target and internal records, respectively, that are located within a distance d of external record k , and we let $A(d)$ be an attenuation function that drops to zero beyond some threshold distance, then the score for external record k may be represented as

$$\text{score}_k = (\alpha_T/N_T) \sum_{d=0}^{d_{\max}} N_{T_k}(d)A(d) + (\alpha_I/N_I) \sum_{d=0}^{d_{\max}} N_{I_k}(d)A(d) \quad (1)$$

Terms associated with each pool are weighted to reflect the role of the pool in scoring. For results presented in this work, the (favorable) target pool weighting factor, α_T , is unity, and the internal pool weighting factor, α_I , is negative unity. In addition, terms for each pool are normalized by the total number of records in the pool to reduce the influence of pool size on scores. The attenuation function returns a full or near full count of records for shorter distances but, for a smoothly decaying form, gives progressively less weight to members of the target and internal pools that are increasingly distant from the external record being scored.

Local density values primarily reflect library composition but will obviously depend on the attenuation function used to calculate them. Accordingly, choice of the parameters that define $A(d)$ has direct bearing on scores and, ultimately, on method effectiveness. In our experiments, we vary both the shape of $A(d)$ as well as the threshold distance beyond which the function is zero. Shapes investigated include a step function, a linear ramp, an inverted parabola, and a Gaussian curve. For a threshold distance of d_{limit} , these functions are defined accordingly:

$$\text{step: } A(d) = \begin{cases} 1 & (0 \leq d \leq d_{\text{limit}}); \\ 0 & (d > d_{\text{limit}}) \end{cases}$$

$$\text{linear ramp: } A(d) = \begin{cases} 1 - d/d_{\text{limit}} & (0 \leq d \leq d_{\text{limit}}); \\ 0 & (d > d_{\text{limit}}) \end{cases}$$

$$\text{inverted parabola: } A(d) = \begin{cases} 1 - (d/d_{\text{limit}})^2 & (0 \leq d \leq d_{\text{limit}}); \\ 0 & (d > d_{\text{limit}}) \end{cases}$$

$$\text{Gaussian: } A(d) = \begin{cases} \exp[-(2.146d/d_{\text{limit}})^2] & (0 \leq d \leq d_{\text{limit}}); \\ 0 & (d > d_{\text{limit}}) \end{cases}$$

The Gaussian function is defined so that it decays to 0.01 by the time the threshold distance is reached.

We note an alternative to eq 1 that yields equivalent scores and is amenable to continuous distance functions is possible. Rather than summing over the counts of records at discrete distances from records being evaluated and then normalizing

by pool size, summation may be performed over the contributions of each target and internal pool record to local density followed by normalization. Iterative updating would then involve simply adding contributions of accepted external records and normalizing by the new internal pool size. Use of explicit counts of nearby target and internal records (eq 1) is useful, however, when a number of different attenuation functions are being evaluated. This is because a single table of distance counts associated with each external record may be used for many experiments. Once a suitable threshold and attenuation function are identified, summation over record contributions to local densities could be used to compute scores. Continuously valued distance functions may be employed with eq 1 simply by assigning distances to a discrete set of "bins".

The present work is concerned with identification of a suitable set of parameters for the MDL keys in combination with a city-block distance measure. Within this system, the maximum possible pairwise distance between records is 166 bits, corresponding to a situation in which all discriminated features present in one structure are absent from another structure and vice versa. In practice, this sort of situation never arises, and the most frequently encountered distances are in the range of 45–50 bits. To ensure that the attenuation functions measure density in a local sense, the threshold should be somewhat less than these most commonly occurring distances. Tests using threshold distances of 10, 20, 30, and 40 bits are described.

Algorithm. LASSOO is a multipass algorithm. It is conceptually simple, involving first construction of relevant pools, then scoring and ranking of external pool records, and finally selection of high-scoring records. At each iteration a set of W best scoring records are selected and moved from the external pool to the internal pool, where W is the "window size". Addition of records to the internal pool will alter local density, increasing it around the new record coordinates and decreasing it slightly elsewhere due to the change in the pool size normalization factor N_I . The updated densities will of course produce new scores for all of the remaining records in the external pool, which, in turn, will result in a new set of rankings. This added level of complexity, however, is exactly the feedback that enables LASSOO to avoid the repeated selection of structurally similar compounds.

When updating local densities, it is necessary to consider only the pairwise distances between each new record being added to the internal pool and the remaining records in the external pool. The observed distance d_{jk} between a newly added internal record j and each external record k is assigned to the correct distance bin d , and the corresponding record count $N_{I_k}(d)$ is incremented by 1. Target pool density does not change between iterations, so it is not necessary to modify this contribution to the scores.

An additional parameter in the algorithm is the window size, W , which is the number of high-scoring records that are accepted into the internal pool at each iteration. A large window size results in fewer iterations and increases speed, but because local densities are updated less frequently, it also increases the chances of adding self-similar records to the internal pool. To address this issue, we present results obtained for window sizes of 10, 30, 100, and 1 000 records as well as results from using an infinite window size. When tie scores arise within a window, all records with the same score are incorporated, even if the window size must be exceeded in order to admit the entire block of records. Alternatively, one could randomly pick tied records until the number specified by window size had been accepted, though this was not done here.

Validation Experiments. The ability of the algorithm to identify target-like compounds is measured by spiking the external pool with records that are taken from the same source as that used for the target pool. Validation required carrying out the procedure several times in order to generate some statistics. Test-case pools for the validation experiments were generated using random subsets of 10 000 records chosen from both the ACD and NCI pools, and the CMC records were

randomly divided into two subsets of 1 000 and 4 487. The NCI records were treated as the internal pool, and the subset of 4 487 CMC records served as the target pool. ACD records were combined with the remaining subset of 1 000 CMC records (merging identical records, where necessary) to yield a spiked test-case external pool of just under 11 000 records.

For each combination of attenuation function, threshold distance, and window size, the LASSOO program was run with 10 different initial random pools (lists of compounds used may be obtained from the authors in electronic form). As the algorithm made its selections from the spiked external pool, cumulative tallies were kept of the number of CMC records extracted. When identical records were present in both the CMC and ACD random subsets, these were attributed to the CMC pool. These cumulative tallies were then averaged across the set of 10 experiments to yield statistically smooth results summarizing the rate at which CMC records were incorporated into the internal pool.

To ascertain how differences in target and internal pool size influence results, additional experiments were carried out wherein the size of the initial internal pool was varied. Selection and processing steps identical to those described above were followed, except that instead of starting with an internal pool of 10 000 NCI records, random subsets of either 4 487 or 20 000 NCI records were chosen. This yields internal and target pools that are initially the same size (4 487) and pools that exhibit more than a 4-fold difference in size (4 487 target records, 20 000 internal records). Experiments with each size combination were carried out 10 different times using favorable parameter combinations identified using the original set of pool sizes.

The value of the present approach for directed diversity selection was further assessed with a series of spiking experiments in which either the target or internal pool density score terms were "turned off". Each run involved a Gaussian attenuation function, a threshold distance of 20 bits, and a window size of 10 records. The diversity of the internal library during these tests was monitored by following the average minimum distance between internal pool records as each external record was added. Average minimum distance is indicative of record separation, with increasing values corresponding to decreasing redundancy. Coverage continuously increases with addition of external records absent from the internal pool. Controls involving unaided selection were also run by simply picking external pool records at random.

We note that the process of ignoring internal density, given by the initial internal library, actually removes the second term from eq 1, and it is therefore not equivalent to simply running LASSOO in its normal two-term mode with an *empty* internal pool at the outset. The latter situation corresponds to *de novo* selection, where a diverse, target-like internal pool is created from scratch. LASSOO may certainly be used for such purposes, although tests performed here always involved the specification of an existing internal pool.

Run-Time Performance. Typical run-times for tests described in this work, i.e., an external pool of ~11 000 records, an internal pool of 10 000, and a target pool of 4 487, took approximately 15 CPU minutes with a window size of 100 (SGI Indigo² R10,000 with 64 Mb). No code optimization was attempted, and the radial distributions of target and internal record counts used for score calculations (eq 1) were read from disk as needed during each iteration. While libraries of 10 000 compounds may be unrealistically small by today's standards, application to libraries with hundreds of thousands of compounds is certainly possible and should take at most a few days with the system we describe.

Results and Discussion

Analysis of Pairwise Distances. LASSOO scores rely on local density distributions arising from the target and internal pools, and these are determined by pairwise distances that link records in the external pool to records in either the target pool or internal pool. Before

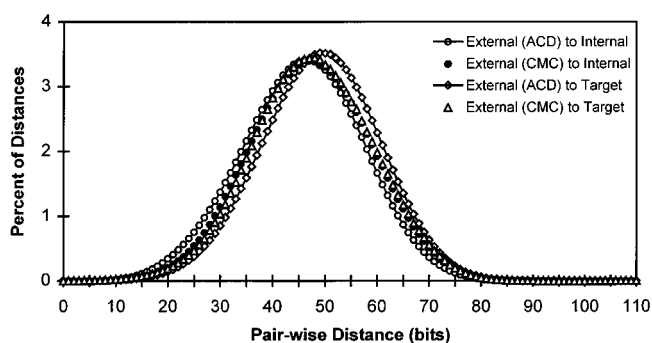


Figure 2. Distribution of pairwise distances observed between external pool records and target and internal pool records. Distances from ACD and CMC records in the spiked external pools are tabulated separately. Averaged data from 10 spiking experiments are shown.

making decisions about which parameter settings to investigate, it is instructive to carry out an analysis of the relevant distance distributions. Histograms of pairwise distances from the initial pools created for the spiking experiments are shown in Figure 2. Average record counts from the 10 sets of random pools were used to generate the histograms, and the ACD and CMC components (10 000 and 1 000 records, respectively) of the spiked external pools are displayed in separate curves. As described previously, internal pools contained 10 000 NCI records and target pools contained 4 487 CMC records.

The positions of the curves along the horizontal axis reflect the overall similarity between each pair of pools and hence the similarity of the libraries from which they were extracted. For example, the leftmost shifted curve corresponds to distances between external records from the ACD library and internal records from the NCI library. With the present means of measuring distance, then, one would conclude that the ACD and NCI libraries are the most similar pair in the collection. On the other hand, distances between external ACD records and target CMC records give rise to the rightmost shifted curve, suggesting that these two libraries are the most dissimilar. Distances between external CMC subsets and remaining CMC target records do not appear most similar. This shows that fairly diverse records comprise the CMC pool, despite the fact that all are classified as drugs. Distances between external CMC subsets and the internal (NCI) library records appear shifted slightly to the left of the curve between CMC external and target records. This is possibly an artifact of the differing size of libraries compared (10 000–1 000 vs 4 487–1 000) and statistical sampling and is of no consequence to our results.

In choosing a reasonable threshold distance, we first note that all distributions peak at around 45–50 bits and that only a small fraction of pairwise distances are less than about 15 bits. This indicates that a threshold distance of 50 would, for a typical external record, incorporate contributions to the local density from about half of all target and internal pool records. Such a large threshold would appear to be in conflict with the concept of *local* density. At the other extreme, threshold distances smaller than about 15 bits might result in little or no density being found for a significant fraction of external records that are too far removed from any

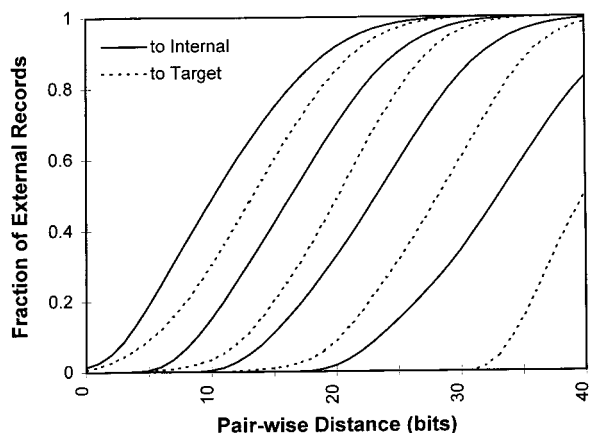


Figure 3. Nearest-neighbor counts from external pool records to differing numbers of internal (solid lines) and target (broken lines) records. Each series of four curves denotes, from left to right, the fraction of external records that are within a given pairwise distance of 1, 10, 100, and 1 000 internal or target records. For example, ~50% of external records are within 10-bit distance of at least one internal record, and ~30% are within this distance of a target record; ~15% have at least 10 internal records within 10 bits, and <5% have 10 or more target records within this distance.

target and internal pool records to calculate meaningful scores; all such records receive identical scores of zero.

Figure 3 addresses these issues in a somewhat different way. For each distance between 0 and 40 bits, the fraction of external pool records that are located within that distance of any 1, 10, 100, or 1 000 internal records (solid lines) and target records (dashed lines) is plotted. For example, about 50% of external records are within 10 bits of some internal record, while only about 30% are that close to any target record. This leaves substantial fractions (50% and 70%) of the external records which would sample no internal or target pool density if a threshold of 10 bits was employed. On the other hand, every external record is within 30 bits of some member of the internal or target pools, and most have at least 100 neighbors within that distance. A threshold of 30 bits, therefore, will assign at least some internal and target density for each external record. Overall, Figures 2 and 3 suggest that a threshold of somewhere between 20 and 30 bits should provide sufficient local density to make an informed assessment of each external record while limiting the amount of information provided by members of the internal and target pools which are rather dissimilar to the external record being scored.

Variation of Threshold Distance and Attenuation Function. Figures 4 and 5 illustrate how rapidly the drug-like CMC records are extracted from spiked external pools when threshold distances of 10, 20, 30, and 40 bits are used, along with a window size of 10 records and a Gaussian attenuation function. Figure 4 shows average results for the first 1 000 records selected, and Figure 5 shows the corresponding behavior over the course of selecting all external records. The uppermost lines denote perfect performance, where all spiked CMC records are selected before any ACD records, and the lower lines indicate the rate of CMC record selection that would be expected by chance.

LASSOO clearly biases selection to favor the drug-like CMC compounds, with up to 5 times as many CMC

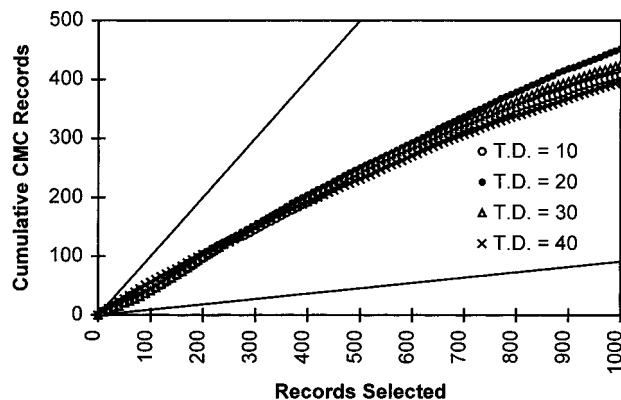


Figure 4. Cumulative number of CMC records selected with different threshold distances for the first 1 000 external records selected. The upper straight line delimits the maximum possible CMC count, and the lower line represents the rate of random expectation. Curves depict average results (10 spiking experiments each) for threshold distances (T.D.) of 10, 20, 30, and 40 bits using a Gaussian attenuation function and window size of 10 records.

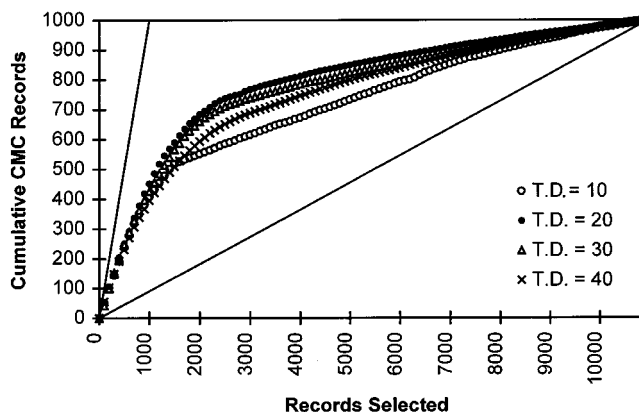


Figure 5. Cumulative number of CMC records selected with different threshold distances over the course of selecting all external records. Except for the range of records shown, this plot is the same as Figure 4.

records selected in the first 1 000 records as compared to chance. A threshold distance of 20 bits appears to provide the best overall enrichment, though for the first few hundred records selected, all threshold distances tested yield similar results. Extending the threshold to 30 then 40 bits progressively degrades performance, and reducing to 10 bits is clearly detrimental after about 1 500 records have been chosen. This same trend is repeated regardless of the attenuation function or window size used.

A conspicuous feature of the curve associated with a 10-bit threshold distance (Figure 5) is the near linear stretch between about 1 500 and 6 000 records selected. This corresponds to the simultaneous selection of approximately 4 700 records which have identical scores of 0. All of these records are more than 10 bits away from any target or internal record, so they cannot be assessed as either target-like or internally redundant. Records selected prior to this block have positive scores and thus are within 10 bits of at least some target records. Records selected after this block may or may not be within the threshold distance of target records, but they are close enough to some internal records to be assigned an overall negative score. Thus, even for the pathologic case of a restrictively small distance

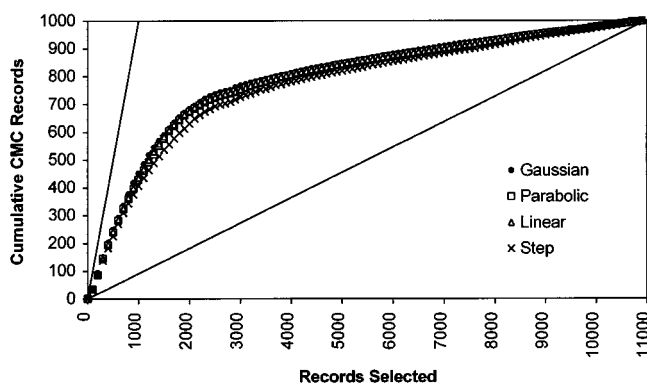


Figure 6. Cumulative number of CMC records selected with different attenuation functions or no attenuation function. Curves depict average results (10 spiking experiments each) for threshold distance of 20 bits and window size of 10 records.

threshold, LASSOO first selects drug-like records which are distinct from existing members of the internal pool, then records with ambiguous characteristics, and finally records which are decidedly similar to internal library members.

Attenuation functions which decay smoothly with distance reduce the influence of target and internal pool records as they become more dissimilar to the external record being scored. They provide a smooth transition from important to irrelevant as record distance moves from 0 toward the threshold value. Figure 6 shows how different attenuation functions influence the recovery of CMC records from the spiked pool when a threshold distance of 20 bits and a window size of 10 records are used. Employing any of the smooth attenuation functions improves performance relative to using an abrupt steplike cutoff, but there is little difference among smooth functional forms. We note that similar trends are seen regardless of which distance threshold is employed. As the Gaussian function appears to be marginally better than the others, we consider this in all further examples.

Variation of Window Size. Window size dictates how many favorably ranking records are transferred into the internal pool at each iteration and therefore how frequently internal pool density is updated to reflect acceptance of external records. Because there is no updating until the end of the iteration, a large window size can lead to many self-similar records being chosen at once. With a smaller window, internal pool density is updated more frequently, so that overpopulated regions in the internal pool get translated to external record scores in a more timely fashion. In the limit, a window size of 1 should lead to near-optimal, nonredundant record selection. Practical considerations, however, favor a large window size, as this translates to fewer iterations and faster run-times. Choosing an acceptable window size thus represents a tradeoff between minimizing redundancy in the records selected and minimizing the time required to carry out the procedure.

As it turns out, window size has very little effect on the rate at which CMC records are recovered from spiked libraries. Window sizes of 10, 30, and 100 records yield curves that are virtually identical to those in Figures 4 and 5, where a window size of 10 was used. Performance is slightly poorer with a window size of

1 000 and poorer still with an infinite window (i.e. a single iteration). For example, after selection of 2 500 records, the average numbers of CMC records recovered for window sizes of 10, 30, 100, 1 000, and infinite are 737.1, 738.4, 738.2, 731.3, and 713.9, respectively (10 test runs, standard deviations of 10 or less in all cases). Large window sizes mean that the algorithm gets no feedback until many records are accepted, and the marginally poorer performance stemming from large (and infinite) window sizes may be attributed to this lack of feedback. In terms of recovering drug-like records, a window size of about 100 seems adequate with the present system.

Variation of Pool Size. Target and internal pools will generally differ in size, either initially or during acquisition of external records, so it is important to investigate how such differences might influence algorithm performance. Results of spiking experiments described above employed initial target and internal pools differing in size by about 2-fold (4 487 target records, 10 000 internal records). For comparison, additional spiking experiments were run using equal-sized target and internal pools of 4 487 records and pools that differed in size by more than a factor of 4 (4 487 target records, 20 000 internal records). The spiked external pools for these tests were the same as before, with 10 000 ACD records and 1 000 CMC records. Using a Gaussian attenuation function and a window size of 10 records, these experiments showed virtually no change in the recovery rates of CMC records when threshold distances of 20, 30, and 40 bits are employed. With a threshold of 10 bits, however, the number of external records that receive tie scores of 0 decreases with increasing initial internal pool size. This is because a more heavily populated internal pool tends to have fewer gaps in the descriptor space, so external records are more likely to have internal record density within 10 bits. As a consequence, records that might otherwise get tie scores of 0 are more frequently differentiated, and results improve very slightly when larger internal pools are employed. With a more suitable threshold distance, however, it appears that widely varying pool sizes have little impact on algorithm performance.

Diversity of Records Selected. Monitoring the recovery of CMC records from spiked libraries illustrates how LASSOO can be used to preferentially extract drug-like compounds, but another distinguishing feature of the algorithm is its ability to maintain diversity in the internal library. As the selection process is carried out, we measure diversity by examining the average distance between each internal record and its nearest neighbor in the same pool. This average minimum distance is a measure of typical inter-record spacing, with large values occurring for pools containing well-separated records and smaller values occurring when more tightly clumped clusters of records are present. Curves showing average minimum distance as a function of the number of records selected are shown in Figure 7. The different curves correspond to random selection (dashed line), normal LASSOO selection (heavy line), the use of only target pool density when scoring (lower line), and the use of only internal pool density when scoring (upper line). Results represent an average of 10 spiking experiments, utilizing a threshold distance

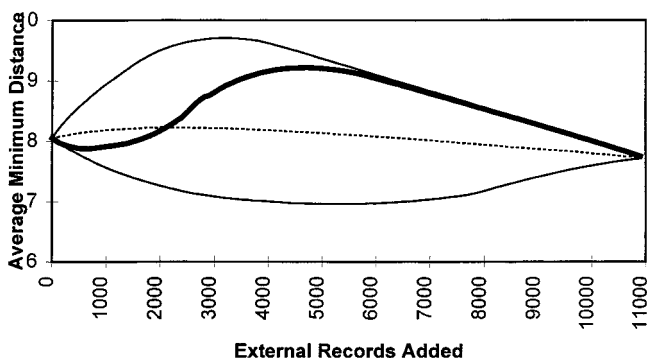


Figure 7. Average minimum distance between internal pool records plotted against the number of external records selected (i.e. added to the internal pool). The heavy curve depicts normal algorithm behavior, and the dotted curve corresponds to random record selection. The uppermost curve corresponds to the situation where only internal pool records factor into scores. Average data from 10 spiking experiments with a threshold distance of 20 bits, a Gaussian attenuation function, and a window size of 10 records are shown.

of 20 bits, a Gaussian attenuation function, and a window size of 10 records.

When only internal pool density is used for scoring, there is no bias in favor of target-like records, and diversity becomes the sole means of ranking external records. The highest-ranking records in this case are relatively far from internal pool members, and over the course of selecting these records, the average minimum distance in the internal pool increases from about 8 bits to more than 9.5 bits (upper curve). As progressively more external records are accepted, gaps between internal records begin to be filled in and the average inter-record spacing drops.

When considering only target pool density (lower curve), the driving force becomes purely that of selecting target-like records. This results in a pronounced lowering of diversity in the internal pool, as the highest-ranking records will be those located within dense, target-like clusters. These tightly clumped collections, which frequently comprise sets of analogues from the CMC pool, are preferentially extracted and placed in the internal pool, lowering the average pairwise record separation. Eventually, the target-like clusters are exhausted and the remaining external records, which offer more diversity by default, are selected.

When both internal and target pool densities are used in scoring, i.e., the normal mode of operation, average record separation drops slightly at first and then climbs in a fairly rapid fashion until diversity coincides with what was obtained when only internal density was considered (heavy curve). This sort of intermediate behavior, combined with the results in Figures 4 and 5, helps to illustrate the intended purpose of LASSOO, that is, balancing quality and diversity.

Scores and Example Ranked Compounds. Output from LASSOO consists of a ranked list of external records, their overall scores, and the separate internal and target pool score components. From the ranked records, the corresponding library compounds may be chosen directly, or in the case where several indistinguishable compounds are associated with a single record, additional criteria such as price, stability, or ease of synthesis may be applied to complete the decision.

Figure 8 contains compounds corresponding to several high-, medium-, and low-ranking external records from one of the spiking experiments. These compounds serve as examples for illustrating some of the underlying trends in the score distributions.

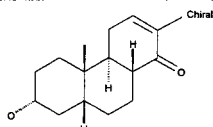
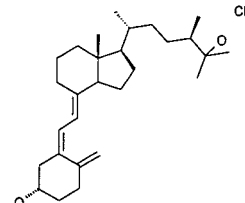
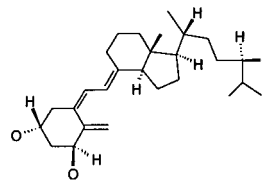
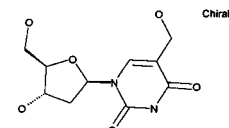
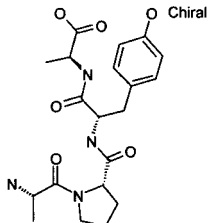
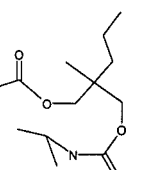
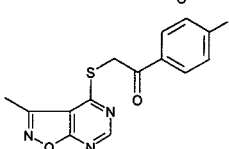
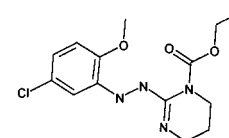
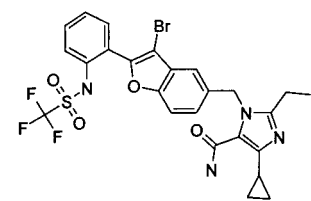
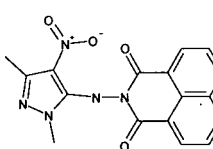
Figure 9 contains plots of overall and component scores as a function of rank for the first 500 records selected in the single spiking experiment. Steroids comprise 49 of the first 50 records, with the CMC portion of the external pool contributing 20 of these steroid records. For all of these highest-ranking records, both target and internal score components are significant, indicating an abundance of steroids in both of these pools. The high-low separation of score components seen with the highest-ranking records is primarily a consequence of the different pool size normalization factors; i.e., the steroids already present in the internal pool represent a smaller fraction of the total records compared to the steroids in the target pool.

A high density of steroids in, for example, the full CMC pool (5 487 records) may be confirmed by examining testosterone. There are seven CMC compounds with MDL key representations that are identical to testosterone (one record in the CMC pool) and a total of 43 records in this pool that differ by 5 or fewer bits from testosterone. Since the target pool is based on CMC compounds, this tight localization of steroid records dominates the selection process at the beginning, with 46 steroid records being extracted before the first nonsteroid compound is encountered. During these early stages, even the nonsteroid compounds exhibit many structural characteristics of true steroids (Figure 8, group A). While the selection of redundant compounds appears to be at odds with the desire for diversity in the internal library, it is difficult to avoid when the target pool contains such a strong bias toward any one class of compound. Use of a target pool with a sparser sampling of these dense clusters should reduce this type of undesirable behavior.

After about 100 records have been selected, a more balanced mix of compound classes begins to be appear. As the selection of the first 500 proceeds, there is a general trend toward fewer and fewer records with large score components. Records with weak target components and still weaker internal components (Figure 8, group B) are eventually selected, not so much because of target-like properties but because they increase the diversity of the internal pool. Beyond the first 500 records, the general pattern of decreasing score component size continues until rank 2494, where a block of 401 records with tie scores of 0 begins. These ambiguous compounds, e.g., Figure 8, group C, are more than 20 bits away from anything in the target or internal pools (for this spiking experiment) and serve only to increase diversity.

Figure 10 contains overall and component scores plotted against rank for the lowest-scoring records in the spiking experiment. Many of these records have target pool scores of 0 and internal pool scores that are quite high, indicating that (1) they are not drug-like and (2) they are highly redundant with other members of the internal pool. Examples compounds are shown in Figure 8, group D.

The very last records selected have non-zero target

| | Structure | Name | Rank | ACD | CMC | Score | Target | Internal |
|---|---|----------------------------|------|------|-----|--------|--------|----------|
| A |  | MFC00198928 | 46 | 27 | - | 4.855 | 9.943 | 5.088 |
| A |  | MFC00210914 | 54 | 33 | - | 4.338 | 7.216 | 2.879 |
| A |  | TACALCITOL [INN] | 57 | - | 23 | 4.276 | 7.455 | 3.180 |
| B |  | MFC00047489 | 221 | 118 | - | 1.210 | 1.340 | 0.130 |
| B |  | MFC00237760 | 306 | 154 | - | 0.868 | 0.942 | 0.074 |
| B |  | CARISOPRODOL [U;INN] | 339 | - | 168 | 0.753 | 0.844 | 0.091 |
| C |  | MFC00102687 | * | | | 0 | 0 | 0 |
| C |  | MFC00177855 | * | | | 0 | 0 | 0 |
| C |  | SAPRISARTAN [U;INN;BAN] | * | | | 0 | 0 | 0 |
| D |  | MFC00113208 | 6154 | 5259 | - | -0.075 | 0.008 | 0.083 |

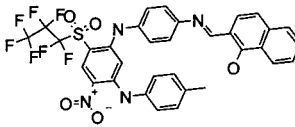
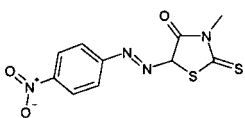
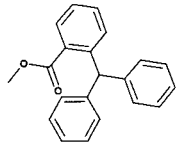
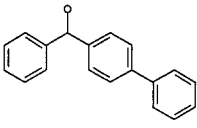

| | Structure | Name | Rank | ACD | CMC | Score | Target | Internal |
|---|---|---------------------|-------|------|------|--------|--------|----------|
| D |  | MFC00196684 | 6335 | 5434 | - | -0.083 | 0.000 | 0.083 |
| D |  | NITRODAN [U;INN] | 6415 | - | 903 | -0.088 | 0.000 | 0.088 |
| E |  | MFC00027557 | 10947 | 9948 | - | -5.817 | 3.216 | 9.033 |
| E |  | MFC00183351 | 10969 | 9970 | - | -7.072 | 3.294 | 10.366 |
| E |  | BENZOIC ACID [U] | 10973 | - | 1000 | -7.661 | 5.673 | 13.334 |

Figure 8. Example CMC and ACD compounds corresponding to high-, medium-, and low-ranking records from a spiking experiment which employed a threshold distance of 20 bits, a Gaussian attenuation function, and a window size of 10 records. Structures and names are as extracted from the source databases. Overall ranking and the rankings within either the CMC or ACD subsets of records are shown, along with overall scores and separated score components; for presentation purposes, score values are multiplied by 1 000. Group A compounds correspond to the highest -ranking, nonsteroid records. Group B compounds are not particularly target-like but are dissimilar to internal records and so increase diversity. Group C compounds are too far (i.e. >20 bits) from any target or internal records to be fairly assessed and so receive tie scores of 0; these records correspond to rankings 2494–2894. Group D compounds are not target-like yet somewhat similar to internal records and thus rank poorly. Group E compounds are somewhat target-like but also have many similar internal records making them redundant; these records rank last.

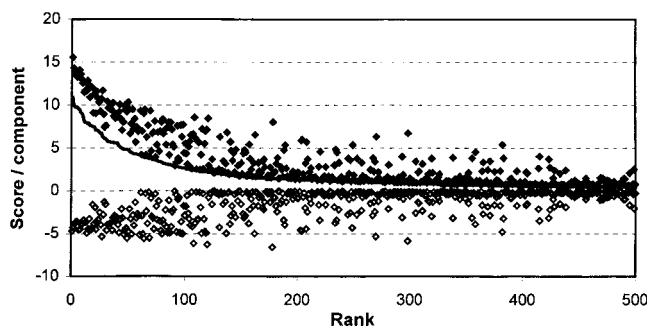


Figure 9. Overall scores (solid line) together with separated target (solid diamonds) and internal score (open diamonds) components for the highest-ranking 500 records from a spiking experiment with a threshold distance of 20 bits, a Gaussian attenuation function, and a window size of 10 records. Scores and component values are multiplied by 1 000 for ease of presentation.

terms, indicating that they are somewhat drug-like, but their exceedingly large internal scores suggest that they also contain highly redundant structural motifs. As illustrated by the example compounds in Figure 8, group E, the last records selected probably contain a subset or composite of moieties commonly found in drugs, but the overall structures are too similar to previously chosen ones to add much uniqueness to the library.

Conclusions

We have demonstrated an effective means of prioritizing compound selection for augmenting an existing

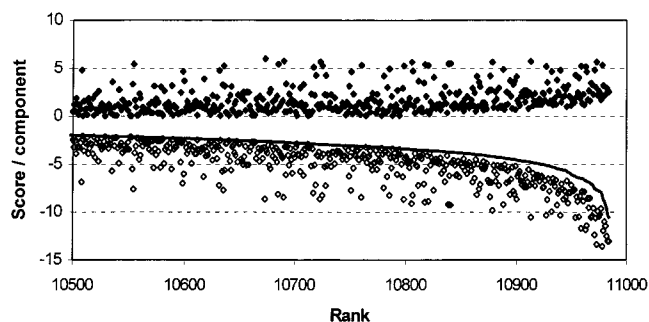


Figure 10. Overall scores (solid line) together with separated target (solid diamonds) and internal score (open diamonds) components for the lowest-ranking 500 records from the same experiment as shown in Figure 9. Scores and component values are multiplied by 1 000.

chemical screening library. The LASSOO algorithm is conceptually simple and depends on only a few parameters. Selection of compounds requires no a priori rules for evaluating what is “good” and what is “bad” but, instead, relies on a target library that automatically encodes a wealth of information about desirable characteristics. At the same time, the composition of the existing library is taken into consideration to avoid addition of chemically redundant compounds.

One critical program parameter is the threshold distance that is used to compute library population densities around candidate compounds. For a system utilizing MDL keys and a city-block distance, a threshold distance of 20 bits performs best, and there is

relatively little sensitivity to other algorithm parameters. Use of a smooth attenuation function improves performance slightly, but there appears to be little difference among the functional forms investigated here.

LASSOO is a general method, and in principle, any set of chemical descriptors may be combined with an appropriate distance function to operate on any collection of compounds which encode desirable or undesirable characteristics. Thus, members of a commercial, proprietary, or virtual chemical library may be evaluated for purchase, use, or synthesis to augment an existing screening library. In addition, a library may be designed from scratch simply by running the algorithm without specifying an initial internal library.

The LASSOO methodology is expected to increase the hit rates of desirable drug-like compounds by enabling the design of diverse, high-quality libraries that are relatively small by today's standards. This capability takes on increasing importance as the number of potential targets grows and researchers come to the realization that screening huge libraries is neither an efficient strategy nor a sufficient means for discovering leads.¹⁷ Diverse libraries enriched in drug-like compounds should supply hits in a variety of assays and make the process of going from a lead to a pharmaceutical candidate more rapid and successful.

We are currently exploring the extension of these techniques to the case of negative reference libraries, which could be composed of any collection of compounds with undesirable properties. Inclusion of a score term that penalizes selection of compounds similar to those in a negative reference pool should then disfavor selection compounds with undesirable attributes.

References

- (1) Gordon, E. M. Libraries of Non-Polymeric Organic Molecules. *Curr. Opin. Biotechnol.* **1995**, *6*, 624–631.
- (2) Ferguson, A. M.; Patterson, D. E.; Garr, C.; Underiner, T. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screen.* **1996**, *1*, 65–73.
- (3) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431–1436.
- (4) Martin, E. J.; Critchlow, R. E. *J. Comb. Chem.* **1999**, *1*, 32–45.
- (5) Gibbon, J. A.; Taylor, E. W.; Braeckman, R. A. In *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*; Gordon, E. M., Kerwin, J. F., Eds.; Wiley-Liss: New York, 1998; pp 453–474.
- (6) Ghose, A. K.; Viswanadhan, A. K.; Wendolowski, J. J. A Knowledge Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Know Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55–68.
- (7) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeny, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Drug Delivery Rev.* **1997**, *23*, 3–25.
- (8) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between “Drug-like” and “Nondrug-like” Molecules? *J. Med. Chem.* **1998**, *41*, 3314–3324.
- (9) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325–3329.
- (10) The ACD, NCI3D, and CMC3D databases are commercially available from MDL Information Systems, Inc., 14600 Catalina St, San Leandro, CA 94577.
- (11) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (12) *ISIS™/Base 2.1.4*; MDL Information Systems, Inc., San Leandro, CA.
- (13) Brown, R. D.; Martin, Y. C. Use of Structure–Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (14) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand–Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (15) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.
- (16) Dixon, S. L.; Koehler, R. T. The Hidden Component of Size in 2D Fragment Descriptors: Side-Effects of Sampling in Bioactive Libraries. *J. Med. Chem.* **1999**, *42*, 2887–2900.
- (17) Dixon, S. L.; Villar, H. O. Bioactive Diversity and Screening Library Selection via Affinity Fingerprinting. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1192–1203.

JM990312G