Designing Compound Subsets: Comparison of Random and Rational Approaches Using Statistical Simulation

Siew Kuen Yeap,*,† Rosalind J. Walley,‡ Mike Snarey,† Willem P. van Hoorn,† and Jonathan S. Mason†

Pfizer Global Research and Development, Sandwich, Kent CT13 9NJ, UK

Received August 30, 2006

Compound subsets, which may be screened where it is not feasible or desirable to screen all available compounds, may be designed using rational or random selection. Literature on the relative performance of random versus rational selection reports conflicting observations, possibly because some random subsets might be more representative than others and perform better than subsets designed by rational means, or vice versa. In order to address this likelihood, we simulated a large number of rationally designed subsets for evaluation against an equally large number of randomly generated subsets. We found that our rationally designed subsets give higher mean hit rates compared to those of the random ones. We also compared subsets comprising random plates with subsets of random compounds and found that, while the mean hit rate of both is the same, the former demonstrates more variation in the hit rate. The choice of compound file, rational subset method, and ratio of the subset size to the compound file size are key factors in the relative performance of random and rational selection, and statistical simulation is a viable way to identify the selection approach appropriate for a subset.

INTRODUCTION

Corporate compound file collections are traditionally screened in their entirety for lead matter; however, as the number of compounds increases with advances in library chemistry, knowledge-based screening that provides time- and cost-effective-enabling approaches becomes more valuable.[1] Where there is at least one chemical lead, ligand-based virtual screening may be applied to identify compounds of similar structure or properties for screening. Where an experimental X-ray or NMR structure or 3-D model of the target is available, structure-based virtual screening via high-throughput docking may be used to prioritize candidates that are likely to bind at that target. In the absence of either a suitable chemical lead or enzyme/receptor structure, or when neither ligand-based nor structure-based virtual screening yields a suitable lead, an option is to screen a diverse subset representation of the file in order to produce "seed" leads to be followed up iteratively using virtual screening, as illustrated in Figure 1. Even when the structure of the ligand or target is known, a subset provides a compound-number-effective way to explore the chemical space of the screening file to identify novel ligands while maintaining serendipity in the process.

The success of file subsets is dependent on the ability to limit the compounds to be subsetted to those most likely to become suitable hits and the means to attain representative coverage of these compounds. The first stage is gained via property and structural filters, such as the rule-of-five[2] and further criteria for "drug-likeness",[3] while the second stage may be achieved via algorithms for clustering or partitioning,[4] or simply by selecting the desired number of compounds at random. Literature reports are available on the relative performance of random versus rational selection, however with conflicting observations. Mathematical modeling suggested that, except for very large sampling ratios, randomly selected compounds cover as much space as rationally selected compounds to the extent that the approaches appear equivalent in many cases.[5] In contrast, maximum diversity methods and hierarchical clustering were shown to enhance chemical diversity for three different databases. These selections were only matched by random selections of at least 3.5 times as many compounds.[6] In another instance in favor of rational design, k-means sampling of AMSOL, VolSurf, and Almond descriptors attained diversity that was matched by random samples twice the size.[7] One explanation for the range of observations might be the variations in the quality of the subsets under investigation. Random selections are inherently random by nature—some subsets would be more representative than others and may perform better than subsets designed by rational means and vice versa. Likewise, methods for rational subsets might have a random element; for instance, random selection within clusters and different rational subsets selected by the same method may perform differently. Results may also vary for different compound files, different rational subset selection methods, and different sets of assays. A "rational diverse/representative" approach is thus often used to produce a consistently "good/average" random pick. A true comparison for a particular rational subset selection method for a single compound file would require the generation of a large number of randomly generated subsets in order to establish how random selection would perform on average, and the generation of an equally large number of rationally designed subsets to determine their

* Corresponding author e-mail: kuen.yeap@pfizer.com.
† Medicinal Informatics, Structure and Design.
‡ Statistical Applications.

10.1021/ci600382m CCC: $37.00
Published on Web 10/06/2007

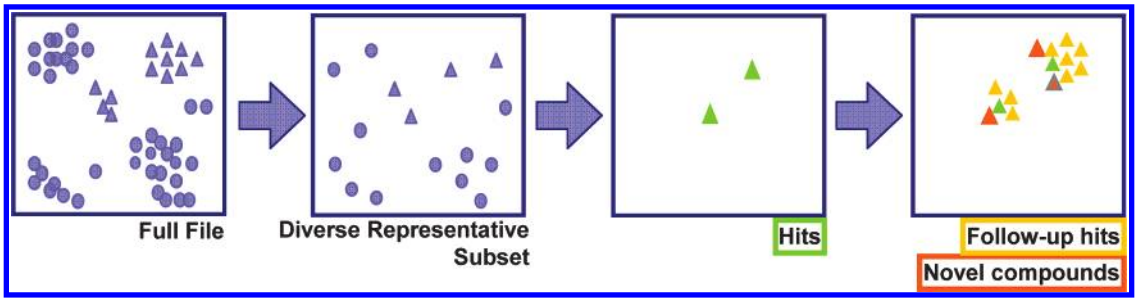

Figure 1. Subset construction. From the full file, the subset is constructed as a once-off set for screening against targets of interest. The hits found in the subset for a particular target might be few; however, virtual screening using the hits as seeds is likely to identify further crops of hits, which might give valuable structure−activity relationship data and offer more potent and attractive compounds. Further followup of the hits using medicinal chemistry can lead to more novel compounds.

relative performance. This is possible using Monte Carlo simulation.

Simulation is a technique widely used to predict outcomes where it is not straightforward to reach the results analytically, for example, in operational research modeling production processes, computational physics, and integration and optimization methods. When the prediction is not deterministic, that is, there is some random element, the simulation is often referred to as a Monte Carlo simulation.[8,9] Where the system to be modeled includes a random component, it is important to consider the impact that changes in this component have on the outcome. Simply looking at the average outcome or restricting analysis to one realization of the random component may be misleading. In the drug discovery process, simulated trial design is used increasingly, thereby providing an *a priori* assessment of the likelihood of various outcomes of increasingly complex clinical trial designs and allowing the selection of the optimal trial design.

The selection of a subset of a compound file can be pragmatically dealt with in two parts: first, the choice of a subset selection method, or a small group of subset selection methods, and then a comparison of candidate subsets selected by these methods. Advantages of dividing the selection process into these two parts include the following:

- An explicit compound selection method provides a rationale for compound replacement. Confidence in the subset selection method rather than just the particular selected subset translates into continued confidence in the subset as compounds run out and have to be replaced.
- A subset selection method is not dismissed because of an "unlucky" choice of subset nor chosen because of a particular "lucky" choice of subset.
- The exact compound file is less critical to the choice of the subset selection method, so it can be carried out less frequently and using a slightly out-of-date compound file.
- Using the strategy of deciding the subset selection method before choosing candidate subsets may reduce the number of candidate subsets selected. Since the simulation for the subset selection method runs comparatively fast, this is expected to reduce the computer processing time required.
- Knowledge of the relative merits of different subset selection methods may lead to further understanding of the distribution of hits in the compound file, thus potentially allowing for further optimization.

This paper focuses on an assessment of subset selection methods rather than on the assessment of particular candidate subsets. It not only addresses expected numbers of hits but also the variability of each subset selection method. The methodology is presented in terms of numbers of hits but, equivalently, could be presented as hit rates and also can be adapted to work with numbers of lead series rather than hits (as long as each compound in the compound file has been assigned to a single series). We describe the design of a historical rational subset and then simulate subsets constructed in this manner along with random compound subsets and random plate subsets. We show that rationally selected subsets are likely to yield more hits than random selections. Also, we show that, although random plate subsets generate the same mean number of hits as random compound subsets, the variation in numbers of hits is increased. We demonstrate the value that statistical analysis adds to subset design, to the extent that it has been incorporated into the design of subsequent screening subsets.

METHODS

Design of the Historical Rational Subsets. A cursory look at any screening file shows that many compounds are not favored as leads.[1] Obvious examples include compounds with missing structures, those defined using Markush structures, or those which contain moieties such as heavy metals or large peptides. In addition, there are duplicate structures in the file, at least at the level of parent desalted structure. The first requirement was to remove such compounds, before refining "lead-likeness" of the remaining file via the application of appropriate property and structural filters and finally selecting a diverse representation that is biased toward actives, as illustrated in Figure 2.

Property and Structural Filters. Small ligands tend to engage in fewer binding interactions and exhibit lower activity in standard high-throughput screening (HTS) protocols. Since their activity is detected most successfully using assays adapted to handle higher concentrations, compounds with a molecular weight of less than 150 were excluded.[10] The desired fate for an attractive HTS hit is lead optimization, a process that tends toward structures with increased molecular weight and lipophilicity.[11] Typical clogP and molecular weight limits for leads are ≤ 4 and ≤ 450, respectively; however, the requirement to comply with both limits have been found to neglect on average 58% of known actives, increasing to 80% for peptidic G-protein-coupled receptor (GPCR) targets, while compliance with one of the limits as long as the other is within clogP ≤ 4.5 or molecular weight ≤ 500 captured a higher proportion of actives, particularly with the peptidic GPCR targets.[12] The latter more generous Ro4.5 criteria were adopted for the subset.

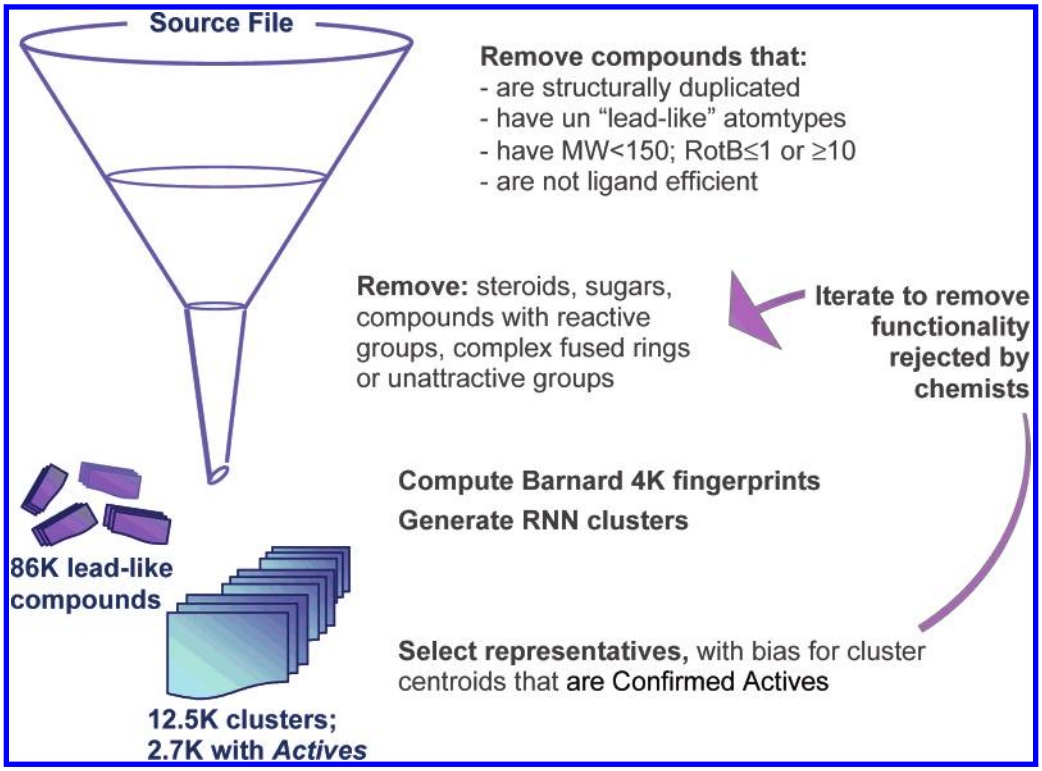

Figure 2. Filtering of the source file. The portion of our file used in this study, termed the source file, was first filtered to remove several undesired categories of compounds. The remaining 86 000 compounds were divided into 12 500-compound clusters, in order to select representatives that are biased toward actives.

Compounds with lower molecular flexibility tend to have better oral bioavailability;[13] therefore, ligand efficient compounds make better leads.[14] A rotatable bond count of $1 \leq \text{RotB} \leq 10$, a ligand efficiency estimate of ≤ 17 heavy atoms (based on an optimal interaction of $\sim$0.4 kcal/atom for a target activity of 10 μM), and proprietary substructure filters (for reactive/toxic groups etc.) were imposed on the source file. Finally, in several iterations guided by chemists' intuition, compounds with other generally undesired groups, comprising steroids, sugars, additional complex fused rings, and reactive moieties, were removed. From the portion of our screening file used for this study, which we term the source file, we obtained a filtered collection of 86 000 compounds deemed suitable for lead optimization.

Although the 86 000-compound file was used for the selection of rational subsets and corresponding random compound subsets, this file was not suitable for the selection of random plate subsets. Very few plates were sufficiently filled with compounds from the 86 000 file, as most had been tested on plates alongside other compounds. Thus, for the selection of random plate subsets, the source compound file was used. For consistency, random compound subsets for comparisons with these random plate subsets were also picked from the source file.

Structural Clustering. An established drug design premise is that structurally related compounds tend to possess similar biological properties. A member of a compound cluster that is active suggests that associated members have a fair chance of being active against the same or similar target. As a result, subset screening is at its most effective when the structure that represents the cluster is the member most likely to exhibit activity.

Clustering was conducted on the Barnard 2-D fingerprints using the Daylight Reciprocal Nearest Neighbor algorithm.[15]

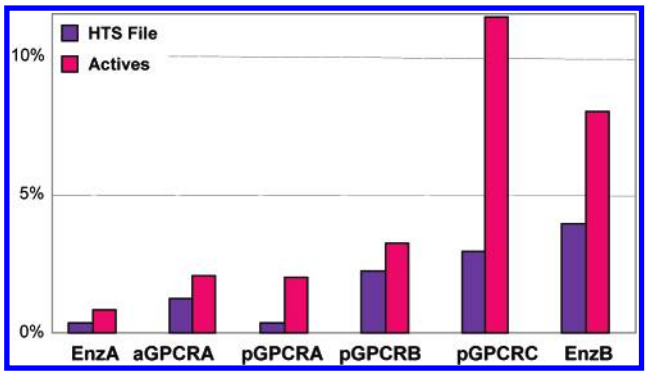

Figure 3. Hit rates determined for the full HTS data set compared to hit rates specifically for confirmed actives. Values for the latter were vastly improved.[16−18]

This gave 12 500 clusters that were used in the simulations of rational and random subsets described later. For the rational subset that was implemented (rather than simulated), however, the centroids were not always selected to represent the clusters. Preference was given to the confirmed active closest to the centroid, on the premise that hit rates for the six historical data sets examined are much higher when confined to confirmed actives, implying that confirmed actives have a greater likelihood to exhibit biological activity (see Figure 3).[16,17] For the 12 500 representative structures that yield a diversity coverage of nearly 15% of the 86 000 filtered compound collection, nearly 2700 are confirmed actives.

Historical Screen Data Used. The data used in the simulation came from five HTS screens of varying targets as summarized in Table 1. All compounds have been tested at least once, with hits defined as those confirmed to exhibit a percentage inhibition above a specified threshold. In the

Table 1. Summary of HTS Data Used in Simulations

screen label[18]	number of compounds used rational subset study	number of compounds used in random plate subset study
aGPCRA	47 746	
pGPCRC	48 761	
EnzA	36 368	
EnzC	38 581	228 753
EnzD	50 539	363 656

simulation work, we assume that the hits are true actives and the remaining compounds are the true inactives.

Investigation of Distribution of Actives in Clusters. Before any simulation work was conducted, the distribution of hits in the different sized clusters was investigated. This was achieved by grouping the compound file according to size of cluster and calculating a hit rate for each cluster (not presented). For simplicity of presentation, the compound file was also divided into two groups: compounds in "small clusters" of 15 compounds or fewer and compounds in "large clusters" of more than 15 compounds. Hit rates were calculated for each group.

Simulation Strategy. As previously indicated, the objective of the simulation was to assess different methods of subset selection rather than a comparison of particular subsets.

For this work, three types of subsets were considered:

- **random compound subsets**: picking compounds at random
- **random plate subsets**: picking plates at random
- **rational subsets**: picking one compound at random from each cluster computed for the historically defined 12 500-compound rational subset

Two comparisons are presented:

- **Random Compound Subsets versus Rational Subsets.** For each screen, the 86 000 filtered compound file was used. For a particular cluster, if none of the compounds in a cluster were screened, the cluster was excluded from the data set. Also for aGPCRA, pGPCRC, and EnzA, only clusters where the 12 500-compound historical subset had been tested were included. For each screen, all rational and random subsets will be the same size (i.e., have the same number of compounds), and this will equal the number of clusters for which there are some data. Due to differences in historical screening, subset sizes vary somewhat from screen to screen. For each screen, 1000 rational and 1000 random subsets were created.
- **Random Compound Subsets versus Random Plate Subsets.** For each screen, only 384-well plates with 350 or more compounds from the source compound file were used in this simulation. Random compound subsets were picked from the same set of compounds as the random plate subsets. Each random plate subset comprised 30 plates, and its partner random compound subset had exactly the same number of compounds. A total of 1000 such pairs were simulated. All subset pairs were not of identical size since each plate did not have exactly the same number of compounds.

Generation of Rational Subsets. To generate the rational subsets, the data were first summarized for each cluster using three summary statistics: the number of compounds in the cluster, the number of compounds screened, and the number of hits found. For each cluster, a virtual list of compounds was made by listing the actives followed by inactives. Note that compound names or structures were unnecessary, for example, for a cluster of five compounds with three actives (Table 2).

Table 2. Example of Data for a Cluster[a]

compound 1	compound 2	compound 3	compound 4	compound 5
active	active	active	inactive	inactive

[a] Here there are three actives and two inactives.

A compound was selected at random from the cluster by selecting a random integer, j, from the set {1, 2, ..., number of compounds in cluster}. If $j \leq$ the number of actives, then the selected compound was recorded as an active, whereas if $j >$ the number of actives, then the selected compound was recorded as an inactive. Figure 4 contains a simplified example of this.

To this point, this is exactly equivalent to drawing compounds randomly one per cluster from a list of compound names for each cluster. However, summarizing the data allowed the simulation to run much faster than if individual compound names were processed.

This approach could be extended to allow for compounds that have not been screened, by estimating the total number of actives in the cluster to be

$$\text{total number of actives in the cluster} = \frac{(\text{actives seen in cluster})}{\text{compounds tested in cluster})} \times \text{total number of compounds in cluster}$$

For the rational pick of just one compound for a cluster, this gives exactly the same probability that this compound is a hit as if we had limited the selection to the compounds that have been screened; so in this situation it is unnecessary. However, these estimates become important when picking multiple compounds for each cluster because the probability that the following compounds are hits will differ depending on whether this estimate is used or not. For example, imagine a cluster of 100 compounds, where one of the 10 compounds screened is a hit. If the first compound selected is a hit and the estimator is not used, the probability of selecting a second hit is zero. If the estimator is used, the probability of the second compound being a hit (given the first compound was a hit) is 9/99.

Generation of Random Compound Subsets. Compared with rational subsets and random plate subsets, it is very straightforward to select a random compound subset of any number of compounds. Indeed, a simulation is not absolutely necessary since the distribution of the number of the hits is described by a standard probability distribution, called the "hypergeometric distribution",[6] which is characterized by the number of actives, size of the subsets, and size of the file used. See Appendix 2 for details. However, in this instance, the simulation of random compound subsets was carried out since it was quick and easy and possibly more convincing to the nonstatistician.

Since the selection of random compound subsets is so straightforward, the random compound subsets to be used for the comparison with the rational subsets and random plate subsets were selected after the other subsets and set to be the same size as them.

	No. Actives	Random no.	Representation of cluster	Result
Cluster 1	2 out of 5	3	Active, Active, Inactive, Inactive, Inactive	Inactive
Cluster 2	0 out of 2	1	Inactive, Inactive	Inactive
Cluster 3	2 out of 4	2	Active, Active, Inactive, Inactive	Active
Cluster 4	1 out of 1	1	Active	Active
Cluster 5	1 out of 5	5	Active, Inactive, Inactive, Inactive, Inactive	Inactive

Figure 4. Example of generation of one rational subset from a compound file of five clusters. For each cluster, one random number is generated between 1 and *n* (the number of compounds in that cluster), and this indicates whether an active or inactive has been selected. As the simulation program works through the clusters of a subset, only the running total for the actives needs to be retained.

	No. actives left	Total no of compounds left	Random no.	Representation of compounds left in whole file	Result
1st compound	6	17	5	Active, Active, Active, Active, Active; Active, Inactive, Inactive, Inactive, Inactive; Inactive, Inactive, Inactive, Inactive, Inactive; Inactive, Inactive	Active
2nd compound	5	16	12	Active, Active, Active, Active, Active; Inactive, Inactive, Inactive, Inactive, Inactive; Inactive, Inactive, Inactive, Inactive, Inactive; Inactive	Inactive
3rd compound	5	15	8	Active, Active, Active, Active, Active; Inactive, Inactive, Inactive, Inactive, Inactive; Inactive, Inactive, Inactive, Inactive, Inactive	Inactive

Figure 5. Example of generating one random compound subset of three compounds from the same compound file of 17 compounds as in Figure 4. (Note that in actual simulation the random and rational subsets are the same size.) For the selection of each compound, one random number is generated between 1 and *n* (the number of compounds in the full file minus the number of compounds already selected). After each selection, the data set is reduced, ensuring that a compound cannot be selected twice. As the simulation program works through a subset, only the running total for the actives and the selected compounds need to be retained.

For the random compound subsets, the summary data set contained only the total number of compounds and the estimated total number of actives. For the simple example in Figure 4, the total number of actives is 6 and the total number of compounds is 17. If for rational screening, the total number of actives in a cluster were to be estimated, the same estimated numbers of actives should be used for the simulation of random compound subsets. Thus, the data set used to simulate random testing represents exactly the same compounds that are used for the rational subset selection. The data can then be represented as in the first row of Figure 5.

Initially, any compound can be selected. Once selected, it is not available for further selection. If an active is selected as the first compound (as exemplified in Figure 5), the number of actives reduces by one before the selection of the next compound. Similarly, if an inactive is selected (as in the selection of the second compound in Figure 5), the number of inactives reduces by one. Thus, a random integer, *j*, was selected from the set {1, 2, ..., available compounds}. If the $j \leq$ the number of actives, the selected compound was recorded as an active, whereas if $j >$ the number of actives, then the selected compound was recorded as an inactive. Numbers of "available compounds" and "available (in)-actives" were each reduced by one. The process was repeated until each subset was complete.

Generation of Random Plate Subsets. The data were summarized for each plate using three summary statistics: the number of compounds on the plate from the file, number of compounds screened, and number of hits found.

A total of 30 plates were selected from the summary data file at random by numbering the plates 1, 2, ..., number of plates and randomly selecting 30 numbers from the set {1, 2, ..., number of plates}. Since each plate included 350−360 compounds, all the subsets were approximately the same size.

For each random plate subset, a corresponding random compound subset was created. This random compound subset was selected from exactly the same list of compounds and was the same size as the corresponding random plate subset. Apart from the change in data set, the random compound subsets were constructed using the same method as for the

	Random no.	Representation of remaing plates					Total actives selected	Total compounds selected
1st plate	5	Plate A 360 compounds: 5 Actives 355 Inactives	Plate B 355 compounds: 20 Actives 335 Inactives	Plate C 350 compounds 0 Actives 350 Inactives	Plate D 358 compounds 5 Actives 353 Inactives	Plate E 353 compounds 13 Actives 340 Inactives	13	353
2nd plate	1	Plate A 360 compounds: 5 Actives 355 Inactives	Plate B 355 compounds: 20 Actives 335 Inactives	Plate C 350 compounds 0 Actives 350 Inactives	Plate D 358 compounds 5 Actives 353 Inactives		13 + 5 = 18	353 + 360 = 713
3rd plate	2	Plate B 355 compounds: 20 Actives 335 Inactives	Plate C 350 compounds 0 Actives 350 Inactives	Plate D 358 compounds 5 Actives 353 Inactives			18 + 0 = 18	713 + 350 = 1063
4th plate	1	Plate B 355 compounds: 20 Actives 335 Inactives	Plate D 358 compounds 5 Actives 353 Inactives				18 + 20 = 38	1063 + 355 = 1418

Figure 6. Example of generation of a random plates subset of four plates from a compound file of five plates. For the selection of each plate, one random number is generated between 1 and *n* (the number of plates in the full file minus the number of plates already selected). After each selection, the data set is reduced simulating the real situation that a plate cannot be selected twice. As the simulation program works through a subset, in addition to the running total for the actives and the running total for the compounds selected, the program needs to retain the information as to which plates have been selected.

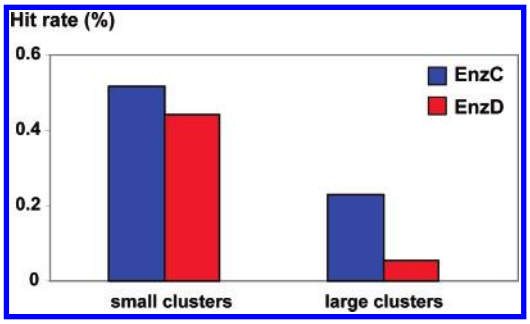

Figure 7. Hit rates for EnzC and EnzD compared for compounds from small clusters (≤15 compounds) and for compounds from larger cluster (>15 compounds). For both screens, hit rates for the smaller clusters outperform the larger ones.

random compound subsets corresponding to the rational subsets (see Figure 6 for an example of the generation of random plate subsets).

Assessment of Simulated Subsets. The results of each subset were summarized using the total number of actives found (equivalently, hit rates could be used). For each subset type and each screen, there were 1000 subsets and thus 1000 values for the total numbers of actives. These values were displayed graphically as histograms, which both indicate the typical number of hits that might be found with a subset and also what might result from a particularly "lucky" or "unlucky" selection.

Approximate normal probability distribution functions were overlaid on the histograms. Details as to how these were selected are outlined in the appendices, and these theoretical distributions and approximations are found in many standard statistical texts.[19]

The mean number of hits for each subset method was then calculated.

RESULTS

Distribution of Actives in Clusters. Figure 7 shows that for the two screens presented, the hit rate was higher in the smaller clusters with 15 compounds or fewer than those with more compounds.[20] There was a similar pattern in the three other screens. This figure suggests that subsets selected by picking more heavily from the smaller clusters will have higher hit rates than those that focus more on the larger clusters. In the compound file, approximately 10% of the clusters have more than 15 compounds, so each rational subset will have 10% of their compounds from these "large clusters." However, 22% of the compounds lie in these large clusters, so on average, each random subset will have 22% of its compounds from these large clusters. There will be slight differences screen to screen because entirely unscreened clusters are excluded from the simulation, and these vary from screen to screen. This consideration of hit rates alone suggests that the rational subsets may outperform the random subsets.

Comparison of Rational and Random Compound Subsets. Figure 8 contains histograms of the numbers of hits of each of the simulated subsets. The histograms show that overall the rational approach outperforms the random approach in terms of numbers of hits. However, the spread of the histograms indicates that there are some random compound subsets that outperform some rational subsets.

Table 3 shows that, for each of the five screens considered, the mean number of hits found for rational subsets was 16−55% larger than for random compound subsets.

Comparison of Random Compound and Random Plate Subsets. A comparison of the histograms in Figure 9 showed that the mean number of hits is the same for the random singleton and random plate subsets (as theory suggests) but that the histograms have a greater spread for the random plates. This suggests that hits cluster on plates. We believe this is most likely to be due to chemical similarities between some compounds on the same plate, although some experimental errors, for example, a blocked pipet, could create a group of apparent actives on a plate and give the same effect. For EnzD particularly, the normal distribution approximation appears to be comparatively poor, and this is primarily because the histogram appears to be skewed.

DISCUSSION

The growth of the Pfizer compound file has led to the need to identify subsets of compounds for use where it is

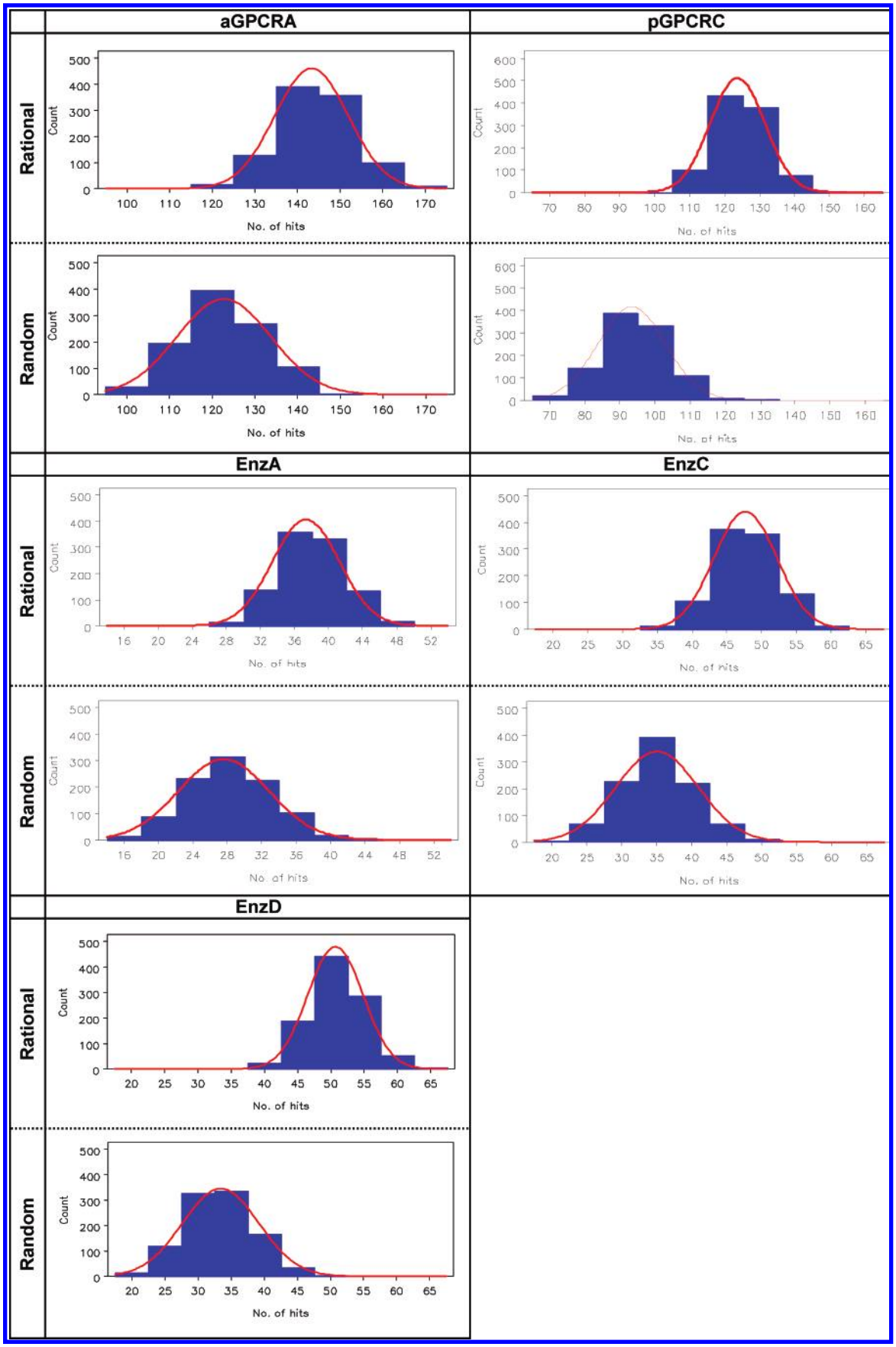

Figure 8. Histograms comparing random compound subsets and rational subsets.

not feasible or desirable to screen all compounds, whether prepared using singleton or combinatorial chemistry methods. The nonfeasibility or desirability may be due to a time constraint, for instance, the urgent need for a tool or lead compounds for a target that does not have an assay ready for large-scale HTS, or a practical one such as the supply or

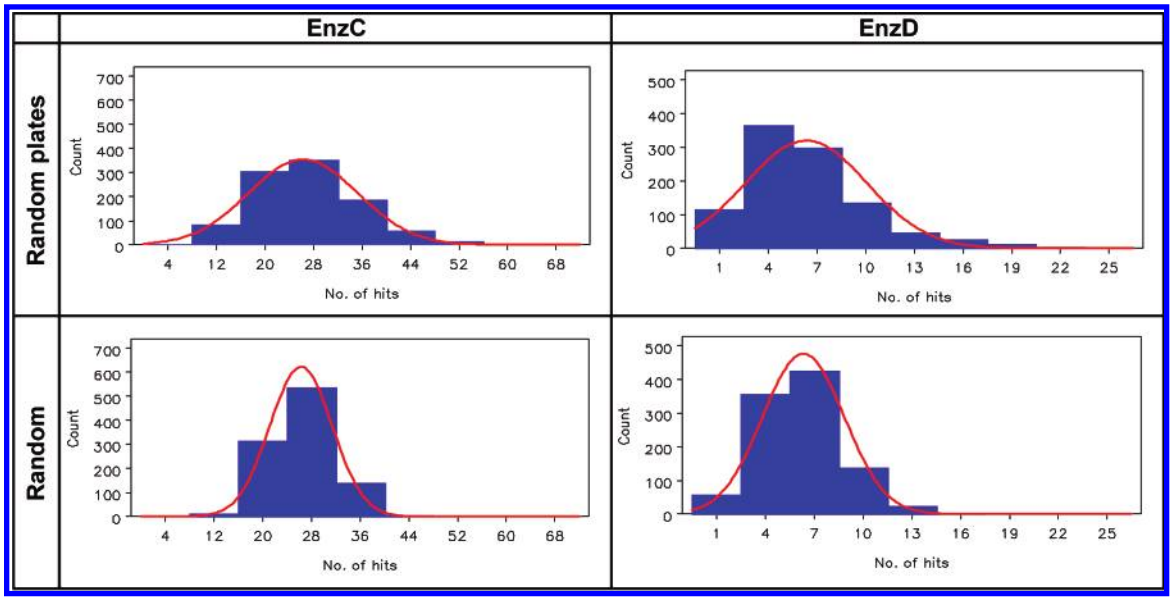

Figure 9. Histograms comparing random compound subsets and random plate subsets.

Table 3. Mean Number of Hits for Each Subset Type

screen	random compound subsets	rational subsets	% increase over random
aGPCRA	123	143	16%
pGPCRC	93	124	33%
EnzA	28	37	32%
EnzC	35	48	37%
EnzD	33	51	55%

toxicity of reagents. In this paper, rationalized cluster-based selection was simulated for 1000 such subsets and evaluated against 1000 subsets comprising random compounds. We show that, using this clustering method for our data, the cluster-based subsets have higher mean hit rates compared to those of the random ones. In addition, subsets comprising random plates are compared with subsets of random compounds. While the mean hit rate of both is the same, the former demonstrates more variation in hit rate.

In this and similar related work, it has been useful to consider the distribution of hits over clusters of different sizes before selecting subsets for any cluster-based rational approach. This can provide an understanding of the relative performance of different subset selection methods. For example, here, the smaller clusters are richer in hits, so selection methods that select proportionally more from the smaller clusters may be expected to outperform other methods. Rational subset selection methods can be thought of as dividing the compound file into subgroups and then selecting n_i from each group. We have also found it helpful to plot hit rates versus n_i. For example, in one case (data not presented), we saw that proportional selection methods picked heavily from one very large cluster, which was largely inactive for the screens under consideration.

It may be tempting to use a subset selection method that puts great emphasis on the part of the compound file shown to have higher hit rates, even to the extent of using compounds from this area alone. Harper et al.[21] refer to such methods as "focused methods", and they contrast them with diverse methods. Focused methods may find many hits, but the risk is of multiple hits from the same chemical series, and thus only a limited number of lead series are identified. The aim of diverse methods is to "cover" chemical space, and this will includes areas where hit rates are considered likely to be low. The rational subset method presented here can be thought of as a diverse method. Advantages of using a diverse method include that hits come from as diverse as possible a representation of the full screening set and that future screening is not so biased by the types of compound structures that have shown activity in past screens. As the type of targets being considered are changing with time, this can be a key advantage (e.g., if all previous screens had been on aminergic GPCRs, for which relatively "leadlike" low molecular weight and clogP compounds give potent activity, a selection based on actives would perform poorly on peptidic GPCR assays, for which the distribution of properties for observed actives is shifted significantly to higher molecular weight and clogP values, to the rule of five border).

The choice of compound file, rational subset method, and ratio of subset size to compound file size are key factors in the relative performance of random and rational testing. We used a "leadlike" subset of the Pfizer file specific for our needs; however, the simulations are as applicable for data sets identified using alternative criteria. We do not suggest that in all situations all types of rational testing will yield the same advantages; however, we advocate the use of some simulation for each case. The value of statistical simulation in subset design is independent of the data set, especially since it is comparatively quick and the amount of simulation could be reduced somewhat by the use of theoretical statistical distributions which for the most part seem to approximate the data well.

The assessment presented in this paper focuses solely on the performance of the subset. Intuitively, if suitable clustering is used and a rational one-compound-per-cluster method used, we would expect individual hits to be at least as diverse as when using random sampling. When subset hits are followed up with virtual screening, hit rates are considerably enriched over experimental equivalents. It is essential that subset screening is followed with iterations of virtual screening to get the most effective use. Clearly, when some information on ligands or the structure of the target protein

is known, a focused/directed virtual screen around the ligands or structure will likely yield the highest number of hits, but the combined use of this set and a diverse set representing the breadth of the full screening file will provide the optimum diversity and yield of actives and enable new "chemotypes" to be found for a target.

APPENDIX 1: RATIONAL SUBSETS

The selection of a compound from a cluster can be modeled as a binomial distribution as follows:

Suppose in cluster i there are H_i hits out of N_i compounds. So, the probability of drawing a hit with one pick follows a binomial distribution with mean H_i/N_i and variance $H_i(1 - H_i)/N_i^2$. This can be approximated by a normal distribution with the same mean and variance (although this approximation is relatively poor; the data fall outside the region where the approximation is considered to be good). The sum of normally distributed variables is normally distributed, so the distribution of the total number of hits in the subset can be approximated by normal distribution with mean $\sum(H_i/N_i)$ and variance $\sum[H_i(1 - H_i)/N_i^2]$ where both sums indicate summing over the clusters.

Although this includes poor approximation en route, the plots suggest that for the data presented overall this is not a bad approximation.

APPENDIX 2: RANDOM COMPOUND SUBSETS

The selection of a random compound subset can be modeled as a hypergeometric distribution. This is an adaptation of the binomial distribution where sampling is without replacement. That is, a compound cannot be selected twice for the same subset.

Suppose a subset of size n is picked at random (without replacement) from a compound file of size N with H hits and the number of hits in the subset is h, then h has mean $H^*(n/N)$. From this, it follows that the mean hit rate for random compound subsets is the same as the full file hit rate, H/N, and

$$\mathrm{Var}(h) = (nH/N)(1 - H/N)(N - n)/(N - 1)$$

when n is small compared to N, $(N - n)/(N - 1) \approx 1$, and the variance becomes

$$\mathrm{Var}(h) \approx n(H/N)(1 - H/N)$$

that is, approximately equal to the binomial variance (sampling with replacement).

The work described in this paper uses the mean and variance from the hypergeometric distribution and approximates the hypergeometric distribution with a normal distribution with the same mean and variance on the plots. Theory suggests that this approximation is good as long as $(n/N) < 0.1$; that is, the subset is less than 10% of the full file.

APPENDIX 3: RANDOM PLATE SUBSETS

For the random plate subsets, as long as each plate has approximately the same number of compounds, the number of hits can be approximated by a normal distribution as follows:

Suppose for the full set of plates the numbers of hits are $\{H_1, H_2, ..., H_M\}$ and the mean, $\bar{H}$, and variance, S_H^2, are calculated. Then, if m plates are drawn, the total number of hits will be approximately normally distributed with mean $m \times \bar{H}$ and variance $m \times S_H^2$. This assumes that the hits on the m plates can be modeled as independent and identically distributed and then applies the Central Limit Theorem.[24] In fact, the assumption is not valid, but again the plots suggest that for the data we have considered overall this is not a bad approximation.

Notice that the mean number of hits per plate $\bar{H} = H/M$, where H is the total number of hits and M the total number of plates. Thus, the mean number of hits of a random plates subset $= m \times \bar{H} = (m/M) \times H$. As in Appendix 1, if n is the number of compounds in the subset and N is the number of compounds in the file and if we assume the plates have equal numbers of compounds, then $m/M = n/N$. Thus, the mean number of hits of a random plate subset is the same as the mean number of hits of a random compound subset.

ACKNOWLEDGMENT

We are grateful to Drs. Alexander Alex and David Price for their roles in the design of the 12 500 rational subset, to Dr. Shyamal Somaroo for his support of our approach and for providing data sets for screens EnzC and EnzD, to Mr. Phil Woodward for providing input to the statistical aspects for the work and the paper, and to Mrs. Fiona Cheyne for help with statistical programming. We also acknowledge the members of the Pfizer Global Subsetting Core Team who were prepared to entertain different approaches and whose wide range of experience provided an excellent sounding board.

Supporting Information Available: Representative Murcko Assemblies[22] for the source file, whether or not present in the 86 000 and 12 500 compound sets. Analysis was conducted after standardizing all non-hydrogen atoms to carbons and all bonds to single.[23] This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) (a) Davis, A. M.; Keeling, D. J.; Steele, J.; Tomkinson, N. P.; Tinker, A. C. Components of Successful Lead Generation. In *Curr. Top. Med. Chem.* **2005**, *4*, 421−439. (b) Karnachi, P.; Brown, F. K. Practical Approaches to Efficient Screening: Information-Rich Screening Protocol. *J. Biomol. Screening* **2004**, *9*, 678−686. (c) Engels, M. F.; Venkatarangan, P. Smart Screening: Approaches to Efficient HTS. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 275−283.

(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(3) Viswanadhan, V. N.; Balan, C.; Hulme, C.; Cheetham, J.; Sun, Y. Knowledge-based Approaches in the Design and Selection of Compound Libraries for Drug Discovery. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 400−406.

(4) (a) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 233−245. (b) Downs, G. M.; Willet, P. Clustering of chemical structure databases for compound selection. In *Advanced Computer-Assisted Techniques in Drug Discovery*; Waterbeemd H., Ed.; Weinhem VCH: Weinheim, Germany, 1995; Vol. 3, pp 111−130.

(5) Young, S. S.; Farmen, M.; Rusinko, A., III. Random Versus Rational Which is Better for General Compound Screening? *Network Sci.* [Online] **1996**, *2*, Article 9. http://netsci.org/Science/Screening/feature09.html (accessed Nov 1, 2006).

(6) Potter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **1998**, *41*, 478−488.

(7) Fontaine, F.; Pastor, M.; Gutierrez-de-Teran, H.; Lozano, J. J.; Sanz, F. Use of alignment-free molecular descriptors in diversity analysis and optimal sampling of molecular libraries. *Mol. Diversity* **2003**, *6*, 135−147.

(8) Hammersley, J. M.; Handscomb, D. C. *Monte Carlo Methods*; John Wiley & Sons: New York, 1964.

(9) Moore, P. G. *Basic Operational Research*; Pitman Publishing: United Kingdom, 1968.

(10) The small ligands may be set aside for a fragment, or Needles, subset, e.g.: Boehm, H. J.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, H.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Muenier-Keller, N.; Mueller, F. Novel Inhibitors of DNA Gyrase: 3D Structure Based Biased Needle Screening, Hit Validation by Biophysical Methods, and 3D Guided Optimisation. *J. Med. Chem.* **2000**, *43*, 2664−2674.

(11) Wenlock, M.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. A Comparison of Physicochemical Property Profiles of Developmental & Marketed Oral Drugs. *J. Med. Chem.* **2003**, *46*, 1250−1256.

(12) Study conducted on a total of 150 000 micromolar actives in the Pfizer compound file and the MACCS Drug Data Report (MDDR; distributed by Elsevier MDL, 2440 Camino Ramon, Suite 300 San Ramon, CA 94583). Similar outcomes were obtained when the analysis was conducted separately for the Pfizer or MDDR compounds. Mason, J. S.; Mills, J. E.; Barker, C.; Loesel, J.; Yeap, S. K.; Snarey, M. Higher-throughput approaches to property and biological profiling, including the use of 3-D pharmacophore fingerprints and applications to virtual screening and target class-focused library design. Abstracts of Papers, 225th ACS National Meeting, New Orleans, LA, March 23−27, 2003; American Chemical Society: Washington, DC, 2003.

(13) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(14) Ligand efficiency is an estimate of binding energy relative to ligand size. At an optimal ligand efficiency of ∼0.4 kcal/atom, a ligand with 17 heavy atoms would exhibit a detectable K_i of 10 μM: Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand Efficiency: A Useful Metric for Lead Selection. *Drug Discovery Today* **2004**, *9*, 430−431.

(15) The 2D fingerprints are developed by Digital Chemistry. http://www.digitalchemistry.co.uk (accessed Nov 1, 2006). The Nearest Neighbour algorithm was implemented by Daylight Chemical Information Systems Inc. http://www.daylight.com (accessed Nov 1, 2006).

(16) The set of 75 000 confirmed actives with affinity under 10 μM was compiled from pre-2001 screens, whilst the test data were six post-2001 HTSs that comprised two enzymes, one aminergic GPCR, and three peptidic GPCRs (prefixed Enz, aGPCR, and pGPCR, respectively).

(17) An additional observation was that, for the four HTS assays, a diverse subset designed without regard for confirmed actives performed no better than a subset picked at random.

(18) The aGPCRA, pGPCRC, and EnzA screens were the same ones used in the previous section (see ref 16).

(19) Hastings N. A. J.; Peacock, J. B. *Statistical Distributions*, 2nd ed.; Butterworths: United Kingdom, 1975; pp 36−40, 78−79, 96−99.

(20) It is unclear why smaller clusters have higher hit rates. It might be because larger clusters usually contain compounds for more common targets, whilst small clusters contain more unusual compounds optimized for a novel target. This view was shared by a reviewer, who observed a similar pattern in his screening file. He also observed that the behavior is more pronounced for assays with lower hit rates, in keeping with higher hit rates in small clusters for such novel targets. Alternatively, a pragmatic explanation is that a low hit is impossible for a small cluster. For example, a single active in a cluster of 15 achieves a hit rate of 7%.

(21) Harper, G.; Pickett, S. D.; Green, D. V. S. Design of a Compound Screening Collection for use in HTS. *Comb. Chem. High Throughout Screening* **2004**, *7*, 63−70.

(22) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(23) Pipeline Pilot, version 6.0.2.0, developed by Scitegic, http://www.scitegic.com (accessed June 7, 2007), was used. An example is the standardized Murcko assembly for sildenafil. The source file gave 37 000 Murcko assemblies, which were further clustered using FCFP4 fingerprints in order to facilitate inspection. Provided are the structures of the resulting 746 cluster centers, the number of Murcko assemblies represented by the cluster, and information regarding whether any of the frameworks in that cluster exists in the 86 000 and 12 500 compound sets.

(24) Lindgren, B. W. *Statistical Theory*, 3rd ed.; Collier Macmillan Publishers: United Kingdom, 1968; pp 157.

CI600382M