The Effect of Structural Redundancy in Validation Sets on Virtual Screening Performance

Robert D. Clark,[*,a,b] *Jennifer K. Shepphird,*[c] *and John Holliday*[d]

Corresponding author's e-mail: bclark@bcmetrics.com

[*] Correspondence to: R.D Clark, 827 Renee Lane, St. Louis, MO 63141 USA. E-mail: bclark@bcmetrics.com.

[a] Biochemical Infometrics, St. Louis, MO 63141 USA.

[b] Indiana University School of Informatics, Bloomington, IN 47408 USA

[c] Tripos International, St. Louis, MO 63144 USA.

[d] Department of Information Studies, The University of Sheffield, Sheffield S1 4DP UK.

SUMMARY

The performance of classification models is often assessed in terms of how well it separates a set of known observations into appropriate classes. If the validation sets used for such analyses are redundant due to bias in sampling, the relevance of the conclusions drawn to prospective work in which new kinds of positives are sought may be compromised. In the case of the various virtual screening techniques used in modern drug discovery, such bias generally appears as over-representation of particular structural subclasses in the test set. We show how clustering by substructural similarity, followed by applying arithmetic and harmonic weighting schemes to receiver operating characteristic (ROC) curves, can be used to identify validation sets that are biased due to such redundancies. This can be accomplished qualitatively by direct examination or quantitatively by comparing the areas under the respective linear or semilog curves (AUCs or pAUCs).

Keywords: docking; circular fingerprints; clustering; ROC; validation

1. INTRODUCTION

Large-scale combinatorial synthesis and high-throughput screening (HTS) once held out the promise of making prediction of biological activity on the basis of chemical structure a matter of purely historical interest, but that possibility has not been realized. Instead, chemometric tools (virtual screens, or vHTSs) have become critically important for determining which of the many possible focused libraries should be pursued; for identifying important lead series that may have been missed by HTS; and for identifying structural series large enough to establish quantitative structure-activity relationships (QSARs). Several different kinds of virtual screen are now in use, each serving a somewhat different purpose in lead discovery and optimization. The primary use of substructural and topological similarity searching is to identify analogs for follow-on synthesis or purchase. Docking and similarity methods based on pharmacophores, shape and other high-level properties, in contrast, are used to identify potential lead- and scaffold hops – alternative chemistries that can serve as a hedge against potential development issues involving ADME and pharmacokinetic problems with lead classes already in hand.

It is tempting to search for a "philosopher's stone" for vHTS methods that will be suitable for all targets, but the reality is that different methods are suitable for different targets and will continue to be so for the foreseeable future. Even within a class of methods – e.g., docking – all tools are not equally well-suited to all targets,[1,2,3,4] so researchers need to be able to compare the performance of different methods on different ligand classes and against different targets. As for any chemometric method, some kind of calibration is necessary to evaluate performance. The area under the curve for receiver operating characteristic plots (ROC AUC) has proven itself particularly useful as a summary statistic for comparing how well various vHTS methods work.[5,6,7,8,9]

An ROC curve is a plot of the fraction of true positives recovered at a given

stringency against the fraction of negative examples that score as well or better than the worst of those true positives – i.e., of the sensitivity α against the specificity β. The curve is drawn using data obtained for a validation set of reference observations made up of known positives – in this case, compounds known to bind to the target protein - and known negative examples – in this case, "decoy" compounds that are known (or, more often, presumed) not to bind.

Unfortunately, the reference sets used for carrying out such evaluations are generally drawn from compound collections accumulated over many years' time. Hence their composition reflects many historical influences, including incidental development series as well as the particular offensive and defensive patent strategies employed over the years; those strategies, in turn, reflect fashions and trends in pharmaceutical development. Such any such data set tends to be structurally "clumpy"[10,11,12] and the bias in sampling represented by that clumpiness compromises many of the otherwise excellent statistical properties of an ROC analysis carried out on truly independent observations.[8,12,13] The lack of robustness that results from any such bias can seriously skew retrospective analyses and mislead researchers as to which method is likely to give the best prospective performance.

This distortion can be addressed by clustering the positive examples into more or less "natural" groups, then including one representative from each group (or subclass) in a revised validation set. [14,15] Unfortunately, doing so often reduces the number of positives to the point that it is hard to conclude anything about the validation analysis with confidence. Moreover, it is not obvious *a priori* which positive will be most representative in a given situation.

An alternative approach is to weight the positives in different classes differently. Two schemes for doing so were recently proposed in a purely theoretical paper by Clark and Webster-Clark[16]: *arithmetic weighting*, which gives every ligand in a class equal weight and gives every class the same overall influence; and

harmonic weighting, which gives earlier hits within each class greater weight and favors larger classes somewhat. That publication showed how the qualitative and quantitative behavior of artificially constructed data sets depended on whether they were biased or unbiased. Here we have analyzed results from two different kinds of vHTS system (docking and functional circular fingerprint similarity) and find that bias can be a problem in real-world validation sets.

The weighting technique described is likely to be applicable to any instance where the there is or may be substantial sampling bias in the calibration set against which a model's performance is evaluated.

2. METHODOLOGY

2.1. Data sets and vHTS.

Ligands, decoys and targets for the docking analyses were taken directly from the Database of Useful Decoys (DUD).[4] Structures for 24 targets – acetylcholinesterase (*ache*); adenosine deaminase (*ada*); ampC β-lactamase (*ampc*); catechol *O*-methyltransferase (*comt*); cyclooxygenase isoforms 1 and 2 (*cox1* and *cox2*); dihydrofolate reductase (*dhfr*); epidermal growth factor receptor (*egfr*); fibroblast growth factor receptor kinase (*fgfr1*); factor Xa (*fXa*); glycinamide ribonucleotide reductase (*gart*); reverse transciptase from human immunodeficiency virus (*hivrt*); hydroxymethylglutarylCoA reductase (*hmga*); human heat shock protein 90 (*hsp90*); enoyl acylcarrier protein reductase (*inha*); mineralocorticoid receptor (*mr*); poly(ADP-ribose) polymerase (*parp*); progesterone receptor (*pr*); *S*-adenosylhomocysteine hydrolase (*sahh*); sarc tyrosine kinase (*sarc*); thrombin; thymidine kinase (*tk*); trypsin; and vascular endothelial growth factor receptor kinase (*vegfr2*) – were downloaded from the UCSF web site[17] and run through Surflex-Dock.[18,19] The DUD data set includes about 37 negative examples (decoys) chosen to roughly match the overall physical properties of each positive example (known active), thereby reducing the risk of artificial enrichment.[20,21] Hence the total number of compounds

docked against each was approximately $N = 37n$, where n is the number of positives connected with each target.[4] When multiple tautomers or protonation states were present in the database, the best-scoring form was used for the ROC analyses described here. Default values were used for all run parameters and no attempt was made to optimize the results. In particular, no effort was made to improve performance by using multiple starting configurations, applying pre- or post-optimization of ligand geometries, or allowing rings to flex.

For circular fingerprint screening, exemplars of six pharmacological classes (Figure 1) taken from the MDL Drug Data Report (MDDR)[22] were used as queries to rank the remainder of the database (ca. 102,000 compounds) in order of Tanimoto similarity to each query.[23,24] Compounds annotated as belonging to the same pharmacological class as the query were considered "positive" for subsequent ROC analyses; the number ranged from 636 to 1130 following standardization with Concord.[25] All compounds in the MDDR (ca. 102,000) not annotated as belonging to the same pharmacological class as the query were used as decoys. vHTS was carried out in SciTegic Pipeline Pilot[26] using the functional circular fingerprints (FCFP_4) derived from their smiles strings. These fingerprints encode the topological relationship between the pharmacophoric features within a molecular structure.[27]

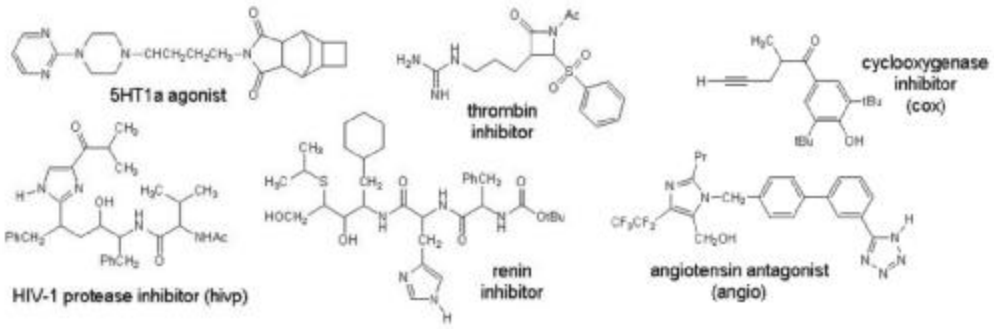

Figure 1. Target structures used for FCFP_4 screens run against the MDDR.

Cluster analysis was carried out by converting structures to SLN format[28] and clustered in SYBYL[29] based on the Tanimoto similarity of their UNITY[30] fingerprints. In contrast to FCFPs, these fingerprints encode the various substructures found within a molecule. Complete-linkage hierarchical agglomerative clustering was used. In this method, the similarity between clusters was taken as the minimum similarity between the fingerprints for any pair of structures in which one is drawn from each of the clusters being compared, which tends to keep clusters "tight" by keeping the similarity between all individuals within a cluster relatively high.[31] Appropriate clustering levels were identified by examination of sorted distances between clustering levels – i.e., the degree of dissimilarity between the pair of clusters consolidated in going from *k* to *k - 1* clusters – with the proviso that each level included increase the number of clusters by two- to four-fold, if possible. The number of positives in each class are listed in Table 1, along with the clustering levels used for each target.

Table 1. Number of individual actives for each class and number of clusters considered.

Target	**Cluster Levels**	**Target**	**Cluster Levels**	**Target**	**Cluster Levels**
trypsin	44[a]	dhfr	201, 23, 9	tk	22
fXa	134, 16, 6	inha	85, 8	hivrt	40
thrombin	65, 12	comt	11	ache	105, 11, 5
cox2	348, 71, 17, 6	src	155, 16, 4		
sahh	17	ada	17		
gart	21	gpb	27	*renin*[b]	641, 48, 20, 6
ampc	21	cox1	25, 6, 4	*angio*	913, 106, 16, 10

mr	15	fgfr1	118, 5	*hivp*	584, 60, 21, 6
parp	33	egfr	444, 27, 11	*thrombin*	752, 96, 23, 7
hsp90	24	vrgfr2	74	*5HT1a*	824, 89, 20, 7
hmga	35	pr	27	*cox*	636, 42, 27, 7

[a] The first clustering level given is for singletons in each case and indicates the number of active ligands provided for the corresponding target.

[b] Ligands for targets highlighted in italics were from the MDDR; others are from the DUD data set.

2.2. Weighting schemes for α.

The uneven distribution of positives across structural class can be addressed by recognizing that not all positives are created equal. This was done here by examining two alternatives to the uniform weighting usually used for each true positive's contribution to α:

uniform: $w_{ij} \propto 1$
arithmetic: $w_{ij} \propto 1/n_j$
harmonic: $w_{ij} \propto 1/i$

where w_{ij} is the weight for the i^{th}-ranked member of the j^{th} cluster and n_j is the number of positives in the *jth* cluster. The plotted α_{ij} in each case is the sum of weights for all positives scoring as well or better than the corresponding observation divided by the sum of weights across all *n* positives in the validation set. For *uniform weighting*, this normalization factor is simply *n*.

Arithmetic weighting gives each class of positive the same overall influence on the AUC statistic, and is equivalent to averaging the results for all possible combinations of subsets in which one example is taken from each cluster.

Harmonic weighting recognizes two practical realities: that larger clusters are more valuable than smaller ones, if only because they can reveal meaningful structure-activity relationships (SARs); and that earlier hits within each class are more informative than are later ones. Overall influence of a class on the AUC increases roughly as the natural logarithm of the cluster size under a harmonic weighting scheme, which is consistent with the fact that small lead series are usually of less practical value than are more fleshed-out series.

For clarity, a logarithmic scale is used for the false positive rate β in the ROC plots, which is problematic for nominal frequencies of zero. This situation arises because the number of negative examples used to *estimate* the values of β is finite. The resulting granularity limits the accuracy with which the actual false positive rate for the pool of all possible structures is being estimated. It is common in such situations to make a continuity correction.[32,33] A value of 0.5/(*N*-*n*), where *N* is the total number of structures in the data set and *n* is the number of positives, is a good estimate β_0 to use when no false positives are observed in a particular sample (i.e., validation set). Here, we set β_0 to the more conservative 1/(*N*-*n*), which provides a direct visual reminder of the size of the decoy set as well as avoiding the need to take the logarithm of zero.

2.3. Summary statistics.

Full ROC plots are useful for making qualitative comparisons but are cumbersome when more than a few curves are involved. Quantitative comparisons are generally made by comparing the areas under the respective curves instead. For the finite data sets used for validation, this area is estimated by summation across all positives:

$$AUC = \frac{1}{\gamma}\sum_{j=1}^{K}\sum_{i=1}^{n_j} w_{ij}(1-\beta_{ij}) = 1 - \frac{1}{\gamma}\sum_{j=1}^{K}\sum_{i=1}^{n_j} w_{ij}\beta_{ij} \qquad (1)$$

where β_{ij} is the estimated false positive rate for i^{th}-best score from the j^{th} class and the normalization factor $\gamma = \Sigma\Sigma w_{ij}$.

As a practical matter, vHTS is generally used as a pre-filter for focused library synthesis programs or biochemical screening, which means that only differences in scores among the top 1-10% for the data set are meaningful. Several investigators have expressed concern about the appropriateness of the classical ROC AUC because of this and have recommended exponential weighting schemes that favor "early hits." [34,35] Harmonic weighting favors early hits naturally, with the best-scoring hit in each class (i.e., cluster or series) receiving twice the weight of the second-best hit, three times the weight of the third-best and so on. This effect is only seen within classes, however.

An alternative approach is to reduce the contribution of "late" hits by applying a logarithmic transform to the *x*-axis of the ROC curve, an action which is effective across all classes. Doing so is broadly consistent with the exponential weighting schemes proposed by others but avoids the need to specify a weighting factor *a priori*. Like the AUC, the logarithmic integral (pAUC) is estimated from a summation across all positives:

$$pAUC = \frac{1}{\gamma}\sum_{j=1}^{K}\sum_{i=1}^{n_j} w_{ij}[-\log_{10}(\beta_i)] = \frac{1}{\gamma}\sum_{j=1}^{K}\sum_{i=1}^{n_j} w_{ij}\log_{10}(\frac{1}{\beta_{ij}}) \qquad (2)$$

The AUC is equal to the average true negative rate (1-β) and the pAUC is equal to the average stringency ($\log_{10}(1/\beta)$).

3. RESULTS AND DISCUSSION

Performance statistics for docking the 24 sets of DUD positives into their respective targets are shown in Figure 2, as are the results for the six FCFP similarity screens; both linear (AUC; black bars) and logarithmic (pAUC; red bars) summations are shown. The docking results are broadly consistent with those reported for other docking programs, with a median AUC of 0.632. Only eight of the docking trials (32%) yielded an AUC greater than or equal to 0.70, and performance on five targets fell below the value of 0.500 expected for guessing at random. Performance on acetyl cholinesterase (*ache*), in fact, was

substantially worse than random. The AUC statistics for the ligand-based screen are nominally much better, with a median value of 0.865; the weakest result was for cyclooxygenase, which had an AUC of 0.59.

The logarithmic integrals (pAUCs) are plotted in Figure 2 as red bars alongside the linear integrals shown in black. In principle, an ROC curve can be thought of as representing a sample drawn from a pair of Gaussian distributions. Were that simple scenario always applicable in practice, applying the log transform – which is monotonic – should not affect the rank-ordering of a series of ROCs. The results presented in Fig. 3 show that this is indeed generally the case, but not always: mineralocorticoid receptor (*mr*), dihydrofolate reductase (*dhfr*), tyrosine kinase SRC (*src*), fibroblast growth factor receptor kinase (*fgfr1*) and epidermal growth factor receptor (*egfr*) stand out as exceptions. Nonetheless, only four docking targets (16%) yield pAUCs that rise above the 1.0 threshold where the average positive is recovered among the top 10% of the data set. Five of the six ligand similarity screens (83%) rise above this level, and two (33%) yield pAUC values above 2.0, indicating that the average positive falls in the top 1% of the data set.

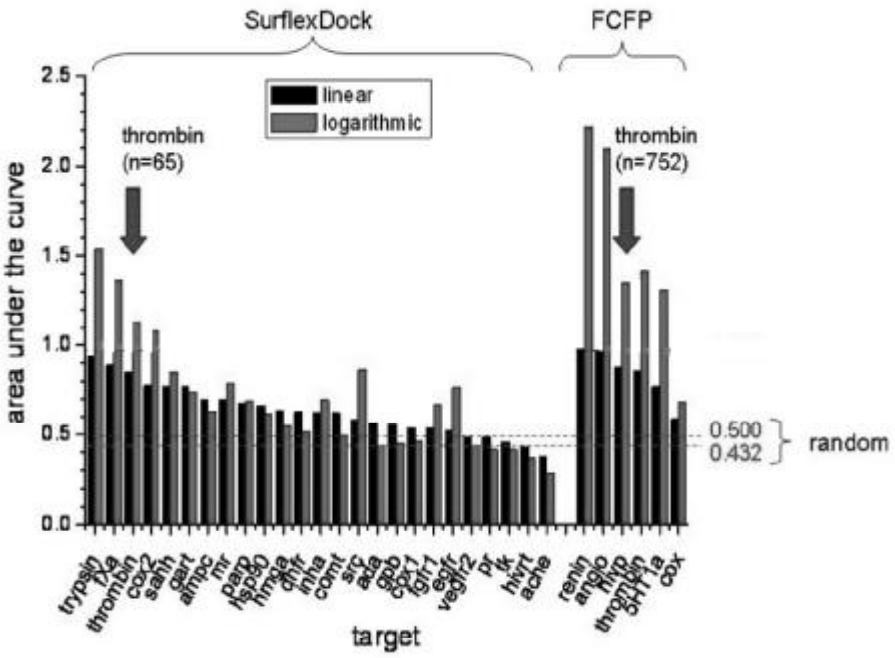

Figure 2. Summary statistics for VHTS using uniform weighting. Areas under the curve were calculated across a range of target types for both docking and FCFP similarity screens. AUCs where ordinates were scaled linearly are shown as black bars, whereas logarithmically scaled integrals (pAUCs) are indicated by red bars. The corresponding values for random guessing are 0.5 and 0.432 ($=1/\log_e(10)$), respectively. The target abbreviations used are taken from Huang et al.[4] for the DUD data sets and those provided in Figure 1 otherwise. Targets are sorted in order of decreasing AUC for each screen type.

The differences in the validation sets used in the two cases make fair comparisons based on summary statistics problematic at best.[4] Comparisons of vHTS performance were not the goal of this work, however. In fact, suboptimal performance is desirable for assessing bias, since it tends to exaggerate the differential effects of weighting on the ROC curves: it would be hard to see any change in the plots if all of the positives outscore all of the decoys. As others have noted, using default settings is unlikely to yield optimal docking performance in general.[15,36] In the particular case of Surflex-Dock, allowing

multiple starting configurations, geometry optimization and ring flexibility – none of which was done here – improves docking scores, particularly for true positives,[37] albeit at the cost of reduced processing speed. Furthermore, only the matched sets of DUD decoys were used here, not the less demanding consolidated set. Some of the matched decoys that score well represent inappropriate protonation states or tautomers, and others are close structural analogs of true positives that one would want to test for activity in any event. Removing such "false false positives" improves the performance statistics substantially, but has no impact on the relative change caused by using a different weighting scheme, which is the focus of this paper.

The plots obtained by applying the various alternative weighting schemes to two artificial data sets[16] are shown in Figure 3. To create the biased data set, higher scores were systematically assigned to positives from larger *ad hoc* clusters. Scores for the unbiased data set, on the other hand, were evenly distributed. The corresponding semilog ROC plots for four different screens are shown in Figure 4 for comparison, with weightings applied based on several levels of UNITY fingerprint clustering in each case. A linear ROC plot for HIV-1 protease inhibitors is shown as an inset.

The arithmetically weighted ROC (blue line) is shifted to the right for the artificial data set that is biased by construction (Figure 3A), but is only sharpened somewhat for the unbiased one (Figure 3B). Harmonic weighting (green line), in contrast, shifts the ROC curve slightly to the right for the artificially biased case but to the left for the unbiased one.[16] The trends seen for HIV protease inhibitors and angiotensin antagonists (Figure 4A and 4B) are qualitatively similar to the result seen in Figure 3A, indicating that these validation sets are biased, a conclusion that is underscored by the increasing shift when the number of positive classes (clusters) is decreased. The curves for $5HT_{1a}$ agonists and cyclooxygenase inhibitors, in contrast, mirror the behavior of those for the unbiased artificial data set (Figure 4C and 4D, respectively).

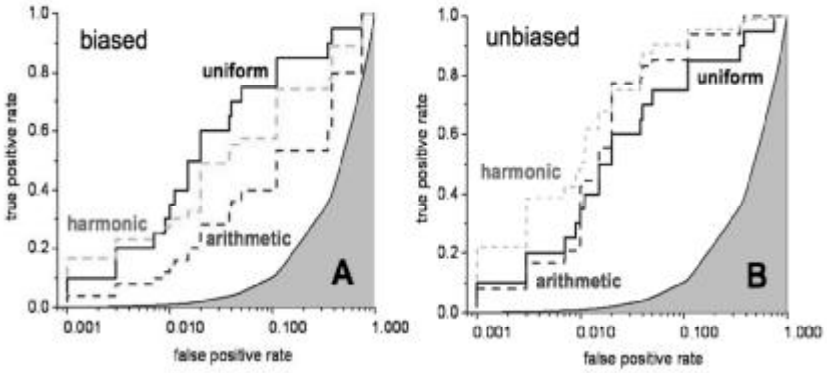

Figure 3. Nominal positives from artificially constructed data sets are plotted using uniform (solid black lines), arithmetic (broken blue lines) of harmonic (broken green lines) weighting schemes. The shaded gray curve corresponds to random recovery of positives. (A) Biased data set, in which high-scoring positives are found in large clusters. (B) Unbiased data set, in which positives are distributed evenly across clusters regardless of score.

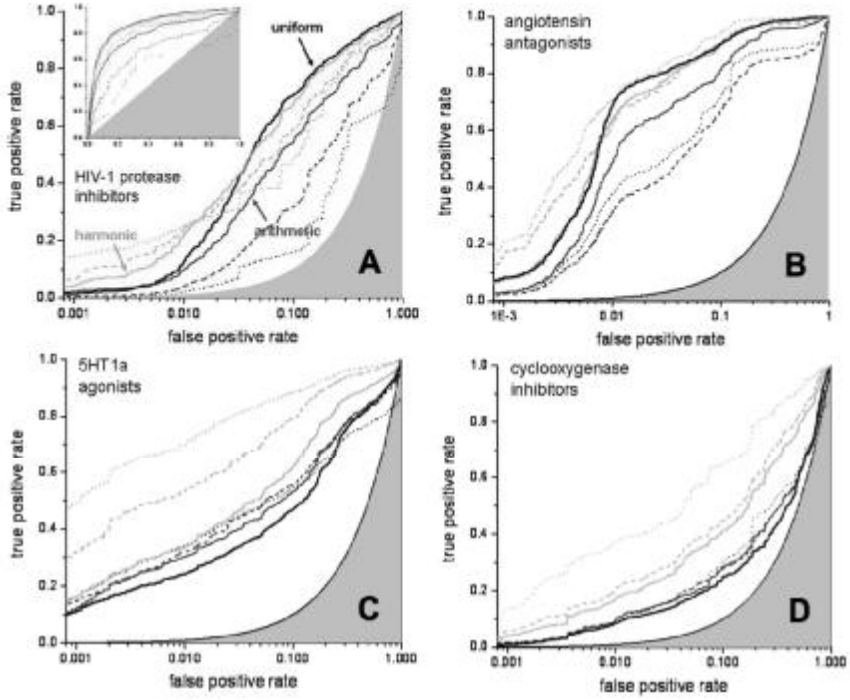

Figure 4. Results for FCFP_4 screens run against the MDDR. The ROC curves obtained are shown with the “true positive” rate α calculated using uniform (black), arithmetic (blue) or harmonic (green) weighting at three clustering levels; the weight of each line

indicates the coarseness of the clustering. The shaded gray curve corresponds to random recovery of positives. (A) HIV-1 protease inhibitors. (B) $5HT_{1a}$ agonists. (C) Angiotensin II AT1 antagonists. (D) Cyclooxygenase inhibitors.

Fig. 5 shows the weighted AUC and pAUC values obtained for those targets for which the number of positives was large enough to make clustering practical. The values shown reflect progressive consolidation of clusters as one moves from left to right, with the right-most bar corresponding to about 10 clusters in each case. As one would hope, the indications of bias that are qualitatively evident in the full ROC curves in Fig. 4 are mirrored in the summary statistics shown Figure 5. The arithmetic AUC (Figure 5A) decreases sharply with increasing cluster size in most cases (the biased ones) where the harmonic AUC (Figure 5C) is relatively stable but is insensitive to cluster size in most cases where the harmonic AUC increases (the unbiased ones) with increasing cluster size. These trends are even more readily apparent in the logarithmic pAUCs (Figure 5B and 5D). Not surprisingly, some validation sets clearly fall into one or the other category vis à vis bias, whereas others (e.g., docking into COX-2 and the FCFP screen for thrombin inhibitors) lie somewhere between the two extremes.

It is possible, though unlikely, that a query or docking target will come from a small cluster of positives. Were this to occur, the ROC curve would be shifted to the *left* by arithmetic weighting and the corresponding AUC and pAUC would increase. This may explain some cases where vHTS performance falls significantly below the value expected for guessing at random. It is not the case for docking into acetylcholinesterase, however; arithmetic weighting only makes the AUC worse (Figure 5A). It seems more likely that poor performance in this case reflects the unusual nature of the ligand, a small one that bears a positive charge but is not a hydrogen bond donor.

The nominal performance statistics are much better for FCFPs, which encode the atomic environments of generalized atom types: hydrogen bond donors and acceptors; positive and negative ionizable centers; aromatic atoms; and

halogens.[27] They are more closely related to substructural fingerprints than are distance-based pharmacophores, though they focus on distinctive groups (piperazines, tetrazoles, amidines etc.) rather than on literal substructures. It is not surprising, then, to find that their ROC statistics are generally biased upwards with respect to the overall structural similarity assessed by UNITY fingerprint similarity (Figure 5). It is interesting to note that the performance statistics for thrombin, a target that appears in both sets of screens, converge once the differing number and relative structural diversity of thrombin positives have been adjusted for by clustering (Figure 5A and 5B). Results for cyclooxygenase are similar, though the interpretation in that case is complicated by the fact that the two enzyme isoforms are split out in the docking validation set but not in the MDDR classifications. That FCFP and Surflex-Dock performances converge suggests that bias is at least as much an attribute of the target – and the known positives that are available – as of the screening method.

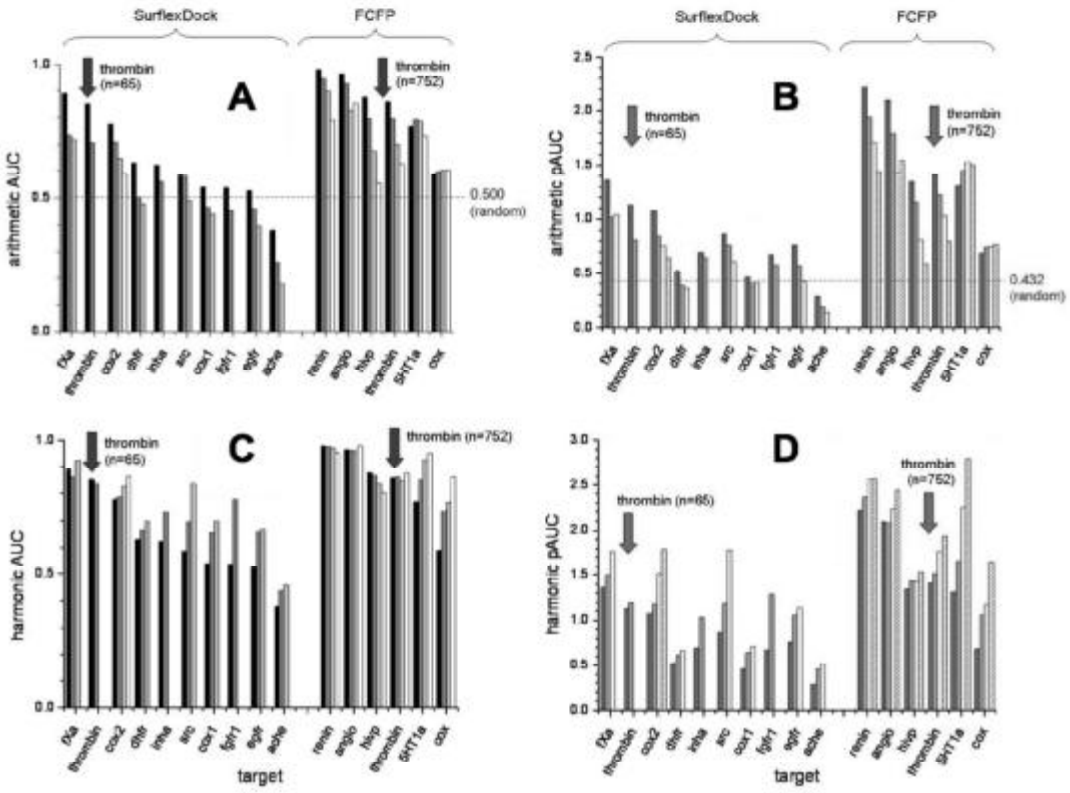

Figure 5. Effect of data set bias on linear and logarithmic AUCs. Data sets large enough to cluster meaningfully were evaluated by linear (black shading to gray) or logarithmic (red shading to yellow) integration of the arithmetically and harmonically weighted ROCs. For each target, the cluster size increases from left to right; conversely, the number of clusters decreases in the same sense. In each case, the lowest clustering level obtained was comprised of about ten clusters. (A) Linear integration of arithmetic weighting. (B) Logarithmic integration of arithmetic weighting. (C) Linear integration of harmonic weighting. (D) Logarithmic integration of harmonic weighting.

The degree and kind of bias detected will necessarily depend on how positives are clustered. The consistency of the trends seen here for each target as the number of clusters is varied suggests that results are not likely to be terribly sensitive to the details involved, e.g., to the particular fingerprint implementation employed or to the similarity metric used.[38,39] That said, fundamentally different types of descriptors can be expected to yield somewhat different results. Atom pair fingerprints, for example, encode the topological separation between atom types[40] rather than the presence or absence of particular substructures that is encoded by substructural fingerprints. Clusters constructed using either type of fingerprint give similar results for *fXa*, for example, but less bias is evident for *cox2* when the less localized atom pair descriptor is used (details not shown). Which approach is most appropriate will be dictated by the kind of novelty that is of most value in the application of interest. In particular, a similarity metric based on central ring structures[41] may an appropriate choice for virtual screening of protein kinase inhibitors, for which direct hydrogen bonding interactions between the ligand scaffold and the target's protein backbone are particularly critical for these targets.[42] More generally, however, the multiple levels of granularity afforded by hierarchical clustering is likely to be more productive.

4. CONCLUSION

If a reasonably large number of known positives are available that represent several different positive subclasses, examining the effect of arithmetic weighting on (p)AUCs makes it possible to minimize the risk of settling on a particular

model because of biased validation results. In this case, the risk is opting for a screening tool that would constitute a slow and expensive way to address the much simpler sub-classification problem of structural or property similarity.[14,20,21] Conversely, examining the effect of harmonic weighting can be used to identify screens that effectively complement the relatively trivial similarity searches that one is likely to run as follow-up to any virtual screen. Taken together, they also help identify validation sets that are *negatively* biased in a particular situation, e.g., because its domain of applicability extends into areas beyond those represented in the test set by accidents of history or because a chosen target or query is unduly incompatible with the positives in the validation set.

The two types of virtual screens to which weighted ROC analysis has been applied here are fundamentally different, yet both exhibit a similar range of bias. Moreover, the biases seen for the shared targets – thrombin and cyclooxygenase – are similar, which supports the intuitive expectation that bias is likely to reflect incidental effects specific to one or a few targets and, hence, need not be a general problem for a particular method. It also implies that related approaches may be useful for identifying sampling bias in other validation and calibration work.

Acknowledgments

The Authors would like to express their gratitude to our anonymous reviewers for providing their analysis of and comments on the original manuscript.

REFERENCES

[1] Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins: Struct. Funct. Bioinform.* **2004**, *57*, 225-242.

[2] Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. III. Assessing Scoring Functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032-3047.

[3] Warren, G. L.; Andrews, C. W.; Capeli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912-5931.

[4] Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789-6801.

[5] Jain, A. N. Morphological Similarity: A 3D Molecular Similarity Method Correlated with Protein-Ligand Recognition. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 199-213

[6] Cuissart, B.; Touffet, F.; Cremilleux, B.; Bureau, R.; Raul, S. The Maximum Common Substructure as a Molecular Depiction in a Supervised Classification Context: Experiments in Quantitative Structure/biodegradability Relationships. *J. Chem. Info. Comput. Sci.* **2002**, *42*, 1043-1052.

[7] Jain, A. N. Ligand-Based Structural Hypotheses for Virtual Screening. *J. Med. Chem.* **2004**, *47*, 947-961.

[8] Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534-2547.

[9] Nicholls, A. What Do We Know and When Do We Know It? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239-255.

[10] Clark, R. D. Getting Past Diversity in Assessing Virtual Library Designs. *J. Brazil. Chem. Soc.* **2002**, *13*, 788-794.

[11] Clark, R. D.; Fox, P. C. Statistical Variation in Progressive Scrambling. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 563-576.

[12] Rohrer, S. G.; Baumann, K. Impact of Benchmark Data Set Topology on the Validation of Virtual Screening Methods: Exploration and quantification by Spatial Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 704-718.

[13] Egan, J. P. *Signal Detection Theory and ROC Analysis*; Academic Press: New York, NY, 1975.

[14] Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD Technique Performance: A Virtual Screening Case Study in the Design of Validation Experiments. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 529-536.

[15] Good, A. C.; Oprea, T. I. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: a Help or Hindrance in Tool Selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169-178.

[16] Clark, R. D.; Webster-Clark, D. J. Managing Bias in ROC Curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141-146

[17] Irwin, J. J. DUD: A Directory of Useful Decoys. http://dud.docking.org (accessed July 15, 2007).

[18] Jain, A.N. Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-based Search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281-306.

[19] *Surflex-Dock*, version 2.3; Biopharmics LLC: San Francisco, CA, 2008.

[20] Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793-806.

[21] Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369-1375.

[22] MDL Drug Data Report. http://www.mdli.com/products/pdfs/mddr_ds.pdf (accessed 2March 25, 2008).

[23] Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J.-J.; Lecchini, S.; Jacoby, E. An Ontology for Pharmaceutical Ligands and its Application for in silico Screening and Library Design. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947-955.

[24] Hert, J.; Willett, P.; Wilton, D. J. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 462 -470.

[25] Pearlman, R. S.; Rusinko, A.; Skell, J. M.; Balducci, R. *Concord*; Tripos International: St. Louis, MO, 2008.

[26] *SciTegic Pipeline Pilot*, version 6.1; Accelrys: San Diego, CA, 2008.

[27] Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-based

Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256-3266.

[28] Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem Inf. Comput. Sci.* **1997**, *37*, 71-79.

[29] *SYBYL*, version 7.8; Tripos International: St. Louis, MO, 2008.

[30] *UNITY*, version 7.8; Tripos International: St. Louis, MO, 2008.

[31] Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley-Interscience: New York, NY, 1990.

[32] Yates, F. Contingency Tables Involving Small Numbers and the χ^2 Test. *J. Royal. Stat. Soc.* **1934**, *1*, 217-235

[33] Snedecor, G.W.; Cochran, W.G. *Statistical Methods*, 8th ed.; Iowa State Press: Ames IA, 1989; pp 119-120.

[34] Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488-508.

[35] Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches . *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395-1406.

[36] Liebeschuetz, J. W. Evaluating Docking Programs: Keeping the Playing Field Level. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 229-238.

[37] Jain, A. N. Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281-306.

[38] Cannon, E. O.; Mitchell, J. B. O. Classifying the World Anti-doping Agency's 2005 Prohibited List Using the Chemistry Development Kit Fingerprint. *Lecture Notes Comput. Sci.* **2006**, *4216*, 173-182.

[39] Haranczyk, M.; Holliday, J. Comparison of Similarity Coefficients for Clustering and Compound Selection. *J. Chem. Inf. Model.* **2008**, *48*, 498-509.

[40] Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Simon K. Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136

[41] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **2008**, *48*, 704-718.

[42] Ghose, A. K.; Herbertz, T.; Pippin, D. A.; Salvino, J. M.; and Mallamo, J. P. Knowledge Based Prediction of Ligand Binding Modes and Rational Inhibitor Design for Kinase Drug Discovery. *J. Med. Chem.* **2008**, *48*, 5149-5171.