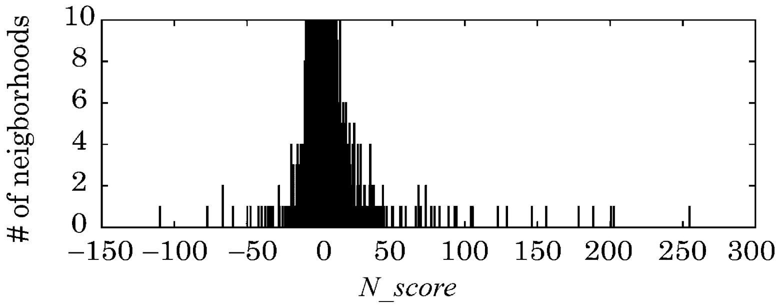
Journal of
Medicinal Chemistry

Brief Article

# Statistical Tools for Virtual Screening

Jennifer R. Krumrine, Andrew T. Maynard, and Charles L. Lerman

*J. Med. Chem.*, **2005**, 48 (23), 7477-7481• DOI: 10.1021/jm0501026 • Publication Date (Web): 25 October 2005



# Statistical Tools for Virtual Screening

Jennifer R. Krumrine, Andrew T. Maynard,[†] and Charles L. Lerman*

*Computational Chemistry & Informatics, Department of Chemistry, AstraZeneca Pharmaceuticals LP, 1800 Concord Pike, Wilmington, Delaware 19850-5437*

*Received February 2, 2005*

In large-scale virtual screening (VS) campaigns, data are often computed for millions of compounds to identify leads, but there remains the task of prioritizing VS “hits” for experimental assays and the dilemma of assessing true/false positives. We present two statistical methods for mining large databases: (1) a general scoring metric based on the VS signal-to-noise level within a compound neighborhood; (2) a neighborhood-based sampling strategy for reducing database size, in lieu of property-based filters.

## Introduction

In structure-based virtual screening campaigns, large compound databases are routinely docked to a protein active site and estimated binding free energies are used to select compounds for experimental testing. The literature reflects that docking and scoring protocols capture low-resolution information about ligand binding but are limited in their ability to accurately predict measured binding affinities.[1,2] The problem of false positives and negatives in virtual screening is especially apparent when docking large databases, where the number of compounds satisfying a “good” docking score, defined by scoring positive-control ligands, can be much larger than assay throughput. For example, in Figure 1 we compare the normalized Glide score[3] distributions of 57 validated, active-site directed, competitive inhibitors of an enzyme target and a database of approximately one million compounds, the vast majority of which are presumed to be inactive.[4,5] Glide discriminates between the two data sets, but there is substantial overlap between the distributions. Imposing a cutoff based on the mean score of the positive controls would leave an unreasonably large number of compounds to assay. Further, assaying only the top-scoring compounds is not a reasonable strategy either because a high percentage of promising compounds would be disregarded. Because binding affinity is only one of many parameters to be considered in the profile of a lead drug, physical property or ADME (absorption, distribution, metabolism, and excretion) predictions are often implemented as pre- or postdocking filters to reduce the number of hits. However, these filters are also limited in accuracy and can preempt the identification of novel leads.

Rather than focus on the rank-ordering of individual compound scores, we exploit the information content of virtual screening (VS) data within neighborhoods of structurally related compounds to attenuate false assignment of positives and negatives. Examination of the docking scores and predicted binding modes of related compounds suggests that neighborhoods with consistently good docking scores exhibit low-resolution, virtual structure−activity relationships (SARs), upon which a neighborhood analysis may be based.

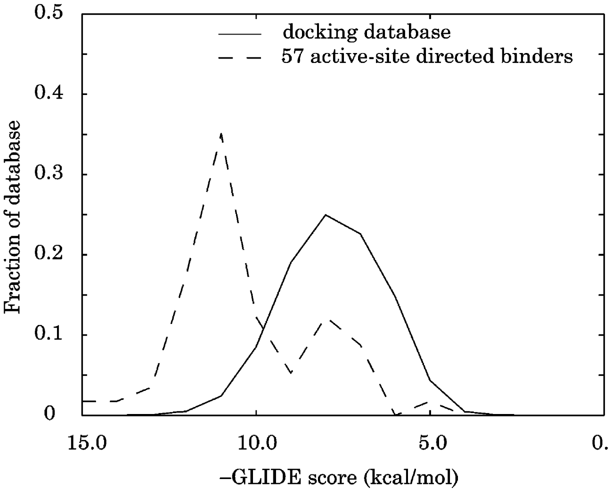


**Figure 1.** Glide docking scores of 57 known, active-site directed ligands are compared to the scores of a database of approximately one million compounds, using normalized histograms. All 57 compounds in the validation set had favorable estimated free energies of binding (Glide-score < 0). The average docking score for the validation set is −2.5 kcal/mol lower than the average of the general database, indicating that the Glide scoring function has some ability to discriminate between compounds that do and do not bind to the target. However, tens of thousands of compounds in the collection meet the criterion of a “good” docking score (defined in this case as −10 kcal/mol), which is far more than can be tested; this is a common problem when docking very large databases that mandates the use of additional filtering or selection criteria.

In section I, we present a method that defines VS hits in terms of the significance of neighborhood virtual SAR. In section II, we present and validate a sampling strategy for efficiently mining large compound databases that can be implemented as an alternative to imposing hard-cut ADME or property-based filters.

* To whom correspondence should be addressed. Phone: 302-886-2577. Fax: 302-886-5382. E-mail: charles.lerman@astrazeneca.com.

[†] Present address: Icagen, 4222 Emperor Blvd, Suite 390, Durham, NC 27703. E-mail: amaynard@icagen.com.

10.1021/jm0501026 CCC: $30.25 © 2005 American Chemical Society
Published on Web 10/25/2005

## I. Neighborhood Scoring: N_score

We originally applied a neighborhood-based analysis to high-throughput screening (HTS) data to assess the likelihood of false positives and negatives. In contrast to evaluating compound activity on an individual basis, activity is assessed in the context of observed activity (signal) within a neighborhood of structurally related compounds. For example, an HTS hit that is in a neighborhood containing mostly inactive compounds is more likely to be a false positive than a hit found in a neighborhood of active compounds (assuming random error). Conversely, a compound that is not a HTS hit, but which is closely related to a large number of active compounds, is more likely to be a false negative than a compound in a neighborhood of inactive compounds. This concept was previously described in a strategy for prioritizing compounds for screening in batches, where known activity or inactivity of a hit's analogues was incorporated as positive or negative feedback during selection cycles.[6]

We assess the likelihood that a compound is a true (false) positive by computing the enrichment (depletion) of activity within the compound neighborhood. In virtual screening, "active" compounds correspond to those with "good" docking scores, where the cutoff is based on the scores of the validation set ligands with known activity (see Figure 1). The binomial distribution $B(X,N,P)$ gives the probability of observing $X$ actives in a sample of $N$ compounds, where the probability of observing $X$ actives at random is $P$, the overall hit rate of the virtual screen. The sample is derived from a structurally related neighborhood. A compound neighborhood that is significantly enriched in activity (i.e., a neighborhood of compounds that are most likely true positives) satisfies the following inequality:

$$\int_X^N B(x,N,P)\,\mathrm{d}x \le p \qquad (1)$$

Similarly, a neighborhood is significantly depleted of hits when

$$\int_0^X B(x,N,P)\,\mathrm{d}x \le p \qquad (2)$$

For convenience, we scale the computed $p$ values to define the following statistic for scoring compounds in the context of neighborhood SARs:[7]

$$\text{N\_score} = \operatorname{sign}(X - N{\cdot}P)\left[-\log \int_{x_1}^{x_2} B(x,N,P)\,\mathrm{d}x\right] \qquad (3)$$

We typically apply a 95% confidence level criterion to classify compounds into three categories: (1) Compounds with N_score > 1.3 ($p$ < 0.05) have neighborhoods that are enriched with actives and are attractive starting points for establishing SAR; hits from these neighborhoods are more likely to be true positives. (2) Compounds with N_score < −1.3 have neighborhoods that are depleted of actives; hits from these neighborhoods are more likely to be false positives. (3) Compounds with 1.3 > N_score > −1.3 are indeterminate; such compounds originate from sparsely populated neighborhoods that lack adequate structural coverage to significantly establish SAR or well-populated neighborhoods that exhibit a level of activity that is similar to the overall VS hit rate $P$.

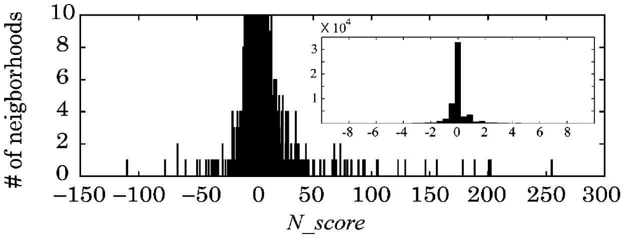


**Figure 2.** Histogram of N_score values for all neighborhoods in the docking database (~1 million compounds). N_score at the extremes (>1.3 and <−1.3) indicate true and false positive compound neighborhoods, respectively. The neighborhoods that are most likely to contain true positives and SAR for establishing leads have the largest positive N_score (e.g., N_score ≫ 1.3). In contrast, the vast majority of neighborhoods, as shown in the inset, fall in the central region between −1.3 and +1.3, corresponding to the less interesting sparse neighborhoods (singletons) or neighborhoods with an SAR signal that is similar to noise.

The first step in computing N_score is to define compound neighborhoods. In this study, it was convenient to cluster the entire docking database, where the resulting clusters define neighborhoods. Daylight fingerprints[8] were used to cluster the database, applying a cutoff of 70% similarity (Tanimoto) to define clusters[9] using a sphere exclusion method.[6,10] As implemented, each compound is assigned uniquely to a cluster and N_score is computed for each cluster.[11] Martin et al. recently investigated whether calculations of molecular similarity translate into factors that lead to biological similarity for the purposes of designing combinatorial libraries and selecting diverse compounds to augment compound collections.[12] Their observations show that two similar compounds, as defined by Daylight fingerprints, are more likely to share biological properties than are two compounds selected at random. The correlation, sometimes low, varies with the threshold used to define similarity. Other investigations, summarized by Martin et al., report varying degrees of correlation between computationally defined structure classes and biological activity, with Unity and Daylight fingerprints tending to outperform other types of descriptors. Given its generality, any meaningful similarity metric or clustering method can be used to compute N_score. We did not investigate how alternative clustering techniques or similarity metrics might amplify different regions of SAR because this was beyond the scope of this communication.

In our docking campaign, the evaluation of N_score enabled the identification of several thousand compound clusters as hits (N_score > 1.3), as shown in Figure 2. Four or fewer representative compounds, depending on availability, were then selected from approximately 300 neighborhoods defined with N_score ≥ 10, comprising approximately 700 compounds, which was compatible with our assay throughput. There was no correlation between cluster size and N_score (plot not shown), although larger clusters have a larger possible range of N_score.

An unexpected result of our docking campaign was that among the experimentally confirmed hits was a compound series that would have been eliminated before docking if common prefilters had been applied. Yet this series was a viable starting point. With minor synthetic modifications that did not diminish the binding affinity, the series passed the prefiltering criterion that it originally failed. While our original motivation for not prefiltering the database was a lack of viable leads from other methods, like HTS, this result suggests to us that a "leave no stone unturned" approach may be desirable for other targets as well.

## II. Managing the Size of Docking Databases Using Sampling by Variables

It has been proposed in the literature that compound databases should be first filtered on the basis of ADME properties, then screened to predict binding to the protein or receptor of interest.[13] Prefiltering is often implemented to limit the size of large databases or in an effort to focus on compounds that are the most leadlike or druglike. The argument is that known druglike compounds are limited in diversity relative to the entire set of known compounds and fall within distributions of relatively simple descriptors (e.g., MW, cLogP, H-bond counts, etc.). Therefore, the accumulated knowledge about properties of orally active drugs can be applied to drug discovery, regardless of the target. However, with recognition that examples of orally bioavailable drugs can be found that violate any number of physical and chemical property filters and confound ADME prediction tools, database prefiltering may preempt the discovery of families of compounds that contain recognition features that are not otherwise represented in a database, as was our experience described in section I.

To retain the chemical diversity of large databases and at the same time enable large-scale virtual screening, we developed a neighborhood sampling strategy. We emphasize that this is a general method for determining how many structurally related compounds need to be screened or tested in order to estimate the overall behavior of the neighborhood within user-defined limits of acceptable error. One can imagine, for example, designing a screening library from a corporate database using this technique to determine how many compounds to include within each structural class. We retrospectively validate this approach in "Sampling Strategy Analysis" below, using the large data set generated in the docking campaign described above. In validating this method, we ask the following question. If we had screened only a select fraction of the database, could we have identified the same neighborhoods as in section I, where the entire database was screened? To answer this question, we analyze the recovery of neighborhoods with high N_score values, found by exhaustive docking above.

Sampling plans are widely applied in commerce to ensure quality control. Within mutually acceptable risk limits, product lots are accepted or rejected on the basis of the quality of the sample that is examined.[14] Sampling plans are constructed by first defining the reasonable risks of accepting an "undesirable" lot and of rejecting a "desirable" lot.

In our case, a "lot" is a neighborhood of related compounds. We determine the sample size $n$ and the mean dock score of the sample compounds, $k$, that will provide statistically meaningful information upon which to accept or reject the neighborhood of compounds. We define the acceptable risk of accepting false positive neighborhoods to be $\alpha = 0.95$ (if the true mean dock score of the neighborhood is acceptable, we require the probability of identifying the neighborhood as a hit to be 95%) and adopt the same criterion for rejecting false negatives: $\beta = 0.95$.

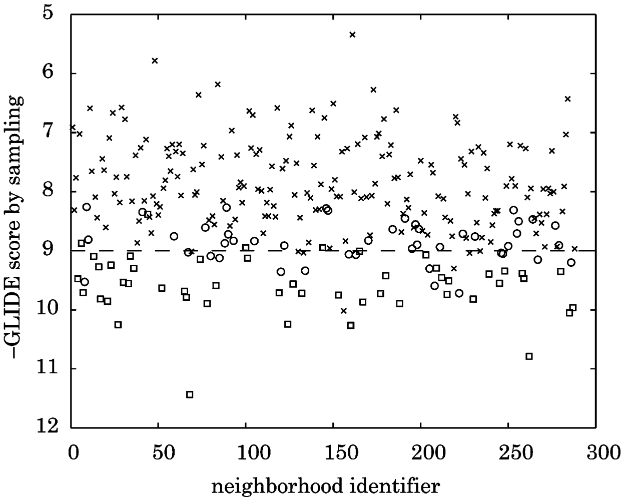


**Figure 3.** Large neighborhoods that are hits by neighborhood analysis (○ = N_score > 1.3 and □ = N_score ≥ 10) and nonhits (× = N_score ≤ 1.3) are plotted by their sampled ($n$ = 30) average dock scores. Only neighborhoods containing 200 or more compounds are shown (288 neighborhoods). The sampled average score for neighborhoods containing 199 or fewer compounds (not shown) are better represented by sampling than large neighborhoods because the sample size $n$ = 30 comprises a larger fraction of the neighborhood.

The criterion for accepting a neighborhood is the probability ($t$-distribution)[15] that the mean dock score of the neighborhood, $\mu$, which is estimated by sampling, is as favorable as a "good" dock score, defined here as −10 kcal/mol:

$$\frac{k - (-10)}{\sigma/\sqrt{n}} = -1.64 \tag{4}$$

Similarly, we define the risk for accepting a false positive neighborhood, where −8 kcal/mol is defined to be the cutoff for an unacceptable dock score:

$$\frac{k - (-8)}{\sigma/\sqrt{n}} = 1.64 \tag{5}$$

By use of eqs 4 and 5, the number of compounds that should be randomly sampled from each neighborhood is a function of the standard deviation: $n = 1.20\sigma^2$.

In our investigation, the mean and maximal $\sigma^2$ over all neighborhoods are 0.9 and 5.8, disregarding all compounds that do not fit in the active site.[16] In practice, one would not know the range of values of $\sigma^2$, which will depend on the docking software that is used. One can estimate a value for $\sigma^2$ to estimate $n$, for example, by docking a representative neighborhood. This procedure is approximate but not misleading, provided $\sigma^2$ is estimated generously. In our investigation, we implement a sample size $n = 30$, which corresponds $\sigma^2 = 25$.

**Sampling Strategy Analysis.** In Figure 3, we show that the neighborhood behavior that is observed when the entire database is screened is well approximated by sampling. "Layers" are observed, where the most favorable neighborhoods, those with N_score > 10 (which are denoted by □), also generally have the lowest mean dock scores by sampling and where neighborhoods with N_score between 1.3 and 10 (denoted by ○) generally score better than neighborhoods with N_score below 1.3 (denoted by ×).

In practice, one would determine how many compounds can be tested with the available assay resources and then pick representative compounds, starting with neighborhoods having the lowest sampled mean dock scores. For example, if resource were available to assay several hundred compounds, this would effectively correspond to a cutoff of $k = -9.0$. Representative compounds would be selected from the 68 neighborhoods below the dotted line in Figure 3, 61 of which have an N_score > 1.3 (circles below the cutoff line). This includes 46 of the 50 neighborhoods with N_score ≥ 10 (squares below the cutoff line).

The computational savings from implementing the sampling strategy can be estimated after the initial clustering. In our investigation, sampling only 30 from each neighborhood or the entire neighborhood if it contains fewer than 30 compounds reduces the size of the docking database by 43%. This is in addition to the approximately 10% database reduction associated with the exclusion of singletons and neighborhoods that contain very few compounds, which is appropriate for both statistical methods presented in this paper.

The required initial database clustering does not diminish the efficiency of the sampling strategy. Glide requires on average 48 s of CPU time to dock a compound with 0−10 rotatable bonds on a 1.3 GHz Linux Pentium III. On a database of one million, the 53% reduction therefore reduces the cost of docking by approximately 294 CPU days times the average number of isomers per compound. In comparison, clustering the database of approximately one million required less than a week on a single processor with our in-house software. Regardless of the number of processors to which one has access, at some point there comes a tradeoff between the number of compounds that can be screened and the number of seconds that can be spent per compound. This sampling strategy is a technique that enables one to apply more accurate (but more time-consuming) virtual screening techniques on a larger database than would otherwise be possible.

## Conclusions

The statistical methods presented here facilitate the analysis and management of massive amounts of data that are generated in large-scale virtual screens. Neighborhood-based analysis of VS data capitalizes on low-resolution, virtual SAR. In this study, N_score captures the signal-to-noise level of SAR associated with a specific docking procedure and molecular force field scoring function (Glide), though N_score can be used with any computed or experimental data. While N_score mitigates random error associated with virtual screening protocols, it cannot compensate for systematic error, underscoring the need for further development of accurate molecular scoring functions and docking poses. Databases with fewer than several hundred thousand compounds can be docked even with modest resources, and the results are prioritized using the N_score analysis described above. For databases exceeding several hundred thousand compounds, where one has only limited computational resources, or for databases containing several million compounds, we recommend a neighborhood-based sampling strategy over property-based filtering methods for managing database size. Massive virtual screening, rather than massive prefiltering, may be essential for identifying viable leads.

**Acknowledgment.** We thank Peter Kenny and James Damewood for contributing ideas and helpful discussion. We thank David Cosgrove for providing our in-house clustering software. We thank Schrodinger, Inc., for the generous trial period license of the Glide docking software, which enabled our large-scale virtual screen.

## References

(1) Bissantz, C.; Folkers G.; Rognan D. Protein-based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(2) Lyne, P. Structure-Based Virtual Screening: An Overview. *Drug Discovery Today* **2002**, *7*, 1047−1055.

(3) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(4) The docking database was prepared using Leatherface[5] for the appropriate generation of ionization, tautomeric, and configurational states. All isomers were docked independently, and the best score was retained as the Glide score for the compound.

(5) Kenny, P. W.; Sadowski, J. *Structure Modification in Chemical Databases. Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2005; Vol. 23, pp 271−284.

(6) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.

(7) For $X < N{\cdot}P$, the $p$ value interval is $[x_1,x_2] = [0,X]$, and for $X \geq N{\cdot}P$, the interval is $[X,N]$, where $X$ is the number of compounds in the neighborhood scoring above a designated threshold, $N$ is the neighborhood size, and $P$ is the hit rate of "actives" in the virtual screen. The sign of the score distinguishes true (+) and false positive (−) domains. N_score > 1.3 indicates a compound neighborhood with a statistically significant number of hits (95% confidence). In MATLAB, N_score = −log 10(min(binocdf($X,N,P$),1−binocdf($X$−1,$N,P$))) sign($X - N{\cdot}P$). Equivalently, N_score can be evaluated using binopdf in a loop, which avoids a potential roundoff error in the binocdf function.

(8) Daylight Chemical Information Systems, Inc.: Irvine, CA; http://www.daylight.com.

(9) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(10) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(11) As an alternative (unpublished) to clustering a database, the neighborhood of each compound in a database is sampled locally and independently. For example, the local neighborhood of a given compound is defined as the set of compounds within a specified Tanimoto radius (e.g., 0.3 units) of the compound. In this manner, neighborhoods are locally sampled for each compound and N_score is computed according to eq 3. Clusters are neither computed nor stored, avoiding potential difficulties in assigning the cluster membership of compounds.

(12) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(13) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.

(14) Mandel, J. *The Statistical Analysis of Experimental Data*; Dover Publications: New York, 1964.

(15) One could decide that acceptable risk limits are neighborhood-specific; for example, if a particular neighborhood contains a large number of compounds that are readily available for testing compared to other neighborhoods that would require extensive synthesis, it may be worthwhile to adopt a sampling plan for neighborhood-specific sampling that is more computationally expensive but provides more information.

(16) We believe that by eliminating those few compounds in any given neighborhood that clash with the receptor or find no reasonable pose, possibly due to inadequate sampling, one derives the clearest picture of the neighborhood behavior; otherwise, one compound with a score of 10 000 would disallow selection of that particular neighborhood, even if all other compounds have excellent dock scores. Further, one might make an exception to prefilter compounds with a very large number of rotatable bonds from the docking database, which can be extremely time-consuming to dock, especially if the neighborhood is reasonably well represented by similar compounds with fewer rotatable bonds.

JM0501026