Retrospective Docking Study of PDE4B Ligands and an Analysis of the Behavior of Selected Scoring Functions

Chidochangu P. Mpamhanga,[†] Beining Chen,*[,†] Iain M. McLay,[‡] Daniel L. Ormsby,[§] and Mika K. Lindvall[‡]

Department of Chemistry, University of Sheffield, Dainton Building, Brookhill, Sheffield S3 7HF, U.K., GlaxoSmithKline Medicines Research Centre, Computational and Structural Sciences, Stevenage, Hertfordshire SG1 2NY, U.K., and Accelrys Ltd., 334 Cambridge Science Park, Cambridge CB4 0WN, U.K.

Received February 4, 2005

Scoring forms a major obstacle to the success of any docking study. In general, fast scoring functions perform poorly when used to determine the relative affinity of ligands for their receptors. In this study, the objective was not to rank compounds with confidence but simply to identify a scoring method which could provide a 4-fold hit enrichment in a screening sample over random selection. To this end, LigandFit, a fast shape matching docking algorithm, was used to dock a variety of known inhibitors of type 4 phosphodiesterase (PDE4B) into its binding site determined crystallographically for a series of pyrazolopyridine inhibitors. The success of identifying good poses with this technique was explored through RMSD comparisons with 19 known inhibitors for which crystallographic structures were available. The effectiveness of five scoring functions (PMF, JAIN, PLP2, LigScore2, and DockScore) was then evaluated through consideration of the success in enriching the top ranked fractions of nine artificial databases, constructed by seeding 1980 inactive ligands (pIC50 < 5) with 20 randomly selected inhibitors (pIC50 > 6.5). PMF and JAIN showed high average enrichment factors (greater than 4 times) in the top 5−10% of the ranked databases. Rank-based consensus scoring was then investigated, and the rational combination of 3 scoring functions resulted in more robust scoring schemes with (cScore)-DPmJ (consensus score of DockScore, PMF, and JAIN) and (cScore)-PPmJ (PLP2, PMF, and JAIN) yielding particularly good results. These cScores are believed to be of greater general application. Finally, the analysis of the behavior of the scoring functions across different chemotypes uncovered the inherent bias of the docking and scoring toward compounds in the same structural family as that employed for the crystal structure, suggesting the need to use multiple versions of the binding site for more successful virtual screening strategies.

1. INTRODUCTION

Advances in structural biology have enabled the determination of many high quality protein structures for potential drug targets. To use this information most effectively for drug discovery it is essential to develop reliable methods for virtual high throughput screening (VHTS).[1−3] Within commercial organizations the key objective is commonly to identify for biological screening a hit-rich subset of compounds from a very large corporate collection (e.g. 1 million compounds). The subset could be 10 000 compounds or more. Virtual screening could be used to rank the collection and hence identify such a subset.

The final goal of the biological screening would be to discover at least one lead series for further investigation. The situation is different for academic institutions, and small biotech companies, in which automation is rare and compounds must be purchased or synthesized for screening. Generally for such organizations it is possible to screen a small number of compounds only (<500). The aim of this present study was to identify a method of VHTS with efficiency suitable to ensure that screening of <500 compounds represented a reasonable chance of success.

It is impossible to predict the hit rate for any particular biological screen when presented with compounds selected at random. However, past experience in GlaxoSmithKline suggests an average hit rate of around 0.05%. For this purpose, a hit would be defined as a genuinely active compound showing a reproducible concentration−response curve. Based upon this figure, and without enrichment, the screening of 500 compounds could be expected to provide a lead on one occasion in four. To screen 500 compounds, with a more reasonable chance of identifying a lead, the database would need to be ranked in such a way that the top 500 have a hit rate of 0.2% (i.e. one active compound in the 500 or a 4-fold enrichment). This would be the minimum enrichment required to provide confidence that the exercise of screening 500 compounds, in the pursuit of a lead, had a reasonable chance of success. In the studies reported here the primary goal was to assess if such enrichment is achievable for the PDE4B protein target. A secondary goal was to explore the series dependency of docking and scoring, specifically with consideration given to how the results depend on the nature of the molecules existing in the source collection and the source of the 3D structural information used.

* Corresponding author e-mail: B.Chen@shef.ac.uk
[†] The University of Sheffield.
[‡] GlaxoSmithKline.
[§] Accelrys Ltd.

10.1021/ci050044x CCC: $30.25
Published on Web 06/22/2005

In general VHTS involves the following steps: (i) analysis and refinement of available 3D structures, with the intention of selecting potential binding sites; (ii) flexible docking of compounds into the selected sites, to computationally search through libraries of compounds for novel leads; and (iii) analysis of hits. This analysis leads to the purchase or synthesis of selected compounds and finally biological screening. VHTS, as currently implemented, falls into two general categories: receptor or ligand based pharmacophore mapping[4] and the flexible docking of ligands into protein binding sites.[5,6] The pharmacophore mapping requires availability of active or binding sets of compounds upon which the pharmacophore model can be trained.

Docking, on the other hand, does not require such a prerequisite and hence lends itself as a potential initial prospective screening methodology for novel targets. Docking is carried out in two stages: (i) the search for optimal ligand poses (conformation and orientation) and (ii) scoring of the ligand pose. The search involves an optimization of an object function (internal scoring function). The object function is usually based on a fast contact-based molecular mechanics partitioned model, used for ligand−protein interaction energy estimation, which can be evaluated rapidly during the docking cycle. Currently, docking is achieved by the application of the following optimization algorithms: genetic algorithms (GOLD,[7] AutoDock3.0,[8] Gambler[9]), evolutionary programming,[10] simulated annealing[11] (AutoDock), Monte Carlo[12] (MCDock), distance geometry,[13] fast shape matching (LigandFit[14], Dock[15]), and incremental construction (FlexX,[16] HammerHead[17]). These algorithms appear to be successful at predicting known protein bound ligand poses.[18] However, there are still some major challenges such as accounting for protein flexibility and the inclusion of solvent molecules during docking, both of which are generally overlooked by current methods.

Once a putative ligand pose has been proposed, a scoring function is used to estimate its relative affinity for the target. The ligand scoring is used to rank compounds, to separate binders from nonbinders in a docked list, rather than as a way of estimating absolute binding energies. Scoring functions are models which attempt to estimate the free energy of interaction between ligands and receptors.[19] These constructs hinge on empirical data such as measured binding constants (Empirical scoring), knowledge stored up in the ligand/protein X-ray complexes in the PDB (knowledge-based scoring), or the experimental values used as force field parameters (Molecular Mechanics-based scoring).[20] Scoring appears to be the weak point in VHTS. Some examples of fast scoring functions include the following: empirically based (PLP,[21] LUDI,[22] JAIN,[23] FlexX,[16] and LigScore2[14]), knowledge based (PMF,[24] DrugScore[25]), and force field based (GOLD,[7] DockScore[14]) functions.

In recent years there have been many comparative studies of docking and scoring methods.[26−29] These studies rely on seeding experiments to evaluate the efficacy of scoring functions, but this approach does have shortcomings as (1) the data used for such experiments is usually obtained from different sources and is therefore not coherent. (2) The random set of compounds selected to represent inactive sets, into which a few known actives are sprinkled, may contain unconfirmed actives. Presented here is an attempt to avoid these limitations by exploiting a set of known PDE4B inhibitors and noninhibitors (GlaxoSmithKline: high throughput screening IC50 data from a scintillation proximity assay (SPA)).

PDE4B is a member of the protein family called cyclic nucleotide phosphodiesterases (PDE), the basic role of which is to hydrolyze adenosine or guanosine 3′,5′-cyclic phosphate (cAMP or cGMP) to 5′-AMP and 5′-GMP. Some PDE4B inhibitors are believed to have potential clinical applications for dermatitis, rheumatoid arthritis, multiple sclerosis, autoimmune diseases, and various gastrointestinal and neurological diseases.[30,31] To our knowledge PDE4B has not been used previously to do retrospective or prospective docking studies in VHTS mode.

For this study, docking was separated from scoring by using a single docking engine, coupled to an internal scoring function (LigandFit/Dockscore) and postdocking scoring with a set of four scoring functions (PMF, PLP2, JAIN, and LigScore2, as implemented in the Accerlys Cerius2 module). The docking was performed using an in-house high-resolution crystal structure, derived from PDE4B cocrystallized with a pyrazolopyridine inhibitor. The proprietary nature of these data restricts any revelation of structural details. This, however, does not limit the ability to investigate and report the performance of the docking and scoring schemes.

2. METHODS

2.1. Computational Tools. Daylight toolkit[32] was available for database manipulation (e.g. SMIRK transformations), SYBYL tool case[33] (for visualization), Catalyst[34] for conformer generation, Cerius2 structure-based design module with Parallel LigandFit[34] (visualization, docking, and scoring), and PERL scripts for text and results processing. The Daylight Toolkit and named visualization software were run on a Silicon Graphics O2 machine. Parallel LigandFit was run on a 56-processor Linux cluster.

2.2. PDE4BB/Inihibitor Complex and the Set of Potential Inhibitors. 2.2.1. Data. A high-resolution (1.65 Å) X-ray structure of PDE4B cocrystallized to a pyrazolopyridine inhibitor was obtained (GSK: in-house data). The coordinates of the inhibitor were removed from the complex. Some water molecules as well as metal ions are retained for the docking (water molecules retained were identified from some comprehensive knowledge of the site obtained from several GSK in-house experimentation). Along with this a set of cocrystallized X-ray structures were used to assess the accuracy of the dockings.

The potential inhibitors, a data set of 5081 small molecules (single parent SMILES) with measured pIC50 (−logIC50) (the activity distribution in the data is shown in Figure 1), were generated during a therapeutic program to identify PDE4 inhibitors and as a consequence the compounds in the set has some interesting properties. Each was selected or synthesized as a result of structural similarity to known PDE4 inhibitors. Indeed half of the set belonged to four broad structural families (Figure 2).

The similarity profiles of these compounds were analyzed using the DAYLIGHT[33] Tanimoto coefficient to calculate average self-similarities within each class and interclass similarities (Table 1). The class self-similarities ranged from average Tanimoto values of 0.57−0.68 indicating reasonably high figures with class 5 yielding a very low value in comparison (0.31) as would be expected. The interclass values indicate lower average values 0.26−0.35.

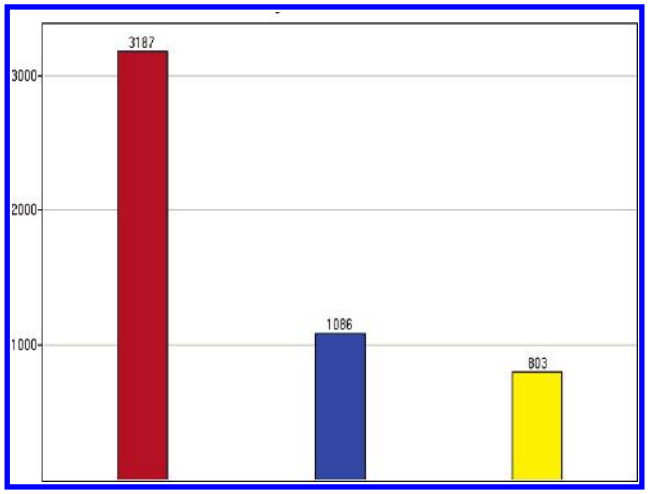

Figure 1. The distribution of activity across the data set (yellow: highly active > 7.0, red: < 6 inactive, blue: medium activity).

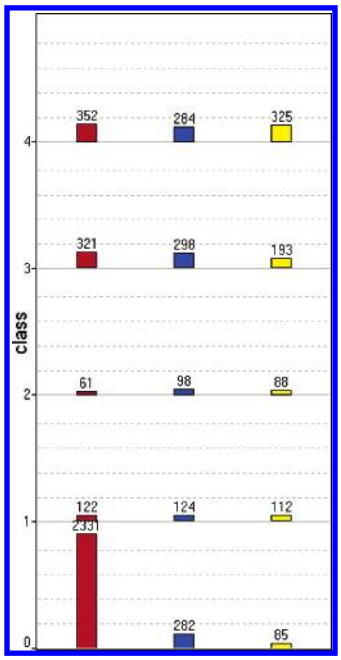

Figure 2. The activity distribution across four large chemical classes (classes: 1 to 4) and all other structurally diverse compounds (class 0) which make up the data set available for docking. The number on top of each column indicates the population of each class separated into highly active (yellow), medium active (blue), and inactives (red).

The nature of this set was believed to make the identification of actives a true test of docking and scoring efficiency. The fact that we are looking for actives among inactives of similar structural types can be likened to searching not so much for a needle in a haystack but for some particular pieces of hay in the haystack.

The data set was processed to produce a "dockable database" by attributing proper protonation states to ionizable groups (amines, amidines, carboxylic acids, and other acidic isosteric forms) at neutral pH, this was done by the use of simple Daylight SMIRK transformations, and finally the 3D geometries for each compound were generated using catConf, a conformational generator from Catalyst which can be used to generate one low-energy conformation for each molecule in the database.[34]

Table 1. Matrix of the Interclass and Intraclass Average Daylight Tanimoto Self-Similarity Values

	class 1	class 2	class 3	class 4	class 0
class 1	0.57				
class 2	0.26	0.61			
class 3	0.26	0.29	0.63		
class 4	0.27	0.33	0.36	0.68	
class 0	0.28	0.35	0.32	0.32	0.31

2.2.2. Preparation of Protein Coordinates and Identification of Hypothetical Binding Site. The high-resolution X-ray structure of PDE4B/pyrazolopyridine inhibitor complex was imported into Cerius2, and the ligand was extracted to leave a cavity.[34] LigandFit protein-based site detection algorithm was employed to identify potential binding sites. The largest site identified was selected and superimposed onto the extracted ligand coordinates; the two matched well. Another protein complex PDE4D2/rolipram (PDB entry 1OYN) is employed here to illustrate this procedure (Figure 3a,b). Our experience with LigandFit revealed that site grid regions identified in this manner must be kept as small as possible for successful results. Large regions, generated from large pockets or by manual extension of the grid, yield poor docking results.

2.3. Virtual Screening. LigandFit a modern docking engine was employed for this work;[14] this algorithm makes use of a cavity detection algorithm for detecting invaginations in the protein as potential active site regions. A shape comparison filter is combined with a Monte Carlo conformational search for generating ligand poses consistent with the active site shape. Candidate poses are minimized in the context of the active site using a grid-based method for evaluating protein−ligand interaction energies.

The docking was carried out with the following nondefault settings in LigandFit: site partitioning 2 in order to fully access the potential docking orientation of the active site, maximum trials variable table values to help the pseudorandom conformational analysis, and the CCF force field option used for the grid energy calculations. The flexible fitting option was selected for generation of alternative conformations on the fly, as was the diverse conformer's option to ensure the solutions generated cover a broad range of conformations with similar low-energy docking scores, and a maximum of 10 top scoring diverse ligand poses were returned for each of the 5081 compounds.[14]

2.4. Ligand Scoring. Docking was separated from general ligand scoring. The LigandFit algorithm uses an internal scoring function, DockScore, to select and return a maximum of 10 dissimilar poses for each compound. DockScore is a simple force field based scoring function which estimates the energy of interaction by summing the ligand/protein interaction energy and the internal energy of the ligand. CFF force field[14] was used to resolve the van der Waals parameters for DockScore. Although 10 poses were returned by LigandFit, the top DockScore pose was used for postdocking scoring. Scoring was performed using a set of scoring functions as implemented in Cerius2.[34] These included PLP2, PMF, JAIN, LigScore2, and DockScore available from the docking process. The putative 3D poses and score results were then stored as an SD file. Each docking was minimized, using DockScore, the only purely molecular mechanics based scoring function employed in this study, and this minimized

pose was then presented to each of the other scoring functions, which were either knowledge based or regression based.

2.5. Sprinkling Experiments (Seeding). In comparative docking and scoring studies it has become customary to evaluate the effectiveness of methods through the seeding of a small number of known active compounds into a large set of presumed inactives and subsequently determining the rate of recovery of actives. In this study, as the activities of the entire set was known, it would have been possible to simply process all the data and explore the recovery rate. However, rather than use that approach it was decided to explore the "robustness" and stability of each scoring method through repeated experiments in which subsets were created with randomly selected actives and inactives.

With this in mind, nine 2000 compound data sets were assembled from the pool of 5081 by random selection of 20 actives and 1980 inactive compounds For the purpose of this exercise active compounds were designated as pIC50 > 7 and inactive as pIC50 < 6. The compounds falling within the medium activity range pIC50 < 7 and > 6 were avoided to remove ambiguity. On average, each of the nine databases had a "random chance screening hit rate of one percent". These sprinkling experiments were set up to explore the ability of fast scoring functions to improve the hit rate of a screen through the successful ranking of compounds measured against a random hit rate of one percent. The repeated experiments, with randomly selected sets, were designed to provide an estimate of the robustness of the scoring function performances (averaged performance factors are reported for each function over these nine different data sets). It should be re-emphasized here that this study represents a true challenge of the effectiveness both of docking and scoring, as it requires identification of actives sprinkled among compounds selected originally because of their structural potential to be inhibitors of PDE4.

3. RESULTS AND DISCUSSION

The set of 5081 compounds was docked into the PDE4B pyrazolopyridine inhibitor binding pocket and several studies performed to study: (i) effectiveness for binding mode prediction; (ii) effectiveness of postdocking scoring as a means of identifying compounds of high activity; and (iii) relationship of binding mode prediction to success or failure in scoring. Consensus scoring was then explored, especially the impact of combining three functions from the different basic models or classes of fast scoring. Finally, the behavior of one combined scoring function across four different chemical classes is also analyzed.

3.1. Prediction of Binding Modes (Virtual Crystallography). The most straightforward method of evaluating the accuracy of a docking scheme is to inspect how closely "the lowest energy pose (conformation and binding mode)" predicted by the object scoring function, DockScore in our case, resembles an experimental binding mode as determined by X-ray crystallography. Ideally, a docking algorithm should return the correct binding modes for all ligands (virtual crystallography). The correct poses would then be used for scoring the molecules for a potency estimate (the virtual screening of virtually crystallized ligands).

In practice, incorrectly predicted binding modes will also be present, and they, along with the correctly predicted modes, will be scored, resulting in a more complicated outcome where enrichment is likely to be reduced.[35] Interestingly, it may be possible still to obtain enrichments but for less obvious reasons, such as differences in size and shape between actives and inactives.[36] The quality of binding modes was evaluated. Importantly, the same docking was used in this analysis as in the virtual screen itself, in order that a direct link could be established between the binding modes and the associated virtual screening outcome. A more thorough docking, customary when the primary objective is to obtain accurate binding mode prediction, was not embarked on.[37] Such studies are typically too resource intensive for virtual screening and are reserved for a smaller number of molecules.

Table 2. Top Scoring DockScore Predicted Binding Modes

ID	class	RMSD
Mol 15	4	1.04
Mol 13	4	1.28
Mol 7	4	1.53
Mol 9	4	1.63
Mol 10	4	1.64
Mol 1	1	1.71
Mol 5	2	1.72
Mol 3	2	2.16
Mol 14	4	2.19
Mol 17	1	3.08
Mol 8	4	3.14
Mol 2	1	4.85
Mol 12	3	5.26
Mol 19	0	5.46
Mol 16	1	5.95
Mol 18	1	6.06
Mol 6	1	6.08
Mol 4	1	7.10
Mol 11	3	7.15

The level of success in binding mode prediction with a particular docking algorithm also varies between targets and between ligand types for a given target. Below is an evaluation of the success in predicting the experimentally observed binding modes of a small subset of the compounds during the virtual screening exercise. The subset is composed of 19 compounds for which crystallographic binding modes were available. Success overall and within individual compound classes was investigated, and the latter analysis was used to check whether the virtual screening success for a given compound class is related to the quality of the binding mode predictions.

An RMSD analysis and visual comparison between the crystallographic and predicted binding modes was carried out for the LigandFit/DockScore selected pose for each of the 19 compounds with crystal structures—this corresponds to the use of only the top scoring pose for scoring in the virtual screen. The results are shown in Table 2. Visual inspection suggests that compounds with RMSD values of up to 2.19 correspond well to the empirical binding modes. This means a reasonably good overall result of nine out of 19 compounds or approximately half is achieved.

However, individual compound classes behave quite differently. The pyrazolopyridines 'class (4)' dock well with six out of the seven examples docking correctly. 'Class (2)' also performed well, though with a small sample set, with two compounds both docking within the RMS criteria.

Other classes perform less well: 'class (1)' performed rather badly with only one compound passing the RMS

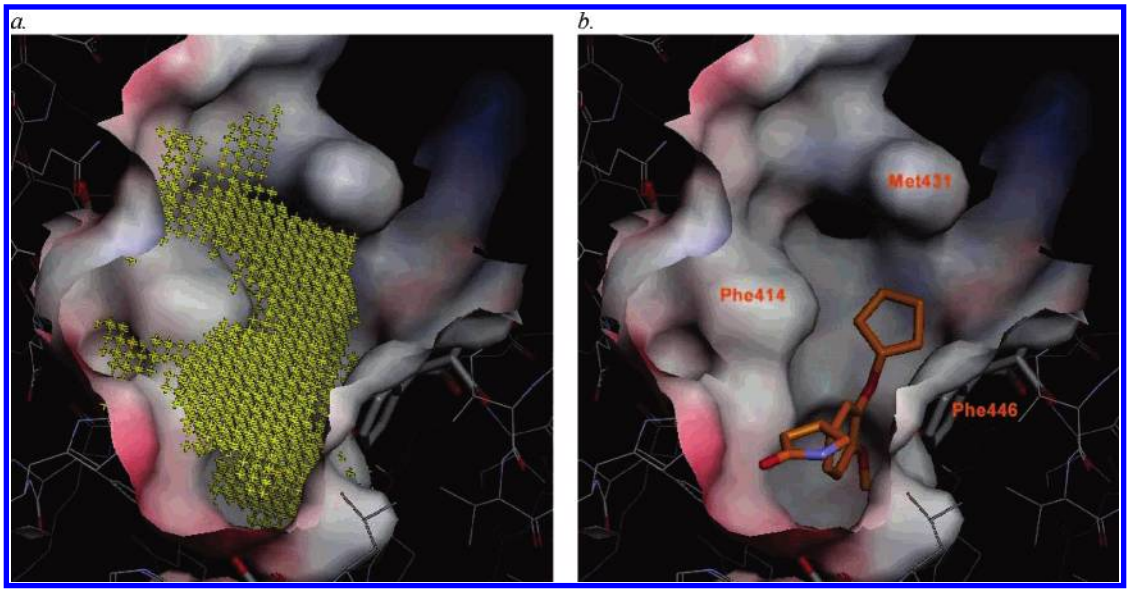

Figure 3. a. The site grid as detected by the LigandFit flood fill algorithm from protein coordinates. b. The cocrystallized X-ray ligand coordinates superimposed onto detected site viewed in Sybyl6.9 (PDB entry 1OYN) (site overlaid with ligand).

criteria and five compounds failing, while the two compounds in 'class (3)' both failed to dock within the RMS criteria. It is noteworthy that the crystal structure used in the docking was from a cocrystal structure with a pyrazolopyridine 'class (4)'. As will be discussed subsequently, pyrazolopyridines was also a compound class where particularly good enrichments were obtained. This indicates the limitation of relying on a single cocrystallized complex for rigid receptor docking.

Further work would be necessary for a deeper understanding of the reasons why compound classes behave differently in binding mode prediction. Certainly the differences in success for "virtual crystallography" between classes 1 and 4 cannot be explained with variation in simple properties such as MW (Figure 12), surface area, numbers of rotatable bonds, numbers of H-bonding groups (acceptors and donors); these properties were very similar for both classes. Variation in bound waters were considered but found to differ little in the binding pocket for the four classes. One explanation for failure of 'class (3)' is that the crystal structures for all compounds in the class invariably showed a cisoidal thiourea group; LigandFit was unable to generate dockings showing such a conformation. Experience with GOLD and GLIDE, using these same compounds, suggests this thiourea behavior may be a common failing of other docking algorithms.

Some possible reasons may include the subtle differences observed in the structure of the active site with different compounds bound and differences in the quality of the force field parameters for different molecule classes. Compound class dependency is harder to deal with for novel targets without prior crystallographic knowledge. However, it is useful to be aware of it as a potential issue for future virtual screening studies, as such behavior has been confirmed here with a well understood target.

Overall, the level of success in binding mode prediction is sufficient to indicate that the favorable virtual screening results reported subsequently have been obtained for the intended reasons, that is, molecules reasonably placed in the active site and scores reflecting binding affinity. In a subsequent study all 10 poses returned by LigandFit were investigated to see if any were close to the crystal structures, i.e., could the crystallographic pose be found at all? When RMSD values were calculated for all poses and the best value retained, 17 out of 19 compounds had a good predicted binding mode with RMSD less than 2.19, most pyrazolopyridines occurred within 1.5 Å from the crystallographic binding mode.

3.2. Assessment of Scoring Functions. Even though 'virtual crystallography' is included in this experiment, the principle objective of virtual screening is that of scoring molecules in a database in order to reduce the number of compounds that must be tested experimentally. As explained previously this work was undertaken to identify a method of VHTS which could provide a method to screen 500 compounds with a reasonable chance of success, where success is the identification of at least one lead. The efficiency of virtual screening can be expressed by taking into account the improvement gained by scoring compared with a random selection from the original database.

The investigation involved the analysis of results of docking and scoring over the nine randomly prepared data sets (sprinkling experiments) using two performance measures, the rate of recall (R) and the enrichment factor (E). The role of these measures is to facilitate the quantification of the relative differences between the original and the ranked or enhanced data set. There are many ways of defining the enrichment factors, but for this study simple definitions were chosen as described by Charifson et al.[38] and Pearlman and Charifson[39] for R and E, respectively.

The first factor, the rate of recall of hits (R), is the ratio of recovered actives or hits ($H_{sampled}$) to the total number active compounds in the database (H_{total}) (eq 1). R can be multiplied by 100 and plotted as a function of the screened percentage of the database, to yield a cumulative recall plot, for each scoring function (Figure 4). These cumulative recall plots 'R-plots' are useful for visualizing, quantifying, and comparing the behavior of molecules for each given scoring scheme at various fractions of the data screened.

$$R = H_{sampled}/H_{total} \quad (1)$$

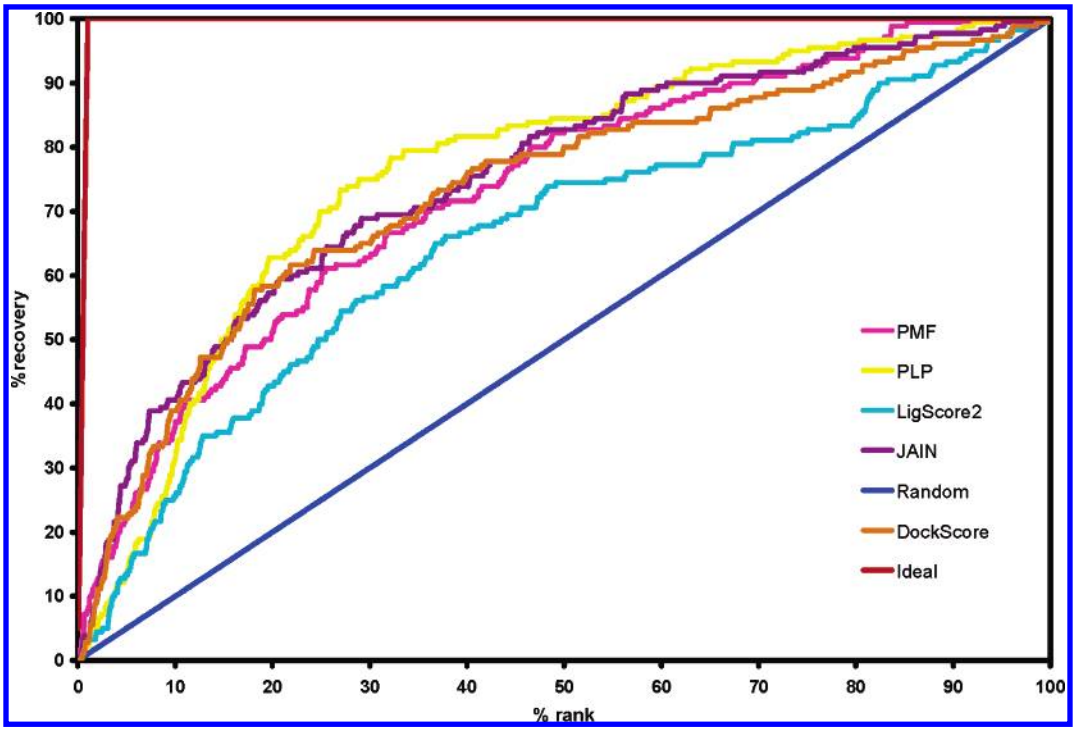

Figure 4. Rate of recall plotted as a function of database percentage for each scoring function. Performance of all the single scoring functions compared with random chance (blue line) and the ideal line (red line).

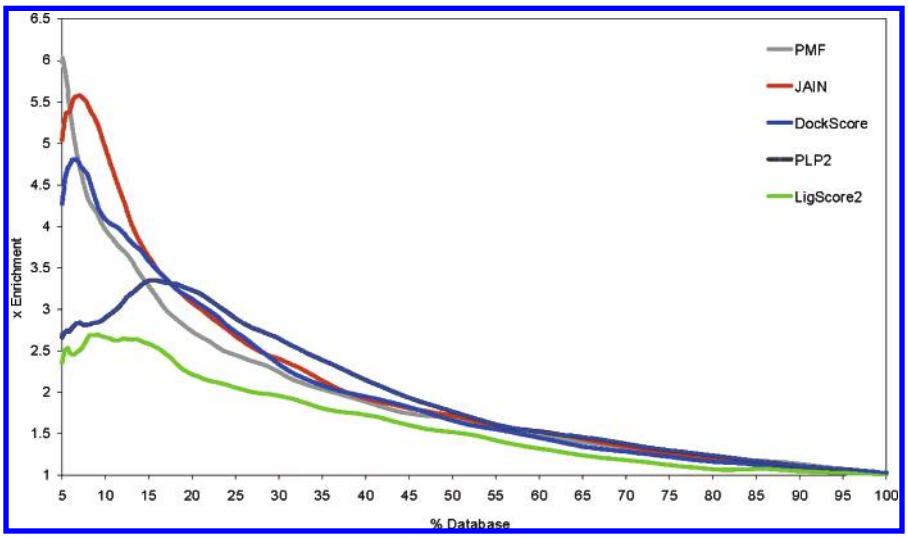

Figure 5. The enrichment factor (E-plots) plotted as function of the fraction of database sampled for the single scoring functions. A moving average is applied with a 5% interval to smoothen the plots.

The second factor chosen is the enrichment (E) (eq 2) where H is the active compounds, N is the number of compounds, and R is the recovery as described in eq 1. This factor describes how much richer inactives the selected fraction is than the initial database. A closer examination of this equation shows that a random data set must yield an enrichment factor of 1. If E is plotted against the percentage of database screened this yields enrichment plots (Figure 5).

$$E = \frac{H_{\text{sampled}}/H_{\text{sampled}}/H_{\text{total}}/H_{\text{total}}}{= \dfrac{R^{*}N_{\text{total}}}{N_{\text{sampled}}}} \qquad (2)$$

Using these relatively simple numerical tools (eqs 1 and 2) one is able to analyze and compare performance of the scoring schemes. The plots, cumulative recall plot (R-plot) and enrichment plot (E-plot), can complement each other to provide visual aids for the analysis. As mentioned, consideration is also given to the underlying variability in the success rates of our simple 'virtual crystallography' experiment and relating it to scoring behavior.

3.2.1. The Rate of Recovery. The results reported here are averages of the enrichment factors of scoring schemes in over nine 2000 compound databases. The average performance figures are believed to give more reliable results (Figure 4). This method also provides a variance (standard deviation) figure which serves to indicate the consistency of each scoring method at recovering structurally different active molecules, in other words the scoring scheme robustness (Table 3).

Table 3. Rate of Recall for Single Scoring Functions

scoring function	5% database	10% database
PLP2	13.9 ± 5.5	31.6 ± 10.0
PMF	22.2 ± 7.5	36.6 ± 8.6
LigScore2	13.3 ± 7.1	25.5 ± 9.8
JAIN	28.3 ± 6.6	40.5 ± 8.1
DockScore	22.2 ± 8.7	38.9 ± 12.6

At a glance the plots show that the rate of recovery (*R*) can be affected by the choice of scoring function. When studying the cumulative recall plot, consider the random and the ideal lines. The random line represents the likelihood of retrieving active compounds at random from the original database before ranking. The ideal line indicates the best possible rate of recall. The goal is to have a scoring scheme which drives the cumulative recall plot away from the random line (blue) toward the optimal (red) (Figure 4).

The top 5−10% fraction is of utmost importance to this exercise because the intention is to use virtual screening as a means of selecting a small number of compounds for experimental screening. At approximately 5% of the ranked database the best single scoring functions (JAIN, PMF, and DockScore) will recover between approximately 22−28% of the active compounds and the rest (LigScore2, PLP2) below 14% (Table 3). The standard deviation (SD) indicates the variation in performance of each of the scoring schemes across the nine different compound sets. A large error (SD) indicates that scoring is affected acutely by the type of compounds making up each database. These results were confirmed by enrichment factor analysis.

3.2.2. Enrichment as Function of Ranked Database Fraction. Enrichment plots for each of the scoring functions (Figure 5) provide a rapid means of analyzing their performance and complement the information gleaned from the cumulative recall plots. To remove the noise and obtain a smoother and more meaningful analysis of the enrichment, the enrichment plots were constructed by taking the mean (a running average) for every 5% interval along the *x*-axis. The (E) plots instantly demonstrate that selecting the top 5% of compounds from the single scoring function ranked database leads to subsets with enrichments of up 2.4 times for LigScore2 to a maximum of 6 for PMF. E-plots for all the scoring functions show that there is an initial increase in the enrichment factor (*E*-value) to a maximum and then decreases as the number of actives falls in the remainder of the database. The *E*-value gradually approaches 1, the random enrichment of the database. The shape of the plot for each scoring function also provides insight into the suitability of each scoring method for virtual screening. It is preferable to select a scoring function which not only attains a high *E*-value but also climbs to a peak as early as possible especially in the top 5−10%. Notice that some single scoring functions have broad shaped E-plots which peak later e.g. LigScore2 and PLP2.

In summary, the variable performances of the single scoring functions (Figure 4) demonstrate some problems encountered by scoring methods. In the randomized sprinkling experiments some data sets have high rate of recalls compared to others as indicated by the large standard deviations for each scoring function (Table 3). Some of the implications of this will be revisited during the analysis of the behavior of scoring across different compound classes.

Table 4. Rate of Recovery for Consensus Scoring Functions

scoring function	5% database	10% database
cScore-PPmJ	35.0 ± 9.7	47.2 ± 9.1
cScore-LPPm	32.2 ± 8.7	42.8 ± 6.7
cScore-LPJ	29.9 ± 8.9	45.0 ± 7.7
cScore_DPmJ	37.2 ± 8.3	51.1 ± 9.0

However, here the Student *t*-test was used to compare pairs of scoring functions in order to investigate if the apparent differences observed for single scoring functions are statistically significant. The 5% point in the database was used for the *t*-test (Table 4). Each 5% point on the cumulative recall plot (*R*-value) is an average over the nine databases. For simplicity, DockScore is compared to all other single scoring functions (for example *c*-values greater than 1 for LigScore2 and PLP2 indicate statistical differences as opposed to JAIN and PMF which are not significantly different to DockScore); see scores in Table 4.

Importantly, for novel targets, an experimenter will not always have at their disposal such a comprehensive test set. Since scoring function efficacy depends on both the protein target and the type of compounds in the database the selection of a specific scoring function for any prospective docking study can become impossible. Rational combination of available individual scoring methods (consensus scoring) may be a way to avoid this problem. The combined scoring functions (cScore- DPmJ and PPmJ) perform in a superior fashion to the single scoring function and by their nature the combination of functions will ameliorate the effect of any particularly unsuitable single function.

3.3. Consensus Scoring Schemes. As reported elsewhere[38,39] data fusion, data triangulation, combined scoring, or consensus scoring is useful as a means of creating more robust functions. These new functions may not be as significantly affected as singular scoring by varying the chemotypes to be scored. We also attempted to investigate the effect of assembling triangular scoring schemes that are constructed based on a rational approach to data fusion. The rational approach is based on combining scoring functions that are as different as possible in the nature of their basic affinity estimation models i.e., coming from the three different classes of scoring functions. The rationale being that, if the component scoring functions of a consensus scheme are dissimilar, then they will capture different aspects of a predicted binding mode. To illustrate this, a correlation of DockScore with each scoring function is performed (Figure 6). These scatter plots merely serve to indicate the differences between the single scoring methods and do not say anything about the performance of each scheme. It appears that correlation (*R*-value) improves if one compares similar scoring methods (e.g. LigScore2 and DockScore both rely on the CFF force field) but for different models (e.g. PMF and DockScore represent scoring methods from different classes) there is a distinct worsening in the *R*-value (Figure 6).

To simplify consensus scoring, a simple rank based score was designed. More sophisticated methods have been reported, including vote based ranking and many others.[38,39] Rank based consensus scores are calculated as follows: for each function used in the consensus score a ranking (*r*) of the database was attained and *r* was subtracted from the total number of compounds in the database (*d*) to give a rank

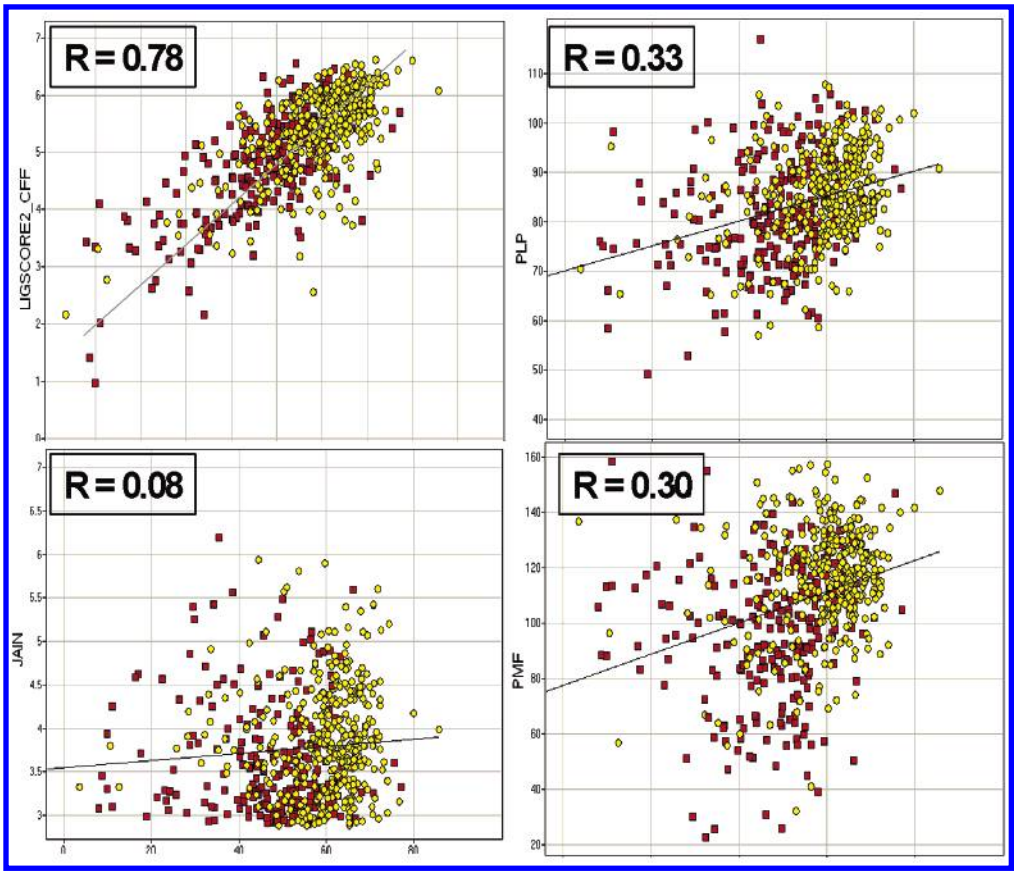

Figure 6. Scatter plots to illustrate the correlation between DockScore on the *x*-axis with each single scoring function. The calculated *R* values for the straight line fit are 0.78, 0.33, 0.30, and 0.08 for LigScore2_CFF, PLP2, PMF, and JAIN, respectively.

based single-function score ($d-r$) for each compound. This score was then averaged over all combined functions to give the combined score (i.e. divided by $n = 3$ for a our triangular scoring schemes) (eq 3).[39]

$$R_{cs} = \sum_{i=1}^{n} \frac{(d-r)^i}{n} \tag{3}$$

Another major difference between this method and the literature reported methods is that, here, scoring is done on a single binding mode. While it may appear desirable to score all poses returned by the docking engine and use the best score for a given molecule/scoring function in a consensus score, it was found that such an approach failed to alter improve the enrichments obtained [Daniel Ormsby, unpublished data].

Thus for the purposes of this work, rational consensus scoring involved combining a knowledge based function (e.g. PMF), an empirical scheme (e.g. JAIN), and a force field based scoring scheme (e.g. DockScore). Semirational schemes involve two related scoring schemes e.g. two empirical (from JAIN, PLP2, and LigScore2) with a statistical scheme (e.g. PMF). Finally nonrational schemes from the same class are compared. Following these guiding principles four new rank-based combined scoring schemes were constructed; a rational (DPmJ), two semirational (LPPm, PPmJ), and a nonrational scheme (LPJ). Please note that L represents LigScore2, P (PLP2), D (DockScore), Pm (PMF0), and J (JAIN).

The results show an overall increased performance for the combined scoring functions over the singular scores (Figure 7). This is indicated by the shift away from the random toward the ideal line for all combined scoring schemes. Further qualitative analysis to support this observation can be obtained by contrasting the recall plot of DockScore with those of each combined score (Figure 7).

Quantitative results are shown in Table 4. The combined score (cScore)-DPmJ shows the greatest rate of recall (37%) and the poorest is (cScore)-LPJ (at 30%) in the top 5% of the database. Compare this with Table 2 for single scoring. An encouraging result is the SD for these data points. The SD does not increase relative to the increasing recall from 5 to 10% for each scoring function.

'Rational construction' of scoring functions appears to provide a successful means of combining the scoring functions. This is supported by the observation that (cScore)-DPmJ the only truly heterogeneous 'rational' consensus scoring scheme shows a marginal improvement over the other three functions. This may require a further study to analyze if this is a general trend.

However it is also encouraging to see that each new combined scoring scheme appears to outperform the best of the single scoring function. Again the *t*-test is used to investigate if the differences seen are significant (Table 5).

As with the single scoring functions, the E-plots for consensus scoring schemes are used to further validate the rates of recovery. Again caution must be taken when reading the enrichment plots because each data point plotted is an average over a 5% interval and direct comparison with the cumulative recall plots may be miss leading. The E-plots reveal interesting results. The new schemes all peak in the top 5−10% of the database, and another positive indicator is the shapes of the E-plots which show no broadening or

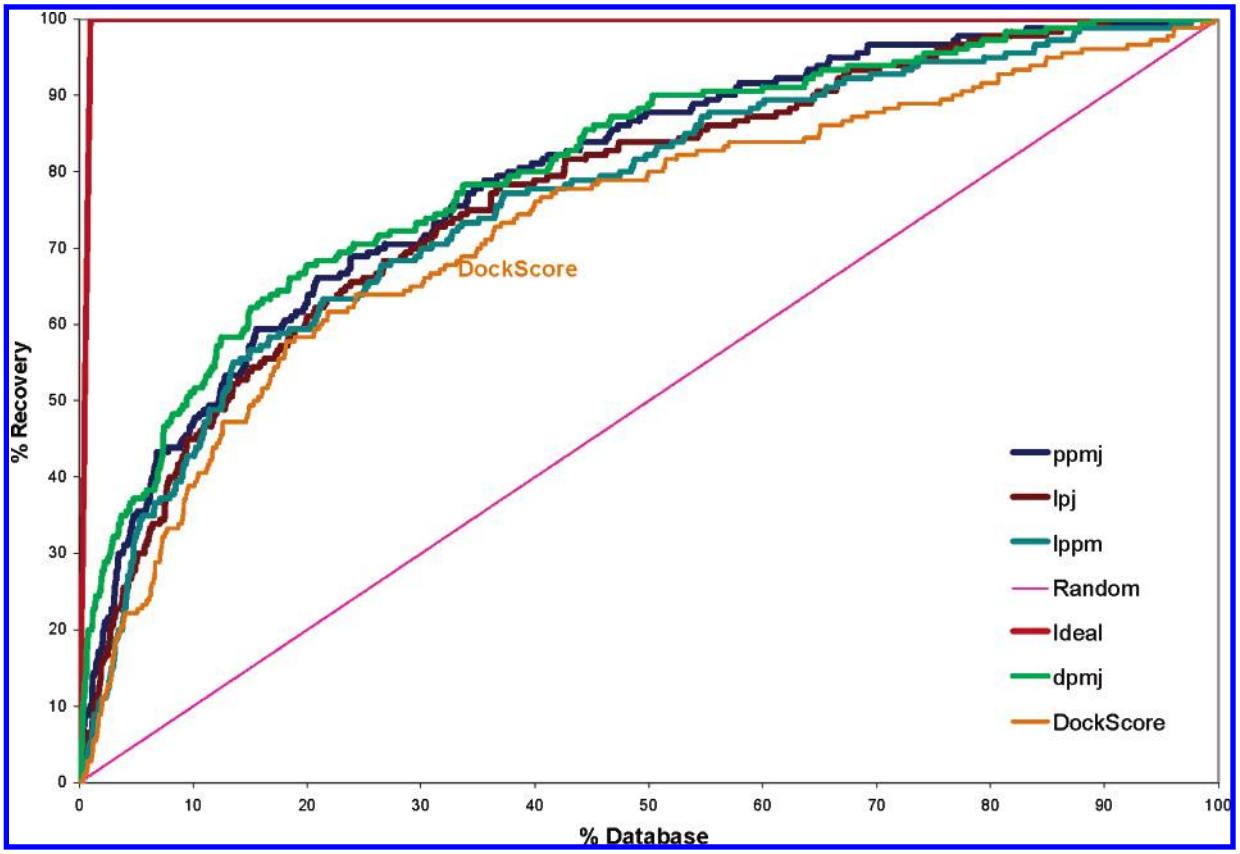

Figure 7. Rate of recall plotted as a function of database percentage for each combined scoring function compared to DockScore.

Table 5. Significance Test (*t*-Tests) Used in This Study[a]

scoring scheme	*t*-value	*c*-value	significant difference
sScores			
DockScore v PLP2	2.43	1.39	yes
DockScore v LigScore2	2.13	1.22	yes
DockScore v JAIN	1.68	0.96	no
DockScore v PMF	0.00	0.00	no
cScores v sScores			
DPmJ v PLP2	7.02	4.02	yes
DPmJ v LigScore2	5.84	3.34	yes
DPmJ v PMF	4.00	2.29	yes
DPmJ v DockScore	3.74	2.14	yes
DPMJ v JAIN	2.51	1.44	yes
cScores			
DPmJ v LPJ	2.05	1.17	yes
DPmJ v LPPm	1.25	0.71	no
DPmJ v PPmJ	0.52	0.30	no

[a] The *t*-test (a one-tail, two-sample assuming unequal variances test) was used to determine the significant differences in the average recall (*R*) for the 5% point in the nine ranked databases between pairs of scoring functions. The *c*-value is the ratio of *t*-value to the *t*-critical value a *c*-value > 1 indicates a significant difference. The one-tail *t*-critical value of 1.746 is read from *t*-tables for 16 degrees of freedom at 95% confidence level. The number of degrees of freedom is calculated by adding the total number of data points in the two populations being compared. sScore:- single scoring functions. cScore:- consensus scoring functions.

spreading as do some singular scoring functions such as PLP2 (Figure 8).

The choice of a correct scoring function for ranking hits obtained from any virtual screening experiment represents a real challenge given the ever increasing numbers of possible scoring functions. The results suggest that an improved and more robust performance can be obtained by consensus scoring (Table 4 and Figure 7). The results demonstrate statistically significant differences in the rate of recall between the best cScore function compared to each of the single scoring functions. The *t*-test (Table 5, cScores v sScores) shows that the DPmJ-sScore pairs have significant differences at the 95% confidence level.

Rational based consensus scoring, above all, yields better results when compared to nonrational combinations. DPmJ, the most rational scheme, appears to outperform all other combinations (Figure 8). Again the *t*-test can be used to analyze the statistical significance of the observed differences (cScores in Table 5). A statistical difference is observed between DPmJ and the nonrational combination LPJ. It is important to appreciate that a more comprehensive study of this nature should include more scoring functions and combinations.

Finally to answer the key question about whether the methods proposed enrich the database satisfactorily (i.e. a targeted enrichment factor of 4), 3 out of 5 sScoring functions (PMF, Jain, and DockScore) achieved this requirement. All consensus scoring schemes used in this paper exceed the set enrichment target with enrichments of 14 for DPmJ, 10 for PPmJ, and 7 for LPJ. Even the poorest performing combination (LPPm, *E*-value 5) achieves this requirement. A real benefit of the combination of scoring schemes is the robustness of the resulting scoring functions; these functions appear more generalizable over the nine databases, designed by the sprinkling or seeding procedure.

3.4. The Behavior of Scoring across Different Compound Structural Classes. One of the more important aspects of docking/scoring in virtual screening is that the technique can be used to identify compounds with diverse structures, a distinct advantage over ligand-based virtual

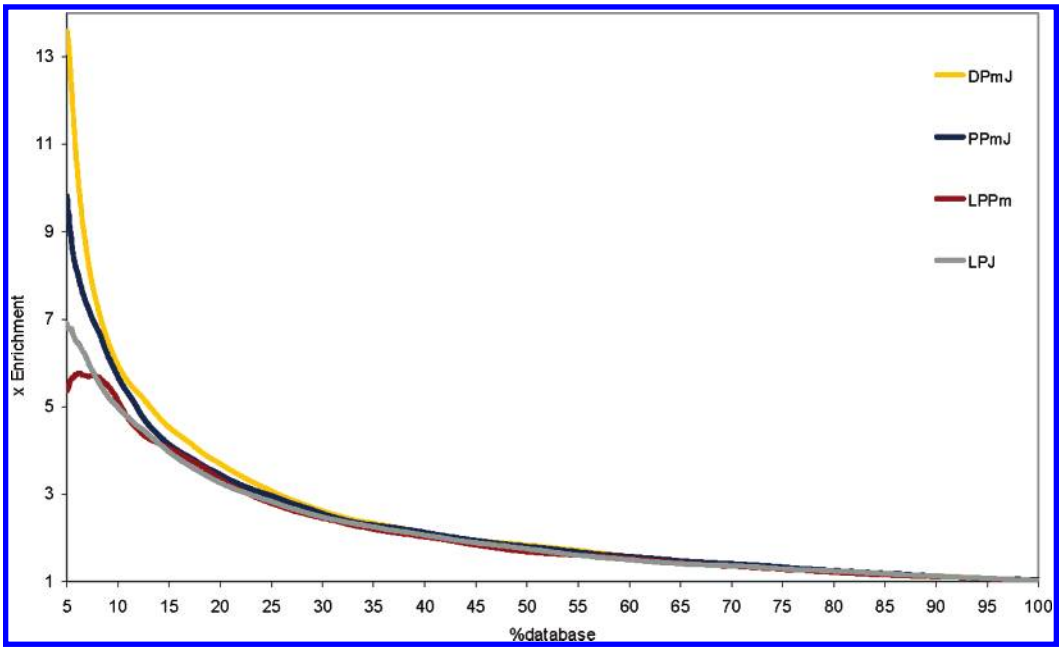

Figure 8. The enrichment plots (E-plots) for the triangular rank based consensus scoring schemes. A moving average is applied with a 5% interval to smooth the plots.

screening techniques or similarity searching. Therefore, the approach has the potential to provide new lead series for known targets. This property prompted an investigation into the way scoring behaves over different compound structural types.

As previously discussed, binding modes for certain compounds types are better predicted than others. This may be due to a number of reasons: inherent failure of the docking for that chemo-type; the difficulty of predicting binding modes incorporating water bridges; the failure of the rigid receptor model, due to the predisposition of the rigid receptor toward certain binding modes as encoded by the initial cocrystallizations; and many other related problems discussed elsewhere.[35,36] Suffice to say, one must be aware of these problems when utilizing docking as a virtual screening tool.

A robust, highly effective 'triangular' consensus scoring scheme, PPmJ, was used not only to compare the scoring of chemo types but also to evaluate the potential utility of scoring as a guide to lead optimization of a series. As mentioned DPmJ appears to perform marginally better than all other combined schemes; however, for the rest of this study PPmJ was preferred because DockScore is a specific internal LigandFit scoring scheme and may not be available in all laboratories. It is hoped that this work may be helpful for general docking studies.

The results and performance over the whole data set of 5081 compounds for PPmJ are shown by the histograms in Figure 9. Here the scoring function value range is divided into regular intervals, or bins, into which all the retrieved compounds are accumulated. The three histograms show the compounds separated into their respective activity groups (highly active (-yellow), medium activity (-blue), and inactive compounds (-red)).

Qualitative analysis of histograms in Figures 9−11t: examination of these histograms is highly informative of the success or failure of the scoring for the various structural classes explored. The scoring function values are partitioned, or binned, into 10 bins at regular intervals. For a successful scoring regime it should be seen that within each data set, or compound class, the number of active molecules increases (yellow) relative to the inactive compounds (red) as the scoring function value increases (i.e. enrichment of active compounds).

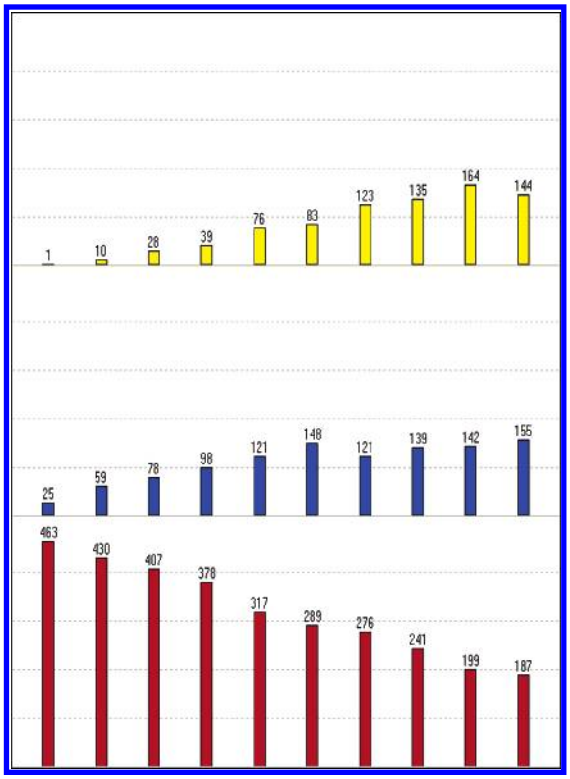

Figure 9. The performance of (cScore)-PPmJ a combined scoring function over all compounds in the database (*x*-axis represents a binning at regular intervals of the score; *y*-axis represents the number of compounds scored in a particular bin) (interval). (Yellow: high activity compounds pIC50 > 7, blue: medium activity compounds, red: inactive compounds pIC50 < 6). The number on top of each column indicates the number of compounds, active or not, recovered.

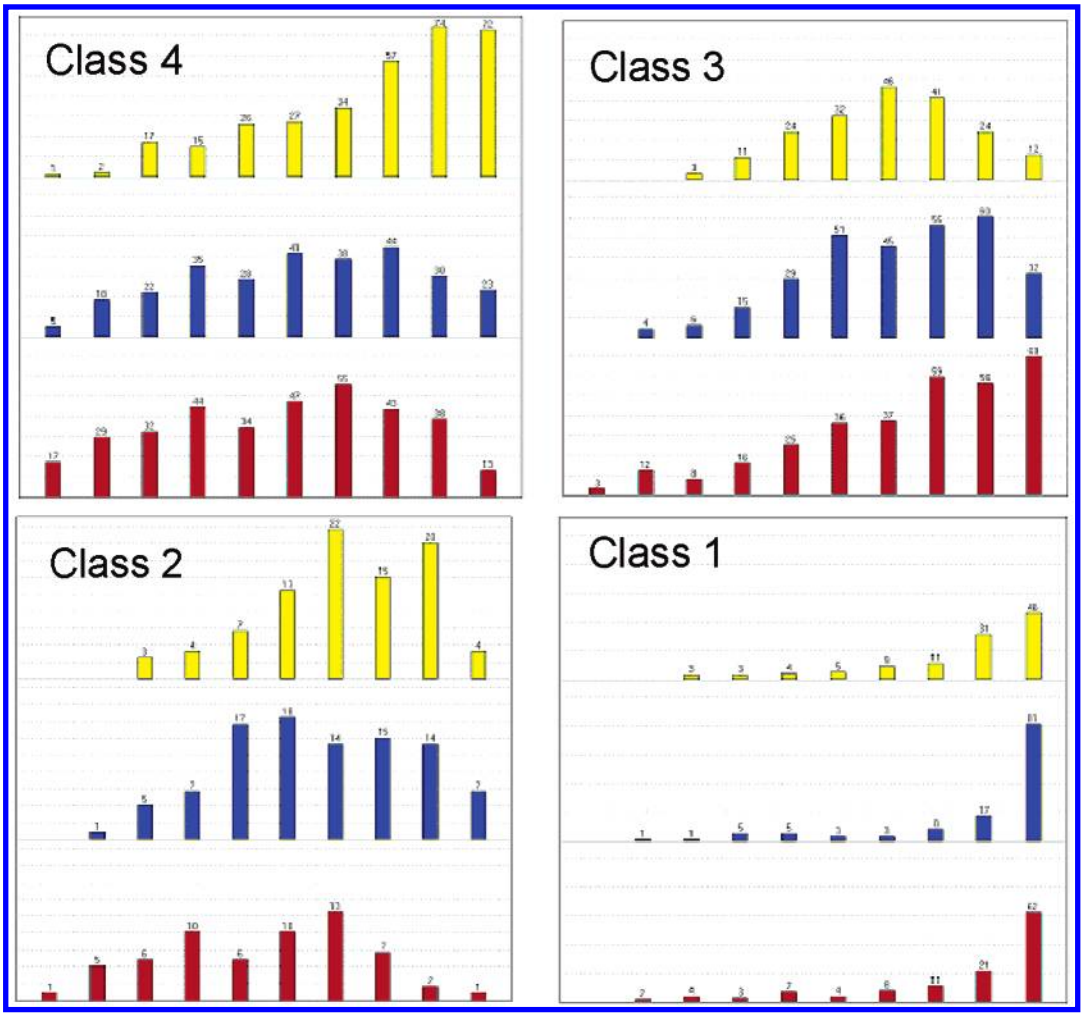

Figure 10. Performance of the consensus score across the four identifiable chemical classes [classes 1, 2, 3, and 4].

To assist with analysis of the results it is necessary to observe the population number in each bin; this is marked on top of each the column. If the increase is accompanied by a decrease or no change in the inactive compounds (red), then there is discernible enrichment. To avoid bias it is essential to consider the population profile of each analysis, see Figures 10and 2. This is especially vital when the analysis is aimed toward smaller classes of a data set which may contain limited numbers of inactive compounds.

The (cScore)-PPmJ performs relatively well over the complete data data set (Figure 9); this is also reflected in the high enrichments and rate of recalls seen in the virtual screening section for this particular scoring scheme. The scoring method is less successful when presented with compounds of medium activity. Generally the usual sprinkle and dock exercises are poor at identifying which active compounds are failing consistently (false negatives). Even when this information is made available, the virtual screen is not usually accompanied by an examination of the virtual crystallography, i.e., identifying which compounds classes are failing to score predicatively and rationalization of such problems with the initial docking 'virtual crystallography'.

Qualitative analysis for each of the four chemical major classes followed the same cue described above. Qualitative enrichment trends can be traced and related to the results from the 'virtual crystallography' experiment.

Class 4. There appears to be a general increase in the numbers of actives from left to right, while the inactives do not increase, i.e., an enrichment trend is seen (Figure 11). This is can be related to the success rate for the binding mode prediction for this class of compounds i.e., 6 out of 7 poses returned by LigandFit/DockScore prediction method were correctly placed within the binding pocket (RMSD values less than 2.5).

It is important to note that the initial experimental cocrystallized PDE4B high-resolution X-ray structure was bound to a class (4) compound. This may mean that docking and scoring is biased toward this class of compounds. Comparison with the success of other compound classes is necessary before reaching any conclusions. The result for this class of compounds suggests that the enrichment observed is for the right reasons i.e., the correct binding mode is scored successfully. In addition, the success of the PPmJ scoring with this chemical class suggests that it could provide useful assistance to a lead optimization project attempting to select compounds for synthesis.

Class 3. Figure 11 for class 3 fails to reveal a useful trend. There is no significant separation of actives from inactives, and indeed the number of inactives actually increases with higher scores. This is also reflected by binding mode prediction results, the two active compounds analyzed are not well predicted i.e., RMSD values for these are greater than 4.00. It should be mentioned that 'class (3)' adopted a

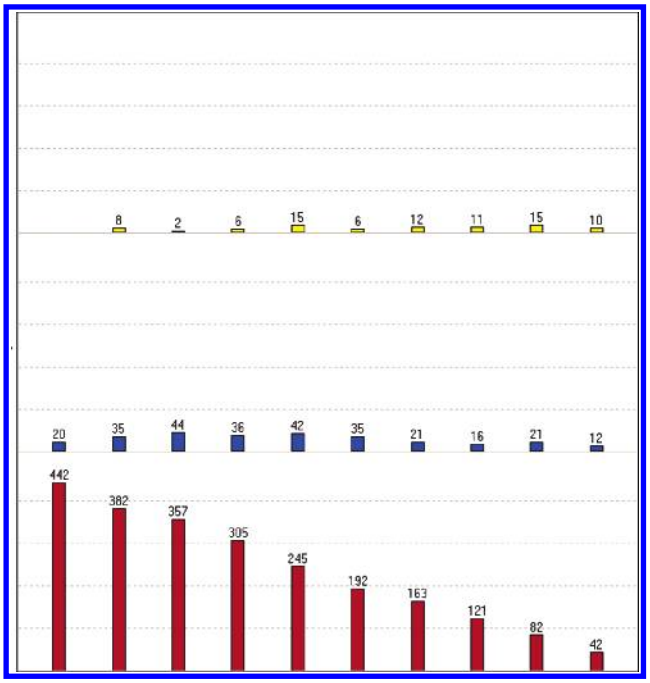

Figure 11. Performance of (cScore)-PPmJ a combined scoring function over class 0, all compounds in the database that do not form part of the four classes (*x*-axis represents a binning at regular intervals of the score, *y*-axis represents the number of compounds scored in a particular bin (interval).

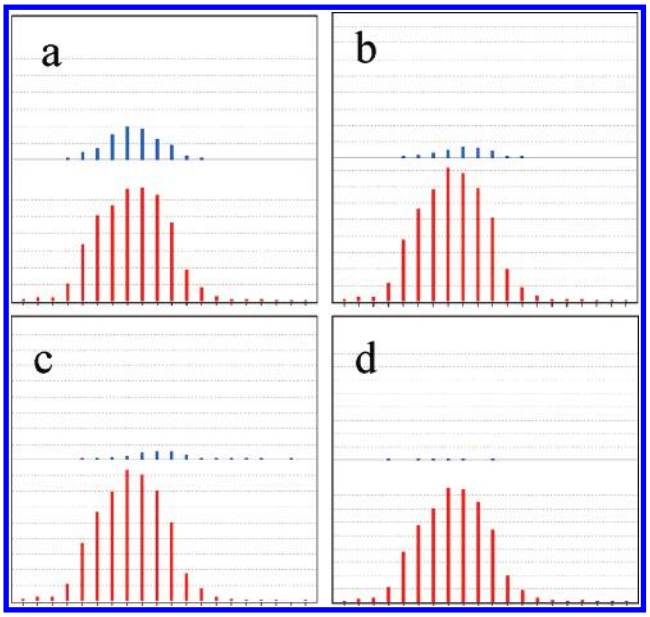

Figure 12. Molecular weight (MW) profiles of the compound classes (a) class 4 (blue) compared to the profile of the remaining molecules in the whole PDE4b (red) data set, (b) class 3, (c) class 1, and (d) class 2. The *x*-axis represents 20 evenly distributed MW intervals and *y*-axis (MW range from 139.12 to 809.97) represents the population of compounds in each interval.

particularly surprising cisoidal thiourea conformation in the crystal structure. Interestingly, it was found that neither LigandFit nor any of the other commonly used docking approaches were able to find this unusual structure as the most favored docking pose. This failure seems to be translated into the poor scoring results for all of the molecules in this class. Again it is good to see that enrichment can be correlated with binding prediction. This class would be a very bad choice for further lead optimization using the cScore for affinity prediction.

Class 2. This class shows a general increase in the number of actives as the score increases (Figure 10), but the inactive molecules seem to be randomly placed into the score bins. Again referring to virtual crystallography experiments reveals good prediction of this class, and two of the compounds available for binding mode prediction show good RMSD values and are positioned well in the binding pocket. Class 2 would be suitable for further optimization through use of the cScore for affinity prediction.

Class 1. This class is interesting. There appears to be a good trend in the actives and even the medium active compounds. However, the inactive compounds are not discriminated against (Figure 10). Consideration of the binding modes shows that the predictions yielded a 1 out of 6 success rate in the virtual crystallography exercise. This implies that the apparent enrichments observed for the active compounds are not related to correct binding mode prediction but rather a result of a general high scoring for this compound class.

Again this calls into question the strategy of using a single cocrystallized complex as the starting point in rigid receptor docking. It may be beneficial to investigate the impact of using more that one structure, where resources are available. This class is a poor choice for future cScore guided lead optimizations in this case but this may not be so if one had started with the class 1 binding pocket.

Class 0. It is important to analyze the performance of scoring over the rest of the compounds which do not form part of the four main classes. The histograms in Figure 11 reveal that the cScore scores the inactive compounds poorly, while the active compounds are randomly placed across the score bins. The combined effect of these two trends is an overall enrichment of actives in the top scoring bins.

It would appear that scoring functions can be used as means of identifying compounds series for lead optimization. This proposal needs further investigation, but based on these results consensus scoring will not only provide a means of enriching a small data set but also a means of proposing structural classes for further optimization. This is enabled by combining a robust scoring scheme for retrospective and prospective analysis of docking results with knowledge of the success of the docking mode prediction. The more successful the prediction of docking mode, the more likely the scoring to be predictive of affinity.

4. CONCLUSIONS

The efficiency of docking program (LigandFit) and the behavior of the selected scoring functions were thoroughly investigated using a high quality data set of PDE4B inhibitors where IC_{50} for each active and inactive compound was uniformly determined. This is unique and different from most of studies of similar kind where randomly sprinkled compounds were simply assumed to be inactives. The possible false positives and negatives are hence reduced in our study. In addition all of the compounds in the original experimental high throughput screening data set were designed to be similar to known inhibitors. This provides a stringent test of the capacity and robustness of various scoring scheme to discriminate between actives and inactives. The consensus

scoring method we proposed has been particularly successful in this regard.

Docking and scoring can provide a means to improve the hit rate, in the top fractions of the scored database, by at least a 4-fold enrichment factor. This is the minimum enrichment prescribed for the exercise of laboratory screening 500 compounds, in the pursuit of at least one lead. The LigandFit and DockScore procedure was able to return the correct experimental binding modes with reasonable accuracy, even using virtual screening parameters for rapid virtual crystallography.

Furthermore, analysis of the returned ensemble (10 putative binding modes) provided nearly all the crystal structure poses. However, for practical purposes, only the top DockScoring binding mode was used as it is difficult to see how, without prior knowledge of the crystal structures, one would extract the correct solution from this ensemble of potential solutions.

Of the single scoring functions analyses, the optimal, for PDE4, were found to be PMF and JAIN. Combinations of scoring functions were found to provide added value, with the rational scoring function of DPmJ and the semirational function PPmJ identified as methods of choice. Combining scoring functions may not only provide us with a straightforward means of selecting scoring methods for novel targets but potentially provides a more robust scoring approach less affected by possible aberrant behavior of a single function with a particular target.

The results for the "class 0" compounds (no class) clearly demonstrate that docking and scoring owes much of its effectiveness not only to identifying the best compounds but also through elimination of the bad. For the class 0 there was little or no preference for potent compounds to be identified as high scoring, but the poor compounds scored extremely poorly.

Docking and Scoring may be used for activity prediction within a chemical class during lead optimization. The most reliable results are obtained when the protein used is derived from a crystal structure for a member of that chemical class.

Finally, the authors recommend the following docking/scoring combinations for VHTS: LigandFit/DockScore for multiple docking into several versions of the protein active site (several cocrystallized ligand-protein binding sites if available) followed by ranking the highest DockScore retrieved poses using rational consensus scoring (DPmJ) or semirational combinations of available scoring methods.

ACKNOWLEDGMENT

The authors gratefully acknowledge Dr. Mike Hahn and his department at GSK for providing the computational and other facilities and Prof. Peter Willett, Dr. Val Gillet, and Dr. Roger Mutter (Sheffield University) for their advice and proof reading. This work is funded by The Department of Health (UK, DH0071/0102).

REFERENCES AND NOTES

(1) Kuntz, I. D. Structure-Based Strategies For Drug Design And Discovery. *Science* **1992**, *257*, 1078−1082.
(2) Diller, D. J.; Merz, K. M., Jr. High-Throughput Docking For Library Design And Library Prioritization. *Proteins* **2001**, *43*, 113−124.
(3) Joseph-McCarthy, D. Computational Approaches To Structure-Based Ligand Design. *Pharmacol. Therapeut.* **1999**, *84*, 179−191.
(4) Gund, P. In *Pharmacophore Perception, Development, and Drug Design*; Gunner, O. F., Ed.; International University Line: California, U.S.A., 2000; Chapter 1, pp 3−11.
(5) Ewing, T. J. A.; Kuntz, I. D. Critical Evaluation Of Search Algorithms For Automated Molecular Docking And Database Screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189.
(6) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles Of Docking: An Overview Of Search Algorithms And A Guide To Scoring Functions. *Proteins* **2002**, *47*, 409−443.
(7) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development And Validation Of A Genetic Algorithm For Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(8) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using A Lamarckian Genetic Algorithm And Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(9) McMartin, C.; Bohacek, R. S. QXP: Powerful, Rapid Computer Algorithms For Structure-Based Drug Design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.
(10) Budin, N.; Majeux, N.; Caflisch, A. Fragment-Based Flexible Ligand Docking By Evolutionary Optimization. *Biol. Chem.* **2001**, *382*, 1365−1372.
(11) Goodsell, D. S.; Olson, A. J. Automated Docking Of Substrates To Proteins By Simulated Annealing. *Proteins* **1990**, *8*, 195−202.
(12) Liu, M.; Wang, S. MCDOCK: A Monte Carlo Simulation Approach To The Molecular Docking Problem. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 435−451.
(13) Leach, A. R.; Smellie, A. S. A. combined model-building and distance-geometry approach to automated conformational analysis and search. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 379−385.
(14) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A Novel Method For The Shape-Directed Rapid Docking Of Ligands To Protein Active Sites. *J. Mol. Graph. Model.* **2002**, *21*, 289−307.
(15) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach To Macromolecule Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(16) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe. G. A. Fast Flexible Docking Method Using An Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(17) Welch, W.; Rupert, J.; JAIN, A. N. Hammerhead: Fast, Full Automated Docking Of Flexible Ligands To Protein Binding Sites. *J. Chem. Biol.* **1996**, *3*, 449−489.
(18) Ewing, T. J. A.; Kuntz, I. D. Critical Evaluation Of Search Algorithms For Automated Molecular Docking And Database Screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189
(19) Ajay A. J.; Murcko, M. A. Computational Methods To Predict Binding Free Energy In Ligand−Receptor Complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(20) Bohm, H.; Stahl, M. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Indiana University-Purdue University and Wiley-VCH: Hoboken, 2002; Vol. 18, Chapter 2, pp 41−87.
(21) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular Recognition Of The Inhibitor AG-1343 By HIV-1 protease: Conformational Flexible Docking By Evolutionary Programming. *Chem. Biol.* **1995**, *2*, 317−324.
(22) Böhm, H. J. The Computer Program LUDI: A New Method For The De Novo Design Of Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61−78.
(23) Jain, A. N. Scoring Noncovalent Protein−Ligand Interactions: A Continuous Differentiable Function Tuned To Compute Binding Affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.
(24) Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik S. W. Evaluation Of PMF Scoring In Docking Weak Ligands To The FK506 Binding Protein. *J. Med. Chem.* **1999**, *42*, 2498−503.
(25) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(26) Tame, J. R. H.; Scoring Functions: A View from the Bench. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 99−109.
(27) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening Of Chemical Databases. 1. Evaluation Of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(28) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(29) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation Of 11 Scoring Functions For Molecular Docking. *J. Med. Chem.* **2002**, 2287−2302.
(30) Allen, D. G.; Coe, D. M.; Cook, C. M.; Dowle, M. D.; Edlin, C. D.; Hamblin, J. N.; Johnson, M. R.; Jones, P. S.; Knowles, R. G.; Lindvall, M. K.; Mitchell, C. J.; Redgrave, A. J.; Trivedi, N.; Ward, P. Pyrazolo-

[3,4-B]Pyridine Compounds, And Their Use As Phosphodiesterase Inhibitors. *PCT Int. Appl.* **2004**, 293.
(31) Raboisson, P.; Lugnier, C.; Muller, C.; Reimund, J.; Schultz, D.; Pinna, G.; Le Bec, A.; Basaran, H.; Desaubry, L.; Gaudiot, F.; Seloum, M.; Bourguignon J. Design, Synthesis And Structure/Activity Relationships Of A Series Of 9-Substituted Adenine Derivatives As Selective Phosphodiesterase Type-4 Inhibitors. *Eur. J. Med. Chem.* **2003**, *38*, 199−214.
(32) Daylight, [Online]; http://www.daylight.com
(33) Tripos, [Online]; http://www. tripos.com.
(34) Accelrys, [Online]; http://www.accelrys.com.
(35) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose P. W. Deciphering Common Failures In Molecular Docking Of Ligand-Protein Complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731−751.
(36) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; 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.
(37) Verdonk, M. L.; Cole, J. C.; Hartshorn M. J.; Murray, C. W.; Taylor, R. D. Improved Protein−Ligand Docking Using GOLD. *Proteins* **2003**, *52*, 609−623.
(38) Pearlman, D. A.; Charifson, P. S. Are Free Energy Calculations Useful In Practice? A Comparison With Rapid Scoring Functions For The P38 MAP Kinase Protein System. *J. Med. Chem.* **2001**, *44*, 3417−3423.
(39) 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.

CI050044X