The Use of Consensus Scoring in Ligand-Based Virtual Screening

J. Christian Baber, William A. Shirley, Yinghong Gao, and Miklos Feher*

Neurocrine Biosciences, 12790 El Camino Real, San Diego, California 92130

Received July 21, 2005

A new consensus approach has been developed for ligand-based virtual screening. It involves combining highly disparate properties in order to improve performance in virtual screening. The properties include structural, 2D pharmacophore and property-based fingerprints, scores derived using BCUT descriptors, and 3D pharmacophore approaches. Different approaches for the combination of all or some of these methods have been tested. Logistic regression and sum ranks were found to be the most advantageous in different pharmaceutical applications. The three major reasons consensus scoring appears to enrich data sets better than single scoring functions are (1) using multiple scoring functions is similar to repeated samplings, in which case the mean is closer to the true value than any single value, (2) due to the better clustering of actives, multiple sampling will recover more actives than inactives, and (3) different methods seem to agree more on the ranking of the actives than on the inactives. Furthermore, consensus results are not only better but are also more consistent across receptor systems.

INTRODUCTION

Virtual screening has now become an essential element in the drug discovery process.[1] Although the most impressive success stories from virtual screening are based on docking with knowledge of the 3D receptor structure,[2,3] there are also successful ligand-based virtual screening approaches.[4,5] Numerous methods for virtual screening have been developed with many techniques having similar characteristics. Consequently, the selection of the most appropriate method for a given application is often challenging. Consensus scoring, or data fusion, methods have a lot to offer in this regard as they can potentially improve performance by merging results from different methods. Consensus scoring has been widely applied in receptor-based screening to combine data from different docking algorithms.[6−9] In ligand-based drug discovery consensus methods have been most widely used to merge similarity scores and molecular descriptors.[10−14] The consensus decisions in these approaches are generally based on voting or some combination of rankings,[10,11] the generation of consensus fingerprints,[12] neural nets,[13] or the use of conditional probability.[14] Additionally, consensus approaches have also been applied in different QSAR studies.[15−17]

In our ligand-based virtual screening process at Neurocrine, different similarity algorithms are typically used in parallel with diversity-based measures and 3D pharmacophores. The format of the results are highly dissimilar: some provide continuous values, others just binary yes/no answers, some score only positive outcomes, and others also differentiate negative results. Our aim in this work was to investigate the feasibility of merging these highly disparate results using a consensus approach to obtain superior performance.

There are different expectations from a successful consensus score. The most important requirement is that based on the common wisdom from several different methods, the errors in single scores would be balanced, allowing active molecules in a drug discovery program to be more efficiently identified. It appears from simulated models[18] that the improvement on using scores in a consensus arrangement might simply arise because the mean value of repeated samplings tends to be closer to the true value. This would be useful in itself; however, our observation was that the methods applied in this current work tend to agree more on the identity of true positives than on negatives, leading to improved enrichments. A successful consensus scoring scheme should also reduce the uncertainties regarding the selection of the appropriate virtual screening approach in a given project. Since a number of methods are combined in consensus scoring, the inclusion of a low performance method should be less of a problem than if this method was applied on its own. The consensus scheme might also incorporate the relative significance of different methods applied, and these weights might change at different stages in the drug discovery process. It is often observed in practical drug discovery projects that different screening methods generalize differently: some techniques are tolerant to structural modifications, whereas others might not be able to identify actives in a structure class even slightly different from those in the training set. It is clearly advantageous if the consensus scheme can generalize well by recognizing different indicators for activity from the disparate methods. On the other hand, inadequacies of scoring functions might be amplified in a consensus scheme if they are significantly correlated,[19] and it is very important to make sure that this does not happen. Finally, it is expected that the scoring scheme can be used in an automated fashion within a drug discovery program. Our aim was to devise a scheme that satisfies all of these criteria.

METHODS

Scoring Systems. A number of individual virtual screening methods were tested and examined for inclusion in the consensus scheme. First, different similarity schemes were

* Corresponding author phone: (858)617-7630; fax: (858)617-7619; e-mail: mfeher@neurocrine.com.

10.1021/ci050296y CCC: $33.50 © 2006 American Chemical Society
Published on Web 11/02/2005

considered. The two criteria for inclusion were that the scheme had to provide high performance when applied on its own, and the selected schemes had to encapsulate molecular information differently from each other. The similarity schemes selected were all implemented in MOE[20] and included the 166-bit MACCS fingerprint,[21] a 3-point pharmacophore-based fingerprint calculated on the molecular graph (typed graph triplets or TGT[20]) and MP61, a molecular property based fingerprint using 61 binary encoded descriptors.[22] These similarity measures encapsulate structural, pharmacophore, and gross property information of the molecule, respectively. During the virtual screening process, each screened molecule was compared to every molecule in the training set using the appropriate metrics and the maximum Tanimoto overlap (maximum average Tanimoto overlap in case of MP61) of their keys assigned as the score for that compound. The process was made fully automated using the scientific vector language (SVL) within MOE.[20]

BCUT descriptors,[23,24] as implemented within the Diverse Solutions (DVS) software,[25] capture similarity/diversity in molecular properties and also encode substructural, topological, and atomic information. A multidimensional chemistry space can be generated by carefully selecting those BCUT descriptors that best represent the structural diversity of a population of compounds. For this work the most appropriate 2D descriptors available in DVS were selected using the ‘auto-choose’ algorithm resulting in a five-dimensional chemistry space that optimally separated an exemplar set of known drugs and internal compounds.

The idea of the BCUT-based scoring technique developed in this work was to identify the location of a compound with unknown activity in BCUT chemistry space and use its proximity to known exemplars to estimate the likelihood of the compound being active. However, rather than use simply the distance to the closest neighbor—as is done in many similarity-based techniques—it was decided that as many close neighbors as possible should be taken into account.

For maximum speed, the process was divided into two steps: generation of a grid and scoring of compounds. The aim of grid creation is to summarize information from multiple active and inactive exemplar compounds. The grid itself is created by dividing the chemistry space axes into 10 equal sections—in a manner similar to the way that DVS assigns BCUT cell numbers. Compounds are then placed on the grid based on their BCUT coordinates. A base contribution is calculated for each compound depending on its activity—with compounds more active than 100 nM having an increasingly positive contribution and those less active than 1 μM an increasingly negative one. Rather than act on a single grid point, the effect of each compound extends to a maximum of three points away in each dimension. The contribution to each grid point within that range is calculated by taking the base contribution and dividing by one plus the square of the distance between the compound and the grid point of interest. These contributions are summed for each of the active and inactive exemplar compounds and normalized to give a final score for each grid point. The calculated grid can then be stored and used for scoring any number of compounds.

Since most of the processing is carried out at the grid creation stage, the scoring of individual compounds is relatively fast and simple. The BCUT coordinates are first calculated for each compound. The scores for each point within three units of these coordinates are then extracted from the grid file. Compound scores are computed using a method similar to the reverse of that used to calculate grid scores. The score assigned to a compound is equal to the normalized sum of the contributions from each of the surrounding grid points. As with grid creation these contributions are calculated by dividing the score of each grid point by one plus the square of the distance to the compound being scored. Positive scores generated for a compound indicate proximity and thus similarity to actives in BCUT chemistry space, whereas negative scores signify that the nearby space is mainly occupied by inactive exemplars. A score of zero can mean one of two things: either that nothing is known about that area of chemistry space—neither active nor inactive exemplars being in close proximity—or that a mix of actives and inactives are present. The calculated scores were tested and found to be insensitive to small changes in grid placement.

The above 2D methods do not capture steric, chirality, and conformation dependent information: this was achieved using 3D pharmacophores. Two program suites were applied for this purpose: Catalyst[26] and CombiCode.[27] These were selected because of their greatly different handling of pharmacophore information and their general lack of correlation in predictions, especially for negative data (see below). Catalyst hypotheses were generated using the HypoGen method[28] using training sets containing both active and inactive molecules, covering at least 4 orders of magnitude in activity. (Unlike in most other applied methods, activity information in the form of a K_i value was incorporated in Catalyst model generation.) Conformers were generated within Catalyst using the ‘best’ method during the development of pharmacophores and the ‘fast’ method when screening validation sets (to gauge ‘real-life’ screening performance). Because the test examples in each project covered large numbers of molecules in different structural classes, pharmacophore behavior could only be characterized using an ensemble approach. Thus pharmacophore hypotheses were clustered and ranked, and the best clusters were accepted for screening. In the screens, the Catalyst score for each compound was equal to the number of satisfied pharmacophore hypotheses.

3D pharmacophore modeling was also performed using ensemble pharmacophores within the CombiCode toolkit as implemented in Python.[27] Within CombiCode, approximately 200 representative low-energy conformers were rapidly generated for each molecule using the CONAN program.[29] Next a rule-based method (encoded in a Feature Definition Language[30,31]) was used to classify the atoms in each conformation as pharmacophore types (hydrogen bond donor, acceptor, positive charge, lumped hydrophobic group, and aromatic ring). 3D pharmacophore signatures (fingerprints) were generated using the Pharmer program[32,33] and were collected in an ensemble.[34] As part of the signature specification, distance bins were defined, and the pharmacophore distances were assigned to these bins. As it was believed that a 1.2 Å resolution corresponds to the coarseness of the conformational analysis, 16 bins were defined between 1.6 and 20.8 Å. To reduce imposed artifacts by the discrete analysis, the boundaries of these bins were kept fuzzy, so that compounds within 0.12 Å of a bin boundary were placed in both bins. Next all 4-point spatial arrangements in the defined tetrahedron distance space (considering chirality) with all possible feature combinations were enumerated, and solutions that violate the triangle inequality (and consequently describe physically impossible tetrahedrons) were discarded. It was found in the current work that the inclusion of 4-point pharmacophores containing primarily hydrophobic points leads to memorization of the training data set, hence only those pharmacophores were retained that had at least two hydrogen-bonding or charge features. The program Signal was used to evaluate the (binary) correlation of each model with activity by use of information content,[34,35] and the top ranked hypotheses were selected for the ensemble models. As an improvement to the predictive ability of the method as well as to reduce the likelihood of spurious models, an extension to the published CombiCode methods that we call Feature-Based Scoring (FBS) was developed. In this process, the ensemble of conformers for each compound is evaluated to find the conformer that simultaneously satisfies the largest number of 4-point models and spatially presented the most features. We scored compounds based on the total number of unique (and desired) pharmacophoric features that their best conformers presented.

Training and Validation Set Selection. The selection of training and validation sets for the four receptors used in this work was carried out in the following manner. Molecules were first protonated/deprotonated using the ‘wash’ process in MOE[20] to ensure a consistent representation. Compounds were then clustered using MACCS fingerprints, so that fewer than 10 compounds remained in singleton clusters. Subsequently, a diversity ranking for each of the compounds was generated using a modified version of MOE’s Diverse Subset Selection routine. In this process the singleton compounds were selected as the starting set, and ranking was achieved by sequentially identifying the compound most different (using MACCS/Tanimoto similarity) from those already selected.

Compounds were then selected for the primary training set according to the determined rank order. Once the required number of compounds had been selected, Diverse Subset Selection (as implemented in MOE) was again applied to rerank the remaining compounds. Compounds were then selected for the secondary training set as described above, and then the process was repeated for the validation set.

Compounds for training and validating different methods were assembled from the Neurocrine corporate database. Antagonists for four different G-protein coupled receptors (GPCRs) were considered: corticotropin-releasing factor receptor-1 (CRF), gonadotropin releasing hormone receptor (GnRH), melanocortin receptor-4 (MC4), and melanin-concentrating hormone receptor (MCH). Three different data sets were generated for each receptor: a primary training set, a secondary training set, and a validation set. Only the primary training sets were used to develop a model for each method of scoring, the secondary training sets were then used to determine how the individual models could best be combined to give a consensus score, and the validation sets were used to examine and compare the performance of both the individual models and the consensus scores. The selection process described above allowed for the most diverse compounds available at each stage to be selected into each set. The actual number of compounds used in each set was chosen based on the number available both in-house and from literature. The target number of molecules for each set was 1000 for each of the training sets and 2000 for the validation set, with 25% actives in training set 1 and 10% actives in training set 2 and the validation set. A larger number of actives was chosen for training set 1 as more exemplars are useful for generating good models for some methods (particularly BCUT-based scoring), although it is generally possible to produce models using far fewer known actives. To enable binary decisions, molecules in the training set were designated as active with K_i values of less than 100 nM, whereas molecules with less than 50% inhibition at 10 μM concentration in binding assays were defined as inactives.

Evaluating Performance. It is quite difficult to judge the true performance of a virtual screening method, as the different figures of merit of the process critically depend on the data set selection. The two extreme tests generally applied in the literature are finding a few actives in a database of druglike decoy molecules[36] or using a highly focused set of actives and inactives.[37] It has been argued that the latter process is more meaningful as the separation of actives and inactives in the former case might arise simply due to 1D properties (such as molecular weight or lipophilicity).[37] It can also be argued that the decoy test is artificial in that the task is rarely to separate low nanomolar ligands from structurally dissimilar inactives. Also, the test of searching among similar molecules is more demanding and approximates real virtual screening better. Hence in this work all methods were tested using this latter approach.

During the course of the work different methods for quantifying enrichment were tested. The enrichment factor (EF) was defined as the proportion of true positives in the predicted positive set divided by the proportion of positives in the background set.[38] EF ranges between its maximum possible value, called theoretical enrichment ($\mathrm{EF}_{\mathrm{theor}}$), and zero, with the value for random selection being 1. This formula is useful to judge the performance within one data set, but the theoretical enrichments are different in different sets. To allow performance comparisons, we defined relative enrichment

$$\mathrm{EF}_{\mathrm{rel}} = \frac{\mathrm{EF} - 1}{\mathrm{EF}_{\mathrm{theor}} - 1} \qquad (1)$$

$\mathrm{EF}_{\mathrm{rel}}$ is 1 for the theoretical enrichment, 0 for random selection, and negative for worse than random selection. Although this value is normalized only for better than random performance, this is acceptable as this is the territory in which we are interested.

Relative enrichments are useful in that they provide a single number to assess the performance in virtual screening. However, enrichments change depending on the recovery rate with higher recovery generally providing worse enrichments than lower values. In this work, we looked at relative enrichments at 50% and 90% recovery. A more informative way to describe performance is to plot the true positive rate as a function of the false positive rate for different score thresholds of the given virtual screening method. This plot is often referred to as the ROC (Receiver Operating Characteristic) plot.[39] The two main advantages of ROC plots are that they consider information on both actives and inactives and the fact that they are independent of the ratio

Table 1. Consensus Scoring Methods Tested in This Work

method	brief description	comments on performance
voting	Pass/fail criteria for each method, decision based on number of passes for molecule.	Overall performance similar to other methods but it is often difficult to get high recovery rates with good enrichment. Not pursued due to arbitrary nature of pass/fail criteria.
rank voting	Each method has a predefined number of votes for activity; the top ranking compounds are assigned those votes.	Best performance found when each method has the number of votes equal to 1.5 times the number of actives. In practical cases, the number of actives is unknown, although the number of compounds required may be.
simple sum ranks	Ranking compounds on each property, ranks are added.	Selected as the method of choice if little data are available.
weighted sum ranks	The ranks from each method weighted by the performance (*R*%) of the method for the given receptor.	Extra effort in determining weights, only marginally better performance than with simple sum ranks. Not pursued further.
multilinear regression	Coefficients determined from enrichments or overall performance (*R*%).	Inappropriate behavior due to discontinuous nature of activity data (only active/inactive labels).
logistic regression	Fitting logistic curve to data, coefficients are determined from a second training set and are dependent on method's performance.	Selected as the method of choice if sufficient data are available. This method had generally the best overall performance.

of actives in the examined set. In the ROC plots random performance corresponds to the diagonal, and a curve with ideal behavior would follow the left and then the top side of the plot. In our work we found that the performance of different methods on different data sets can be compared more easily if we characterize them using a quantity we called randomness (denoted by *R*):

$$R(\%) = 100\cdot\frac{\text{area over ROC curve}}{\text{area over diagonal}} \qquad (2)$$

Randomness defined in this manner is 0% for an ideal behavior (all actives recovered before any of the inactives), 100% for a completely random behavior, and greater than 100% in the case where a method performs worse than random. An advantage of the randomness descriptor is that it is integral in nature, i.e., it characterizes the overall performance of each method at all recovery rates of the actives. In contrast, the enrichment factors give the performance of the methods at a given recovery rate that may be useful to compare and select methods in specific virtual screening situations.

Combining Scores. When trying to fit different types of scores together, it is important to consider output ranges and if necessary make them compatible. The scores from the three similarity measures are already normalized, due to the nature of the definition of Tanimoto overlap. As described above, predicted activity in BCUT scoring was properly normalized, and, although negative values were also possible (predicted inactivity), these do not lead to incompatibility with other scores. The greatest difficulty is with the 3D pharmacophore scores as each of the hypotheses (single pharmacophores) make binary decisions on whether to accept or reject a compound. In order for pharmacophore ensembles in both methods to generate more possibilities to distinguish compounds, the number of satisfied hypotheses was used as the total pharmacophore score instead of simple pass/fail decisions. As the number of hypotheses/features depended on the receptor system under consideration, it ranged between 3 and 15 for the two methods and different receptors.

Different algorithms were evaluated to select the best methods for generating consensus scores, and these are briefly summarized in Table 1. In essence, the two best methods selected in this study were sum ranks and logistic regression. Sum ranks, often applied in data fusion,[10] are derived by ranking compounds on every property and simply adding the ranks. Generally somewhat better results were obtained using logistic regression,[40] which is the regression method of choice if the dependent variable is discrete because the variance of the dependent variable is different for different values of the independent variables. This is different from the assumption in classical regression that the error term must be independent of the variables. Furthermore, in logistic regression the error terms are not assumed to be normally distributed. The logistic regression model can be thought of as a specific nonlinear linear regression model, in which the logistic curve relates the independent variable *x* to the estimated probability, *P*, in the following expression

$$P = 1/[1 + \exp(-a - bx)] \qquad (3)$$

where *a* and *b* are the parameters of the model that are being fitted. This represents a sigmoidal relationship between probability and the coefficients. In this approach, the optimal coefficients are determined using the maximum likelihood estimation method.[40] It is important to note that all performance data using logistic regression in this work was obtained after properly fitting the coefficients for the given receptor and set of methods applied, i.e., coefficients were never transferred from one model to another. Sum ranks require no training (at least when the weights of all methods are considered equal), whereas the coefficients for the logistic regression were derived using an independent (second) training set.

In tasks related to regression, collinearity of the variables often presents problems, partly due to redundancy and partly because it may introduce poor fit statistics or instability in fit coefficients.[41] Due to the nature of the data in this work (continuous data with skewed distribution mixed with categorical data) only nonparametric statistics are appropriate to describe its correlation properties. Thus, instead of the more usual product/moment (Pearson) method, correlation of the descriptors for each receptor system was examined using Spearman rank correlation, as implemented within the NAG statistical library[42] as an Excel add-in.[43] This implementation also includes a correction for tied ranks. Spearman ρ is analogous to the Pearson correlation coefficient *r* and approximates it well[44] for continuous descriptors (such as

Table 2. Spearman Rank Correlation Coefficients (ρ) for the Methods Applied in This Work, Averaged over the Receptor Systems[a]

	MACCS	TGT	MP61	BCUTs	Catalyst	Pharmer
MACCS	1.00					
TGT	0.66	1.00				
MP61	0.71	0.64	1.00			
BCUTs	0.49	0.52	0.49	1.00		
Catalyst	0.54	0.59	0.55	0.48	1.00	
Pharmer	0.50	0.52	0.49	0.46	0.52	1.00

[a] The four receptor systems considered were GnRH, MCH, MC4, and CRF. The maximum ρ value for an individual receptor was 0.75, the minimum 0.17. See text for further details.

the fingerprints and BCUT properties in this work). The ρ values averaged over the four receptors in this work are given in Table 2. It should be noted that the correlation coefficients for the Spearman rank correlation matrices for the individual receptors showed a similar distribution, with the highest overall ρ value being 0.75 (analogous to a Pearson r^2 of 0.57). In conclusion, the applied methods generally correlate poorly with each other, and hence their application in a consensus arrangement is justified.

RESULTS AND DISCUSSIONS

Performance of Individual Methods. The performance of the individual methods applied in this work can be compared using the ROC curves shown in Figure 1 and the tabulated performance indicators in Table 3. The measures shown in this table for each receptor are the quantity randomness, as defined by eq 2, as well as the relative enrichments at 50% and 90% recovery of the actives. These values are shown for the four GPCRs considered, with their means also given. As can be seen, fingerprint-based methods performed far better than the property-based BCUT's method or 3D pharmacophores. In particular, MACCS keys were consistently better than any other measure, followed by the 2D pharmacophore and property fingerprints. It must be mentioned, however, that the selection of training and test sets in this work was based on maximum diversity as measured by MACCS similarity, which might have somewhat improved the performance of this metric in comparison with others. From the two 3D pharmacophore-based schemes, Pharmer appeared to perform somewhat better on the tested data sets than Catalyst, although some of its shortcomings (such as a break-down at high recovery rates) are noted below. The BCUTs method generally provided the least enrichment, although for the CRF receptor it performed better than either of the 3D pharmacophore methods. These results confirm previous observations[45−47] that MACCS similarity contains the greatest amount of information relevant to ligand−receptor binding and that 2D similarity measures generally perform better than 3D pharmacophores.

The calculated relative enrichments reflect the same trends as the *R*% values, although the relative performance of the methods strongly depends on the recovery rate considered. When the performance is measured for the recovery of only half of the active molecules, the structural and 2D pharmacophore keys display nearly theoretical enrichment with practically no difference between them, whereas at 90% recovery of the actives, differences between these methods are much more pronounced.

One odd case occurred with Pharmer for the CRF receptor. In this example, Pharmer performed very well for the first 75% of actives. However, as no matching pharmacophoric features could be found for approximately the last 20% of actives, these were assigned the same score as approximately 90% of the inactives−resulting in a straight line on the ROC chart. This also led to a relative enrichment of 0 at 90% recovery of the actives. A similar behavior but to a much lesser extent could also be seen for Pharmer for the MC4 and GnRH receptors as well as in case of the Catalyst method for the MC4 and MCH receptors and the BCUTs method for CRF. This behavior might arise as models generated using the actives in the training set are unable to make correct predictions for certain classes in the test set. As the training and test sets were selected carefully in this work, it is likely that this effect is due to the limited capability of these pharmacophore models to generalize around a small number of exemplars. An alternative explanation is that a certain pharmacophore might be common among a certain set of actives and inactives with inactivity being caused by some other property (e.g. the inactives having some extra features). In case of the Pharmer method, one further possibility to explain this failure is that it selects pharmacophore hypotheses based on the individual performance rather than the collective performance of the ensemble. It is likely that hypothesis ensemble selection should improve the performance and diminish the described effect.

It is also interesting to inspect the consistency of the performance of these methods across different receptor systems, as presented in Table 3. As an example, the BCUTs method has almost half the *R*% value as Catalyst in case of the CRF receptor, comparable to it for MC4 and it is more than twice as bad for GnRH. Similar relative differences in the behavior of the other methods can be observed for different receptors. However, the superiority of the 2D fingerprints over BCUTs and the 3D pharmacophores appears to be consistent across all receptors studied.

Overall Performance of Consensus Schemes. Two schemes to combine scores from different methods were tested in detail in this work: logistic regression and sum ranks. Results with these schemes are presented in Tables 3 and 4, and the performance of the consensus score in comparison with single scores is displayed in Figure 1. It is immediately clear from both the table and the figure that the consensus scores generally worked better than any of the single scores. The only example where this was not the case is the MC4 receptor. Here MACCS keys produced almost theoretical enrichment, and the introduction of a consensus with other worse performing scores somewhat reduces the overall performance.

It is interesting to compare the results with the two applied consensus schemes. It was found that sum ranks generally provided an improvement over single methods (with the exception of the MC4 case) but were generally inferior to logistic regression. In terms of the randomness property (*R*%), as defined in eq 2, logistic regression consistently outperformed sum ranks by 0.5−1.5% across all method combinations and receptor systems, which corresponded to the randomness of the former being 50−75% of that of the latter in the same systems.

Leave-One-Out and Leave-Two-Out Experiments. Although the methods applied in this work correlate weakly

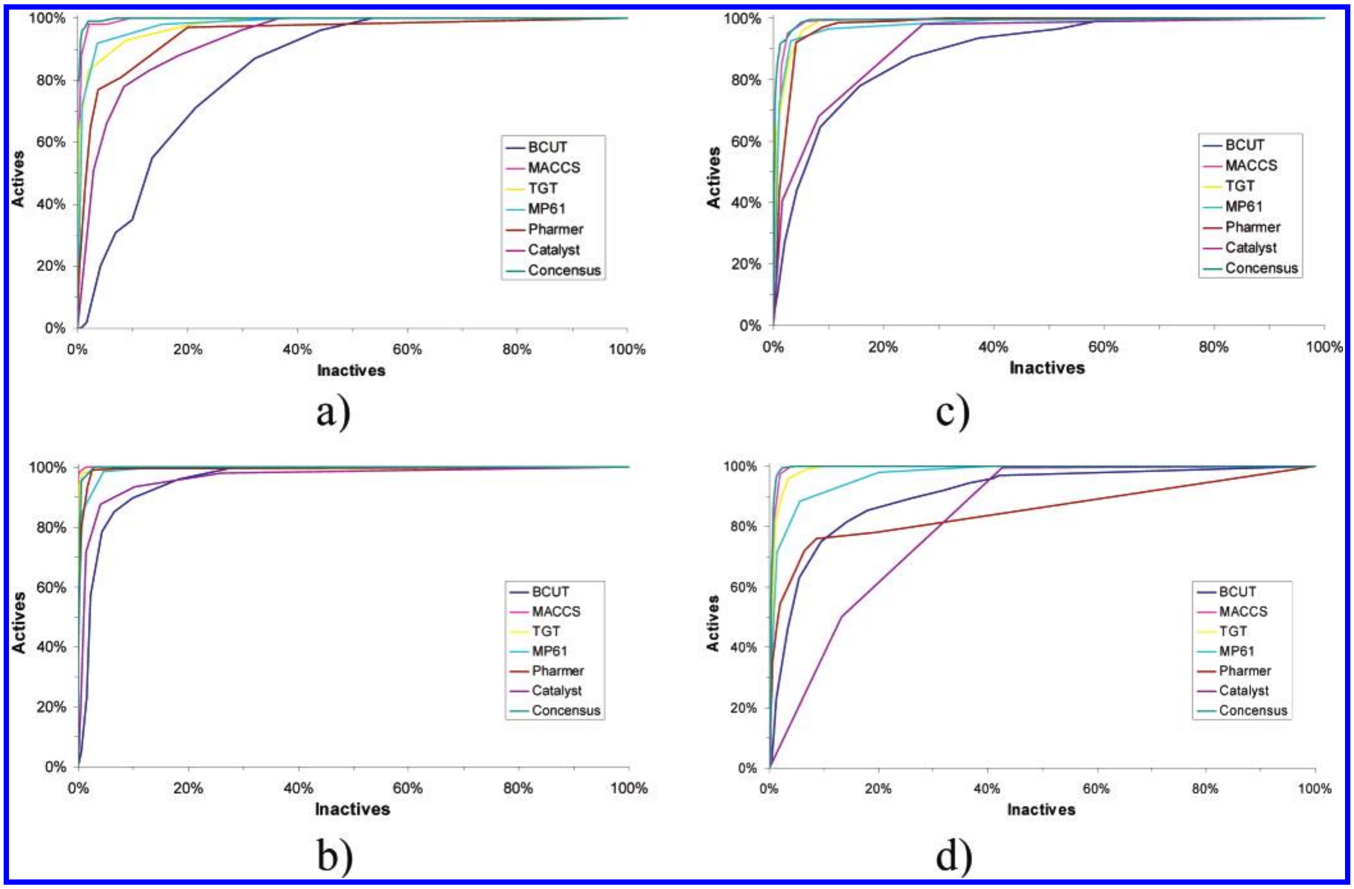

Figure 1. ROC curves (true positives against false positives) for different virtual screening methods—MACCS similarity, 2D pharmacophore (TGT) similarity, property fingerprint (MP61) similarity, BCUT descriptors, Catalyst and Pharmer pharmacophores—as well as the consensus score derived from these using logistic regression for the receptors: (a) GnRH, (b) MC4, (c) MCH, and (d) CRF.

Table 3. Performance of Different Virtual Screening Approaches on the Validation Sets[a]

	GnRH			MC4			MCH			CRF			mean		
	R (%)	EF$_{rel}$ 50%	EF$_{rel}$ 90%	R (%)	EF$_{rel}$ 50%	EF$_{rel}$ 90%	R (%)	EF$_{rel}$ 50%	EF$_{rel}$ 90%	R (%)	EF$_{rel}$ 50%	EF$_{rel}$ 90%	R (%)	EF$_{rel}$ 50%	EF$_{rel}$ 90%
MACCS	0.8	0.96	0.83	0.0	1.00	1.00	2.0	0.91	0.81	1.2	0.93	0.88	**1.0**	0.95	0.88
TGT	3.9	0.94	0.46	0.8	1.00	1.00	2.4	0.92	0.73	1.8	0.91	0.80	**2.2**	0.94	0.75
MP61	3.5	0.82	0.59	0.7	1.00	0.92	3.7	0.86	0.76	5.5	0.82	0.41	**3.3**	0.87	0.67
BCUTs	31.4	0.13	0.08	7.9	0.69	0.43	21.9	0.45	0.17	17.3	0.55	0.19	**19.6**	0.46	0.22
Pharmer	10.5	0.58	0.16	1.3	0.99	0.85	4.0	0.79	0.68	29.6	0.74	0.00	**11.3**	0.77	0.42
Catalyst	13.5	0.45	0.13	6.9	0.83	0.45	16.5	0.42	0.21	35.0	0.22	0.12	**18.0**	0.48	0.23
logistic regression all	0.3	1.00	0.91	0.4	0.98	0.96	1.2	0.97	0.89	0.85	0.96	0.91	**0.7**	0.98	0.92
sum rank all	0.5	0.98	0.93	0.2	1.00	1.00	1.6	0.95	0.81	1.5	0.93	0.86	**0.9**	0.96	0.90

[a] R is randomness, expressed as a percentage of the area over the ROC curve relative to the total area above the diagonal, as defined by eq 2, EF$_{rel}$ 50% is relative enrichment (as defined by eq 1) for the recovery of 50% of the active compounds, and EF$_{rel}$ 90% is relative enrichment for the recovery of 90% of the actives. For the description of the symbols representing different methods and all other information, see text.

with each other in terms of ranking molecules (see Table 2), they may nevertheless contribute to the consensus model to a different extent. In this context, leave-one-out experiments help to establish the importance of individual models and identify methods that may contribute less to the consensus or may be redundant.

The results of the leave-one-out experiments are given in Table 4, together with the results with the single scoring functions for comparison. First, it can be seen that leaving out any one method, even the best performing MACCS similarity, has relatively little impact on the results. It has been shown above that methods in this work were not linearly correlated in a pairwise manner (at least in terms of Spearman rank correlation). Given the fact that the exclusion of any of the methods from the full consensus has approximately the same impact, the effect may be due to either multicollinearity (i.e. that the linear correlation of any five of the methods might be able to explain the sixth) or simply to the fact that five methods applied in consensus had such a high performance that no significant improvement on adding a sixth method could be observed. (This behavior might also be expected due to statistical reasons, as described below.).

To rule out multicollinearity, selected leave-two-out experiments were also performed. In these experiments, combinations not involving one of the 3D pharmacophore methods are the most interesting from a practical perspective, given that these are time-consuming to calculate. Of the two tested, Catalyst is more widely applied, hence experiments excluding Pharmer and another method are presented in

Table 4. Performance (*R*%) of a Combination of Different Virtual Screening Approaches on the Validation Sets[a]

method	GnRH	MC4	MCH	CRF	mean	range[b]
MACCS	0.8	0.0	2.0	1.2	**1.0**	1.9
TGT	3.9	0.8	2.4	1.8	**2.2**	3.1
MP61	3.5	0.7	3.7	5.5	**3.3**	4.8
BCUTs	31.4	7.9	21.9	17.3	**19.6**	23.4
Pharmer	10.5	1.3	4.0	29.6	**11.3**	28.2
Catalyst	13.5	6.9	16.5	35.0	**18.0**	28.1
MACCS+TGT	0.6 (1.3)	0.0 (0.3)	1.9 (1.6)	1.0 (0.9)	**0.9 (1.0)**	1.8
MACCS+MP61	0.5 (0.9)	0.0 (0.1)	1.8 (2.0)	1.1 (1.9)	**0.9 (1.2)**	1.8
MACCS+Pharmer	0.6 (1.2)	0.0 (0.1)	1.3 (1.4)	1.2 (3.0)	**0.8 (1.4)**	1.3
MACCS+Catalyst	0.6 (2.3)	0.0 (0.5)	1.7 (2.7)	1.2 (5.2)	**0.9 (2.7)**	1.6
TGT+MP61	1.7 (1.7)	0.2 (0.4)	2.1 (2.0)	1.5 (1.5)	**1.4 (1.4)**	1.9
BCUTs+Catalyst	14.2 (15.9)	2.5 (2.6)	11.9 (12.8)	8.7 (8.4)	**9.3 (9.9)**	11.7
MACCS+TGT+MP61	0.6 (0.9)	0.0 (0.2)	1.8 (1.6)	1.2 (1.1)	**0.9 (0.9)**	1.7
MACCS+TGT+MP61+BCUTs	0.6 (1.6)	0.0 (1.1)	1.6 (1.8)	0.8 (1.1)	**0.8 (1.4)**	1.5
MACCS+TGT+MP61+Catalyst	0.5 (0.5)	0.0 (1.2)	1.5 (1.4)	1.2 (1.6)	**0.8 (1.2)**	1.5
TGT+MP61+Catalyst+BCUTs	1.0 (1.7)	0.2 (1.4)	1.8 (2.6)	1.0 (1.7)	**1.0 (1.8)**	1.7
all-MACCS	0.6 (1.0)	0.2 (0.3)	1.3 (1.9)	1.1 (1.9)	**0.8 (1.3)**	1.1
all-TGT	0.2 (1.2)	0.2 (0.1)	1.2 (2.1)	0.9 (2.1)	**0.6 (1.4)**	1.0
all-MP61	0.6 (1.3)	0.5 (0.3)	1.3 (1.9)	1.3 (2.0)	**0.9 (1.4)**	0.9
all-BCUTs	0.3 (0.3)	0.4 (0.3)	1.1 (1.2)	0.9 (1.7)	**0.7 (0.8)**	0.9
all-Catalyst	0.4 (1.0)	0.0 (0.1)	1.2 (1.5)	0.9 (1.3)	**0.6 (1.0)**	1.2
all-Pharmer	0.5 (0.8)	0.0 (0.2)	1.5 (1.8)	0.9 (1.3)	**0.7 (1.0)**	1.5
all	0.3 (0.5)	0.4 (0.2)	1.2 (1.6)	0.8 (1.5)	**0.7 (0.9)**	0.9

[a] *R* is randomness, expressed as a percentage of the area over the ROC curve relative to the total area above the diagonal, as defined by eq 2. Where two numbers are given, the one in the brackets was obtained by using sum-ranks, the other one using logistic regression. For the description of the symbols representing different methods and all other information, see text. [b] Range of *R*% values from logistic regression.

Table 4. It can be seen that the performance generally decreases on leaving out Pharmer and another method (in comparison to just leaving out only that single method) by a small amount. There are a few exceptions, however, when a small improvement can be observed on leaving out Pharmer, especially in the CRF and MC4 cases. These issues are probably connected to memorization problems, discussed above.

Another interesting leave-two-out experiment involves leaving out MACCS and another descriptor. It was found that, although the other applied descriptors generally cannot fully compensate the loss of MACCS, the decrease in consensus performance is generally small. When MACCS is removed from the consensus score, not surprisingly, the other fingerprint methods (MP61 and TGT) become the most important followed by Pharmer and then BCUT and Catalyst. This generally follows the same order as the performance of individual methods with the most notable exception being the increased importance of MP61. This agrees quite well with the pair wise results where the MACCS+MP61 combination was found to perform relatively poorly indicating that the two methods may use some of the same information. If this is the case, then leaving both methods out of the consensus should have a greater effect than expected from leaving either one out individually. In conclusion, it can be seen from these tests that multicollinearity among the descriptors applied is unlikely to be an issue.

Pairwise Tests. The performance of all possible method pairs was also investigated, with some of the most interesting results presented in Table 4. From these results it can be concluded that the highest enrichments could be achieved if one of the methods employed was MACCS. This is not too surprising, given that MACCS consistently provided the highest enrichments from the methods tested in this work. However, it is interesting to note that the method most complementary to MACCS was found to be Pharmer followed by TGT and Catalyst even though the 3D pharmacophore methods did not perform particularly well by themselves. One other conclusion from the results is that using two methods in consensus generally leads to some improvement even if the individual methods do not perform too well. This can be seen very clearly from the Catalyst/BCUTs combination in Table 4 (other data not shown). In this case, both Catalyst and BCUTS perform relatively poorly—with mean randomness scores of 18% and 19.6%, respectively. However, a consensus score generated from both methods has a much lower mean randomness (9.3%)—almost half the value of the best of the two component methods. A further interesting observation from the results is that certain method pairs perform better than other pairs, even if the individual methods themselves are worse (as is the case e.g. for the MCH receptor and the MACCS/MP61 pair, compared to the MACCS/TGT pair).

Using Only Fingerprint Methods. Further tests were performed involving only the 2D fingerprint methods. This is important for practical reasons, especially because of their simplicity, ease of use, computational speed, and performance. Results with the simultaneous use of MACCS, TGT, and MP61 fingerprints are displayed in Table 4. This combination appears to provide high enrichments, only minimally worse than the consensus involving all methods (and even slightly better in case of MC4). As these fingerprints require only the 2D structure, large numbers of compounds can be screened rapidly with them.

In pharmaceutical applications, it is often necessary to move away from problematic molecular cores (to avoid physicochemical problems, unfavorable pharmacological properties, or IP issues). MACCS keys were originally developed for database look-up[48] in order to find 'similar' molecules quickly. Thus, despite its high performance in database enrichment, this method might not be suitable in

Table 5. Performance (R%) of Different Approaches under More Demanding Conditions[a]

	GnRH	MC4	MCH	CRF	mean	range
MACCS	25.7	5.2	17.6	5.9	13.6	20.5
TGT	40.3	13.6	20.7	9.0	20.9	31.3
MP61	34.2	14.8	30.1	40.5	29.9	25.7
MACCS+TGT	24.4 (28.5)	4.2 (4.6)	13.8 (14.3)	3.6 (3.6)	11.5 (12.8)	20.8 (25.0)
MACCS+MP61	24.2 (23.5)	3.9 (4.9)	15.5 (16.7)	–[b] (14.8)	12.4[c] (14.9)	20.2[c] (18.6)
TGT+MP61	27.8 (27.5)	8.8 (7.9)	17.8 (18.2)	7.1 (11.9)	15.4 (16.4)	20.7 (19.6)
MACCS+TGT+MP61	23.5 (23.0)	3.6 (4.1)	13.3 (14.0)	–[b] (6.2)	11.0[c] (11.8)	19.9[c] (19.0)

[a] R is randomness, expressed as a percentage of the area over the ROC curve relative to the total area above the diagonal, as defined by eq 2. Where two numbers are given, the one in the brackets was obtained by using sum-ranks, the other one using logistic regression. The training set contained only 20 compounds, selected using diversity in all 3 fingerprints and the criterion for activity in the validation set was $< 1\ \mu$M. For the description of the symbols representing different methods and all other information, see text. [b] Logistic regression failed in these cases, due to very high standard error determined for the MP61coefficient. [c] Mean and range do not include CRF, due to the failure of the logistic fit.

applications where the aim is to enrich the results with molecules that are structurally dissimilar from the original structures. In this regard, the TGT/MP61 combination is highly favorable: neither method would preferentially select molecules structurally similar to the original one (even though these are similar e.g. in a pharmacophore sense). The results in Table 4 indicate that these two methods in combination performed only slightly worse than MACCS similarity. These properties are easy to calculate and hence they can be fruitfully applied in projects in which lead hopping is desired. Indeed such experiments have been performed in Neurocrine using a combination of this scoring and virtual screening as well as de novo design[49] and have identified a number of useful new hits.

Performance under More Demanding Conditions. To test the performance of consensus scoring under more difficult conditions, tests using modified training and validation sets were also undertaken. The training set in this case included only 20 compounds (this might emulate the beginning of lead-finding in a medicinal chemistry project) that were selected from the original training set based on diversity in all three fingerprints. In the validation set, compounds were accepted as actives if their K_i's were below 1 μM. As expected, the performance of all methods under these conditions was significantly lower than in the previous experiments (see Table 5). As before, in most cases the performance with consensus scores was better than that of individual methods and logistic regression outperformed sum ranks. The notable exception was CRF for all consensus scores including MP61. In these cases the sum rank was worse than the best individual method. Interestingly, the existence of this problem was indicated in two of the three cases by the fact that the logistic regression itself also failed: the error of the fitted coefficients for MP61 was much greater than the coefficients themselves. Hence it would appear that failure to fit a logistic regression curve might be an indicator for the lack of performance with this consensus scheme.

How Does Consensus Scoring Work? Although consensus scoring has gained wide acceptance from the docking community and in certain other areas of QSAR and virtual screening, relatively few studies deal with the question how consensus scoring enriches data sets. The issue was studied by Wang and Wang[18] using an idealized computer simulation. The conclusion from these experiments was that the performance increase in consensus scoring can be explained simply by the fact that the mean of repeated samplings tends to be closer to the true value than any single sampling. Based on their simulation they concluded that the improvement in performance scales with the square root of the number of applied scoring functions and that no improvement beyond the variance of the data can be expected if more than 3–4 scoring functions are simultaneously used.

To understand the applicability of the simulation results[18] to the current work, it is important to consider the assumptions used in those simulations. These include the following: (1) The error in binding affinity estimation is solely determined by the scoring function. (2) The error in binding affinity estimation is a random number with normal distribution, centered at zero. (3) All scoring functions are completely orthogonal. Unfortunately, it is easy to see that none of these assumptions is satisfied in the current work. First, none of our methods attempted to approximate the binding energy, e.g. the 2D fingerprint methods only indicate the likelihood of activity. Second, the errors from different scores in this work are not centered at zero. Third, as shown using Spearman rank correlation above, while the methods applied in this work do not correlate strongly with each other, they are far from being orthogonal. Furthermore, the scoring functions in the simulations had equal performance, with the magnitude of the random error being large compared to the randomly assigned binding affinities. In contrast, real scoring functions, such as the ones applied in this work, might have vastly different performances. As a result of all these points, some differences in performance on using multiple scoring functions are expected in a real case compared to the purely statistical improvement in the simulations.

Using a simple voting scheme we can gain some insight into why the efficiency of scoring improves when different methods are used in consensus. Let us assume that each method can cast a vote for activity for as many of the top compounds as there are known active compounds in those data sets. A similar general behavior under these circumstances was observed in all studied data sets, and the results are exemplified by the GnRH case, as displayed in Table 6, which was compiled somewhat similarly to the simulated results in Table 4 in ref 18. On comparing the trends in the two tables, we can see a gradual decrease in the number and percentage of compounds that get zero votes in both works and an even faster decrease in the number of actives with zero votes. On the opposite end of the scale, the number of compounds receiving activity votes from all methods decreases on increasing the number of scoring functions, whereas the ratio of actives among these increases. There is

Table 6. Hit Rates Observed for the GnRH Receptor in a "Rank-by-Vote" Experiment (Voting for Activity)[a]

	number of scoring functions[b]											
votes	1		2		3		4		5		6	
0	12 (1899)[c]	0.6%	6 (1876)	0.3%	3 (1845)	0.2%	2 (1811)	0.1%	1 (1764)	0.1%	1 (1696)	0.1%
1	88 (101)	87.1%	21 (47)	44.7%	12 (65)	18.5%	7 (82)	8.5%	1 (116)	0.9%	1 (172)	0.6%
2			73 (77)	94.8%	26 (27)	96.3%	13 (25)	52.0%	9 (25)	36.0%	9 (37)	24.3%
3					59 (63)	93.7%	39 (43)	90.7%	30 (34)	88.2%	23 (27)	85.2%
4							39 (39)	100.0%	47 (49)	95.9%	43 (45)	95.6%
5									12 (12)	100.0%	21 (21)	100.0%
6											2 (2)	100.0%

[a] The data set contains 100 known actives and 1900 inactives. For each method, the results are ranked and an activity vote is cast for the top 100 compounds. This table demonstrates the amount of overlap between such predictions from different methods as a function of the number of scoring functions applied. [b] To get results for different numbers of methods, scoring functions were added consecutively according to their quality, observed in the data sets in this work (see Table 3) in the order MACCS > TGT > MP61 > Pharmer > Catalyst > BCUTs. This was done in order to avoid an apparent improvement with an increasing number of scoring functions simply because the added scoring function is better than any of the previous ones. [c] The first number denotes how many true actives were captured by the given number of votes, whereas the number in brackets gives the total number of compounds with that many votes. Thus e.g. the first entry means that 'by having only one scoring function (MACCS), there are 12 true actives among the 1899 compounds that received zero votes, corresponding to 0.6% of such compounds'.

Table 7. Median Values for the Pairwise Tanimoto Similarities within the Actives and the Inactives, as Calculated Using Different Methods and for Different Receptors[a]

		GnRH	MCH	MC4	CRF
MACCS	actives	0.78	0.74	0.79	0.65
	inactives	0.52	0.51	0.44	0.46
TGT	actives	0.84	0.91	0.88	0.84
	inactives	0.76	0.77	0.71	0.74
MP61	actives	0.89	0.91	0.91	0.86
	inactives	0.82	0.83	0.80	0.81

[a] Mean values were also calculated and are all within 0.02 of the median. See text for the description of the applied methods and other details.

a notable difference, however, in the simulations: the ratio of actives among the total number of hits that received all of the votes reaches a maximum of ∼60% with 5 functions, whereas it rapidly reaches 100% in this work. Similarly, the ratio of true hits receiving the maximum minus one vote also tends toward 100% on increasing the number of scoring functions. If the performance improvement on applying an increasing number of scoring functions solely arose as a result of the fact that the mean of repeated samplings tends to be closer to the true value than any single sampling, we would expect a slow increase in performance in hit rates with a high number of votes. Clearly, the applied methods contain uncorrelated systematic errors, and hence the improvement in performance on adding an extra method to the consensus score is greater than would be if the errors were due purely to random noise.

The calculated pairwise Tanimoto similarities among the active and inactive molecules are collected in Table 7. This shows that the internal similarities among actives are consistently and significantly higher than among inactives, despite the fact that the inactive molecules were originally synthesized and tested because the medicinal chemists thought that they were likely to be active. Because actives are more tightly clustered than inactives, multiple samplings should repeatedly select many actives but not necessarily the same inactives. This behavior is displayed schematically in Figure 2. It is believed that this explanation based on the different clustering of actives and inactives as well as the purely statistical explanation (as given earlier[18]) are together responsible for the performance of consensus scoring.

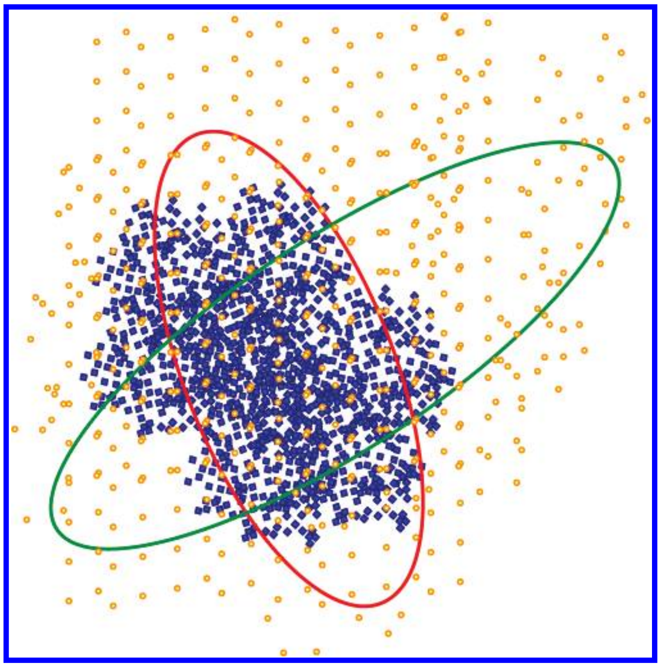

Figure 2. A schematic representation of how consensus scoring might improve screening performance. In arbitrary property space, inactives (shown as orange dots) are thought to be more randomly distributed than actives (shown as blue diamonds) that will cluster together. If a cluster of actives is sampled by two different scoring functions (shown as green and red ellipsoids), these will both recover actives. Since useful scoring functions perform well, different methods will vote for some of the same actives. The cross section of the two ellipsoids (i.e. the compounds picked up by both scoring functions) will be enriched in actives more than if the actives and inactives were equally well distributed. However, even in the latter case there would be enrichment, due to the effect of multiple sampling.

In this regard, it is important to point out an apparent contradiction. For a scoring function to have any practical utility in a given situation, it must have better than random performance, and hence all such functions have an increased likelihood of selecting the same actives. Thus, in statistical terms, real scoring functions must be somewhat correlated, unlike the ones used in the simulations. In the simulations the magnitude of the error was large compared to the value being predicted, leading to the scoring functions being completely independent and hence possibly voting for completely different actives. This slight correlation should

Table 8. Concordance of the Methods Used in This Work for Actives and Inactives in Different Data Sets[a]

	actives	inactives
GnRH	0.82	0.54
MC4	0.90	0.60
MCH	0.94	0.45
CRF	0.49	0.47
average	**0.79**	**0.51**

[a] Using Kendall's method of concordance (*W* value), identifying the concordance of votes by the 6 methods used in this work (MACCS, TGT, MP61, BCUTs, Catalyst, and Pharmer). The computed values of *W* are all highly significant at the 1% level ($p<0.0001$), indicating that the null hypothesis of there being no agreement between the six rankings may be rejected with reasonably high confidence. Further details are given in the text.

ensure that there will be an overlap in the sampled actives among different methods that will contribute to a performance increase on top of the purely statistical one, as illustrated schematically in Figure 2. On the other hand, it is important that the applied methods must not be strongly correlated, particularly in their errors, otherwise they are no longer independent. It is likely that properly considering the level of correlation when selecting methods that work well together is key to improving operational performance in consensus scoring, especially by ensuring that systematic errors are not correlated between methods.

A recent publication[50] on receptor-based consensus scoring appears to support the above arguments. The authors used Flex-X to dock molecules into the blood coagulation factor Xa pocket and then rescored these poses using several different scoring functions. Consensus scoring was applied in the form of a voting scheme, finding the intersection of hit lists from different methods. The authors found[50] that consensus scoring failed to improve the hit rates achieved by single scoring functions. This failure can be understood by recognizing that some of the above-described requirements were not met. First, there was a systematic component to the error, common across all methods: molecules were docked by a single docking engine, and hence errors in ligand placement affected all scoring functions. For example, any inactives for which 'good' poses were found were likely to be recognized by the different scoring functions as such. This introduces some correlation in false positives. On the other hand, as stated by the authors, there was little correlation in true positives between many of the applied scoring functions (namely Flex-X and PMF hit lists barely overlapped with others). These findings together explain why consensus scoring brought little or no improvement in that work.[50]

Although our scoring functions do not attempt to approximate the actual binding affinity, it is interesting to see how well they agree in ranking active and inactive compounds. These were estimated (see Table 8) using Kendall's method of concordance, *W*,[44,51] a value that was specifically developed to gauge the degree of agreement between different rankings. It can be seen that the methods applied in this work appear to provide more similar ranking for active compounds than for inactives. This difference strongly supports the previous point and explains the high performance of consensus scoring, although it is unknown at this point how general this latter observation is. This also shows that the errors have less correlation than correct predictions, amplifying the strengths of these methods but not their weaknesses.

A further advantage of using consensus scoring may be seen by examining the errors across different receptors, as shown by the ranges for *R*% value (the inverse of the quality of prediction) in Table 4. The results show that consensus scores are generally more tightly grouped around the mean error than single techniques, and the range across various receptors is smaller. This means that the consensus results are not only better but they are also more consistent across receptor systems. Thus, when using consensus schemes, the user is less dependent on picking the correct method for each receptor than when using single methods.

CONCLUSIONS

The use of different ligand-based virtual screening methods using consensus scoring is described in this work. The described work is novel, given the high diversity of the methods employed in consensus. The results show that consensus scoring provides enrichments consistently higher than its component methods. The performance from consensus scores also displays far lower variation across receptors than individual methods, reducing the importance of selecting precisely the right method for a given system. Hence, consensus scoring is promising in ligand-based virtual screening applications. Further work will be required to identify other possible virtual screening methods that would complement the ones described in this work and also the optimal combination of methods to be applied in different real-life scenarios. It would also be interesting to be able to predict from the results with single scoring functions how well a combination of methods will perform in a consensus arrangement.

In pharmaceutical practice, virtual screening might be used in different scenarios. The two applications where the described methods might be particularly useful are lead optimization and lead hopping (identifying structurally different compounds that bind to the same target). In the lead optimization stage, a high amount of data is already available, making it possible to generate highly predictive consensus models. Most of the results presented indeed simulate this stage in drug discovery, at which hundreds of active and inactive molecules that fall into a small number of structural classes are known. Due to the high performance of the MACCS method, using one or several other methods in consensus with MACCS scoring was found to perform well in this situation. Best performances were obtained using the three fingerprints together or in combination with one 3D pharmacophore method. A completely different scenario is presented when lead-hopping is the objective. In these cases, the application of MACCS scoring is undesirable because it generally identifies compounds structurally similar to the query compounds. In such cases, the use of different combinations of methods excluding MACCS scores might provide the best results. In this regard, the use of the TGT/MP61 combination in particular has so far appeared to be most promising in our drug discovery work.

It is also important at this point to reflect at the different applicability of the two consensus schemes tested: logistic regression and sum ranks. In the initial phases of a lead finding or scaffold hopping project there is relatively little data available. Sum ranks are useful at this point to rank and prioritize compounds for synthesis/testing or compound purchase, especially that only the scores from the individual methods are required. What this method does not provide are absolute prediction criteria for activity to indicate, what percentage of the virtual library might be worth testing, or the ability to compare directly different sets without having to combine them and recalculate the ranks. This kind of information is readily available when logistic regression is applied. The enrichments with logistic regression are also somewhat higher, but the downside is the large amount of data required to establish a model, which might not be available in the initial stages of the project.

The ligand-based scoring system in the current work was developed especially for GPCR-type receptors, since 3D structures are generally unavailable for this class. However, none of the applied techniques would limit the application of ligand-based consensus scoring to GPCRs, indeed it would be applicable to any other possible binding target. The enrichments attained in this work using consensus scoring were high, despite the fact that by performing the tests using highly similar actives and inactives lower enrichments are expected than in tests involving decoy ligands as inactives.[37] In fact, although the enrichments in this work are not directly comparable to previous receptor-based docking studies, our results appear to be closer to the theoretical maxima. Thus, it might be advantageous to apply ligand-based consensus scoring even if the receptor structure is available, perhaps even in consensus with some docking score. We showed in this paper that it is possible to merge successfully scores from disparate sources. It is expected that such an application will further improve the efficiency of receptor-based virtual screening.

ACKNOWLEDGMENT

The assistance of Jun Ma in the performance of some of the calculations and the compilation of results is gratefully acknowledged.

REFERENCES AND NOTES

(1) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.

(2) Green, D. V. Virtual screening of virtual libraries. *Prog. Med. Chem.* **2003**, *41*, 61−97.

(3) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *20*, 1047−1055.

(4) Stahura, F. L.; Bajorath, J. Virtual Screening Methods that Complement HTS. *Comb. Chem. High Throughput Screen* **2004**, 259−269.

(5) Langer, T.; Hoffmann, R. D. Virtual screening: an effective tool for lead structure discovery? *Curr. Pharm. Des.* **2001**, 509−27.

(6) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(7) Paul, N.; Rognan, D. ConsDock: A New Program for the Consensus Analysis of Protein−Ligand Interactions. *Proteins* **2002**, *47*, 521−533.

(8) Clark, R. D.; Strizev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002**, *20*, 281−295.

(9) Gohlke, H.; Klebe, G. Statistical Potentials and scoring functions applied to protein−ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231−235.

(10) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.

(11) Salim, N.; Holliday, J.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.

(12) Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 21−29.

(13) Baurin, N.; Mozziconacci, J. C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276−285.

(14) Raymond, J. W.; Jalaie, M.; Bradley, M. P. Conditional probability: a new fusion method for merging disparate virtual screening results. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 601−9.

(15) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004**, *47*, 2356−2364.

(16) Dearden, J. C. In silico prediction of drug toxicity. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 119−127.

(17) Mattioni, B. E.; Kauffman, G. W.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 949−963.

(18) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.

(19) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nature Rev. Drug Discovery* **2004**, *3*, 935−948.

(20) MOE (Molecular Operating Environment) version 2004.03; Chemical Computing Group Inc.: 1010 Sherbrooke St. West, Suite 910, Montreal, Quebec H3A 2R7, Canada.

(21) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL Keys as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(22) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151−1157.

(23) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(24) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 339−353.

(25) DiverseSolutions, version 6.2; Optive Research, Inc., 12331-A Riata Trace Parkway, Suite 110, Austin, TX 78727, U.S.A.

(26) Catalyst, version 9.1; Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, U.S.A.

(27) CombiCode is a suite of program and libraries in Python provided in an interface known as Pyxie. It is available from Deltagen Research Laboratories.

(28) Li, H.; Sutter, J.; Hoffmann, R. HypoGen: An Automated System for Generating 3D Predictive Pharmacophore Models. In *Pharmacophore Perception, Development and Use in Drug Design*; International University Line: La Jolla, CA, 2000.

(29) Smellie, A.; Stanton, R.; Henne, R.; Teig; S. Conformational Analysis by Intersection: Conan. *J. Comput. Chem.* **2003**, *24*, 10−20.

(30) Bush, B. L.; Sheridan, R. P. PATTY: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.

(31) Greene, J. W.; Kahn, S.; Sprague, P.; Savoj, H.; Teig, S. L. Chemical function queries for 3D Database search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(32) Barnum, D.; Greene, J. W.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(33) Penzotti, J.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A computational ensemble pharmacophore model for identifying substrates of P-Glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.

(34) Lanctot, J. K.; Putta, S.; Lemmen, C.; Greene, J. W. Using ensembles to classify compounds for drug discovery. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2163−2169.

(35) Srinivasan, J.; Castellino, A.; Bradley, E. K.; Eksterowicz, J. E.; Grootenuis, P. D. J.; Putta, S.; Stanton, R. V. Evaluation of a novel shape based computational filter for lead evolution: Application to Thrombin inhibitors. *J. Med. Chem.* **2002**, *45*, 2494−2500.

(36) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(37) 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.

(38) Pearlman, D. A.; Charifson, P. S. Improved Scoring of Ligand-Protein Interactions Using OWFEG Energy Grids. *J. Med. Chem.* **2001**, *44*, 502−511.

(39) Zweig, M. H.; Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*, 561−577.

(40) Tabachnick, B. G.; Fidell, L. S. *Using multivariate statistics*, 4th ed.; Allyn & Bacon: Boston, MA, 2000; ISBN: 0321056779.

(41) Livingstone, D. *Data analysis for chemists*; Oxford University Press: Oxford, 1995.

(42) NAG Fortran Library, Routine G02BNF, Numerical Algorithm Group Ltd., Oxford, U.K.

(43) NAG Statistical Add-Ins for Excel, Numerical Algorithm Group Ltd., Oxford, U.K.

(44) Norman, G. R.; Streiner, D. L. *Biostatistics: The bare essentials*, 2nd ed.; Dekker Inc.: Hamilton, 2000.

(45) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(46) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(47) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atomic Environments, Information-Based Feature Selection and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(48) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(49) To be published.

(50) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 333−344.

(51) NAG Fortran Library, Routine G08DAF, Numerical Algorithm Group Ltd., Oxford, U.K.

CI050296Y