How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information

Johannes Kirchmair,[†,‡] Simona Distinto,[†,‡] Patrick Markt,[†] Daniela Schuster,[†] Gudrun M. Spitzer,[§] Klaus R. Liedl,[§] and Gerhard Wolber*[,†,‡]

Department of Pharmaceutical Chemistry, Faculty of Chemistry and Pharmacy and Center for Molecular Biosciences (CMBI), University of Innsbruck, Innrain 52, A-6020 Innsbruck, Austria, Inte:Ligand Software-Entwicklungs- and Consulting GmbH, Clemens Maria Hofbauer-Gasse 6, A-2344 Maria Enzersdorf, Austria, and Institute of Theoretical Chemistry, Faculty of Chemistry and Pharmacy and Center for Molecular Biosciences (CMBI), University of Innsbruck, Innrain 52, A-6020 Innsbruck, Austria

Received November 13, 2008

Shape-based molecular similarity approaches have been established as important and popular virtual screening techniques. Recent applications have shown successful screening campaigns using different parameters and query selection. It is common sense that pure volume overlap scoring (or "shape-based screening") underrepresents chemical or pharmacophoric information of a molecule. Using the "Directory of Useful Decoys" (DUD) as a benchmark set, we systematically evaluate how (i) the choice of query conformations, (ii) the selection of the active compound to be used as a query structure, and (iii) the inclusion of chemical information (i.e., the pharmacophoric properties of the query molecule) affect screening performance. Varying these parameters bears remarkable potential for improvements and delivers the best screening performance reported using these tools so far. From these insights, guidelines on how to reach optimum performance during virtual screening are developed.

INTRODUCTION

In recent years, shape-based virtual screening methods have become increasingly popular in the field of computer-aided drug discovery and are established as important tools for virtual screening. At present, the highly optimized screening platform ROCS (Rapid Overlay of Chemical Structures)[1−3] is considered the de facto industry standard for shape-based, ligand-centric virtual screening. It uses a Gaussian function to define molecular volumes of small organic molecules. The available range of other shape-based screening tools and algorithms includes the algorithm CatShape as implemented in CATALYST[4] (application example by Singh et al.[5]), the Phase-Shape module as implemented in the Schrödinger product suite,[6] PARASHIFT,[7] and Hex[8] (as applied, e.g., by Perez-Nueno et al.[9,10]), ShaEP,[11] and the USR[12] (ultrafast shape recognition) algorithm. Putta and Beroza provide an excellent review on shape-based screening methods, alignment procedures, and current challenges.[13]

Successful studies using shape-based screening technology include the identification of ZipA-FtsZ protein−protein interaction inhibitors using ROCS[3] as well as the identification of a novel class of cannabinoid receptor 1 antagonists.[14] Freitas et al.[15] applied ROCS for screening on cruzain inhibitors, aiming at improved selectivity over cathepsin L. Shape-based screening showed excellent performance and was able to correctly classify the first 37 hits of a rank-ordered list. Bologa[16] et al. used ROCS for investigations on 30 selective G-protein coupled receptor agonists. Recently, Perez-Neuno et al.[10] published a comparative study on ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 (a CXC chemokine receptor) and CCR5 (chemokine (C−C motif) receptor 5) receptors using shape-based virtual screening (PARASHIFT, ROCS, and Hex) and ligand−receptor docking. Muchmore et al.[17] developed a probabilistic framework for the interpretation of similarity measures in order to directly correlate the similarity value to a quantitative expectation that two molecules are equally active. They investigated the performance of ROCS and nine other similarity methods in their ability to identify analogues and to perform scaffold hopping. Venhorst et al.[18] compared the scaffold hopping ability of their novel scoring function based on molecular interaction fingerprints against the scoring functions included to GlideXP,[19] GOLD,[20,21] ROCS, and a Bayesian classifier on two pharmaceutically relevant targets. The authors found that ROCS performs well in enriching compounds during virtual screening. Scaffold hopping performance increased when using their fingerprint-based scoring function. In a recent study, Hawkins et al.[22] found that ROCS performs at least as well as most docking programs in enriching active compounds. Sutherland et al.[23] have directly compared the performance of ROCS to well-established docking programs on eight proteins and found that ROCS performs competitive or better than the docking algorithms investigated. McGaughey et al.[24] investigated several different virtual screening approaches and found ROCS to perform very well in terms of enrichment. In our recent comparative study on structure-based pharmacophores and shape-based ROCS also we found comparable performance of both approaches,[25]

* Corresponding author phone: +43-512-507-5252; fax: +43-512-507-5269; e-mail: gerhard.wolber@uibk.ac.at.
† Department of Pharmaceutical Chemistry, University of Innsbruck.
‡ Inte:Ligand Software-Entwicklungs- and Consulting GmbH.
§ Institute of Theoretical Chemistry, University of Innsbruck.

10.1021/ci8004226 CCC: $40.75
Published on Web 02/18/2009

which corroborates the high potential and power of such approaches for virtual screening.

There are also studies available demonstrating synergistic effects of shape-based virtual screening in combination with pharmacophore-based searching techniques: Sykes et al.[26] have published a study on the successful prediction of nonspecific binding of drugs to hepatic microsomes by a molecular modeling approach that combines pharmacophore-based and shape-based screening. The technique was able to identify the 18 classified strong binders within the 22 highest ranked hits of the 56 compounds containing molecular library. One of the latest approaches uses the combination of structure-based pharmacophore modeling applying LigandScout[27,28] with ROCS shape-based screening for enhanced screening performance.[29,30]

Since the basic idea of shape-based screening is the generation of a complementary image of the binding site by considering the shape of a ligand, the assumption that the reference ligand (also called query ligand) needs to be represented in the bioactive conformation is straightforward. Studies based on crystal conformation queries include several examples presented by Hawkins et al.[22] and the work of Sutherland et al.[23] Lee et al.[31] use shape-based screening to find the appropriate receptor conformation for a single-receptor conformation docking in order to avoid CPU time expensive docking to a protein conformational ensemble.

However, information on the bioactive conformation of a ligand bound to a protein may not be available. In such cases, a single low-energy conformation can be used as query for shape-based screening. Examples of successful screening using calculated query conformations include experiments by Hawkins et al.[22] and Freitas et al.[15] Perez-Nueno et al.[10] used calculated docking poses as a starting point for shape-based screening. There are a few examples available indicating that the availability of the bioactive conformation is not an essential precondition for shape-based screening, which is why such methods may be classified as a ligand-based approach.[22] The developers of ROCS claim having evidence that using a low-energy conformation of their conformational model generator OMEGA[32] instead of the bioactive conformation has essentially no impact on the screening performance of their screening platform.[22] While the impact of conformational sampling of databases that are screened with ROCS has been studied thoroughly,[33] there is no systematic investigation available providing a clear statement or quantitative estimates on the impact of the query conformation(s) on virtual screening performance. Even more, ROCS supports multiconformer queries for virtual screening. However, also on the importance of this feature there were only limited data available so far.[10]

The scoring functions included in ROCS are based on two distinct aspects: the shape similarity ("ShapeTanimoto") and chemical pattern ("ColorScore") similarity (see the Methods section for more detail). Several studies indicate that ColorScore and "ComboScore" (i.e., the combination of ShapeTanimoto and ColorScore) obtain superior performance during virtual screening compared to ShapeTanimoto.[10,24,26] Until now there was no study available rendering a global image of the performance of these scoring functions. Even more, the question arises whether the combination of ShapeTanimoto and ColorScore with exactly equal weights corresponds to optimum performance or if there is space for improvement. In this work, we provide an answer on this point.

In their recent work, Hawkins et al.[22] point out that—while docking usually suffers from a high false positive rate (i.e., inactive compounds are not correctly identified as being inactive during virtual screening)—the problem of shape-based screening is mainly its inferior false negative rate (i.e., a low score is incorrectly assigned to active molecules and therefore these molecules are not retrieved in the early stage[33] of virtual screening). The basic assumption of shape-based similarity screening is that molecules of shape and chemistry comparable to known active agents have a significant probability of also showing activity. Therefore, actives with shape differing from that of known actives are likely to be missed during virtual screening. However, larger compounds may be able to bind even tighter to a protein by presenting a larger surface and more interaction spots to the target and smaller molecules are generally favorable when searching for small lead structure cores with high potential for optimization by medicinal chemistry approaches. Hence, considering multiple compounds of distinct shapes may be favorable for shape-based virtual screening.

In this study, we investigate the impact of the query conformation on the performance of shape-based screening and analyze the possible gain in performance by using conformational ensembles of query structures for screening. Furthermore, we study the performance of ShapeTanimoto, ScaledColor, and ComboScore and highlight possible improvements for the latter one. We also quantify the benefit of using multiple compounds for screening in parallel and provide guidelines for best performance of ROCS. All results are put in direct relation to the results obtained with the same data set using the DOCK algorithm as published by Huang et al.[34]

METHODS

The ROCS Shape-Based Virtual Screening Technology. ROCS is a shape-based method for rapid similarity analysis of molecules. Thereby, the volume overlap of two molecules is assessed by Gaussians, which are parametrized according to the hard-sphere volume of heavy atoms. The use of Gaussians allows for fast calculation of overlaps between two atoms, as the product of two Gaussians is again a Gaussian. In the first step, ROCS assesses the centers of mass of the query and the candidate compound and subsequently starts aligning their principal components of intertia.[22] In the second step, the initial orientations of the two molecules are optimized applying a solid-body optimization algorithm in order to maximize volume overlap. As of version 2.3, ROCS considers both shape and a color force field for optimization of the overlap.[35]

The basic scoring function implemented in ROCS is the ShapeTanimoto score, which is a quantitative measure for the shape overlap of two molecules. By focusing on the optimum shape overlap, however, one is likely to run into the problem that compounds—in particular if they are of distinct dimensions—are aligned in an unfavorable way, with corresponding chemical functions not matched. In order to obtain activity on a certain target, not only shape but also appropriate chemical functionality are crucial for a com-

pound, which is why ROCS has implemented a color score to support the alignment process and evaluate chemical feature-based similarity. The ROCS color force field is composed of SMARTS patterns for the characterization of chemical functions in combination with a rule set that describes how such functions interact. ROCS provides two color force field flavors—the ImplicitMillsDean and the ExplicitMillsDean force field. Six different types encode the chemical functionality of molecules: hydrogen-bond donors, hydrogen-bond acceptors, hydrophobes, anions, cations, and rings. As the ImplicitMillsDean force field includes also a basic pK_a model assuming pH 7, charges are assigned accordingly in an automated way, and users do not need to protonate molecules accordingly before virtual screening. ColorScore is the scoring function that quantifies the matching of these chemical functionalities between the query compound and the molecule being screened. It is not independent from the query molecule; the query self-color represents the maximum ColorScore value that can be obtained. Hence, the ColorScore of the actual compound being screening divided by ColorScore of the query allows for scaling score values between 0 and 1, which is referred to as "ScaledColor" score.

In this way, ROCS is able to maximize both molecular shape overlay and chemical functionality overlap. The ComboScore function puts exactly equal weights on its both components, the ShapeTanimoto and the ScaledColor score. Both components obtain values between 1 and 0 and are summed up for the ComboScore. Hence, ComboScore values range from 2 to 0, where 2 stands for the best possible overlap and 0 for no similarity. Overall, this screening and scoring approach is straightforward, based on a few basic assumptions, and—as also observed with related techniques like, e.g., pharmacophore modeling[36]—this elementariness is a key feature for the global robustness of such methods.

The Directory of Useful Decoys (DUD). The DUD[34] is based on the idea of providing a public, well-defined and—as far as possible—unbiased data set of annotated active compounds and decoys for the validation of docking programs. Currently, it is considered the industry standard for the validation of docking protocols, and recently it has been applied also for the evaluation of related structure-based and ligand-based virtual screening methods.[37] It comprises 2950 known ligands for 40 different, pharmaceutically relevant targets. 36 physically similar yet topologically dissimilar decoys (putative inactive compounds) have been selected for every of these 2950 ligands, summing up to a database of 98,266 compounds. Dissimilarities of ligands and decoys have been calculated based on fingerprints, and five basic molecular descriptors have been used to select decoys that match molecular weight, number of hydrogen-bond acceptors, number of hydrogen-bond donors, number of rotatable bonds, and logP of the ligands. Therefore, the DUD represents—so far—one of the best and most comprehensive collections of actives and decoys, with structural data on the target available.

Software and Hardware Setup. ROCS virtual screening has been performed on five Intel 6600 Core2Duo processor machines with 2GB RAM running Linux Fedora Core 6. OMEGA 2.0 was applied for multiconformational database generation using default settings, ROCS 2.3.1 was used as screening platform, and Pipeline Pilot 6.0 was applied for compound clustering and pharmacophoric fingerprint similarity assessments. Query conformations were calculated with CORINA 3.00,[38−40] SYBYL 8.0,[41] and OMEGA 2.0. Perl scripts were applied for data elaboration, and GnuPlot 4.0 was used for receiver operating characteristic (ROC) curve plotting.

Workflow Description. Considering guidelines for reproducible and unbiased performance assessments that we have published recently,[33] we have followed a straightforward and transparent study design: we selected the DUD deliberately in order to avoid any input bias on our part. Manual data curing and manipulation was reduced to an absolute minimum, and each working step is concisely described in the following section.

The DUD has been downloaded directly from dud.docking.org/r2 (DUD release 2, October 22, 2006) in mol2 file format. As OMEGA has proven reliable performance for the generation of conformational ensembles for 3D virtual screening,[25,42] we chose OpenEye's conformational space sampling tool using default settings for our investigations on ROCS. Screening with ROCS was performed based on default settings—with settings under investigation adapted as stated below. Since the DUD compound selection includes several duplicates (e.g., several tautomers per compound), we used a Perl script to remove all but the highest ranked occurrence (according to the scoring function investigated) of each molecule from the report files obtained from ROCS virtual screening—a procedure that is applied frequently also for the analysis of docking results. DUD errata, as published on Dec 7, 2007, on http://wiki.compbio.ucsf.edu/wiki/index.php/Main_Page have been removed from these report files. Subsequently, Perl scripts were applied to calculate statistical measures and to execute plotting of receiver operating characteristic (ROC) curves[43] using the GnuPlot engine. Again, we tried to exclude any bias as far as possible by sticking to well-established performance measures and cutoff values (see below).

Performance Metrics. The general aim of virtual screening methods is to retrieve a significant larger fraction of true positives from a molecular database compared to a random compound selection. If a virtual screening method selects n molecules from a database with N entries (A compounds of the database being active), the selected hit list comprises active compounds (true positive compounds, TP) and decoys (false positive compounds, FP). Active molecules that are not retrieved by the virtual screening method are defined false negatives (FN), whereas the unselected database decoys represent the true negatives (TN). The enrichment factor (EF) represents one of the most prominent performance descriptors in virtual screening.[44−46] It takes into account the improvement of the hit rate by a virtual screening protocol compared to a random selection (eq 1).

$$\mathrm{EF} = \frac{TP/n}{A/N} \tag{1}$$

One disadvantage of the EF is its high dependency on the ratio of active molecules of the database screened.[43,47] This bottleneck makes the descriptor unsuitable for the direct comparison of performance investigations and screening runs that are based on test sets of different ratio between actives and inactives. In our case, however, the ratio between actives

Table 1. Overall Performance for ROCS on All 40 DUD Targets Using CORINA, OMEGA, SYBYL, and X-ray Conformations

	AUC av | sd[a]		EF1% av | sd[a]		EF5% av | sd[a]		EF10% av | sd[a]		EF20% av | sd[a]	
Corina	0.71	0.22	18.5	13.0	8.5	6.0	5.0	3.0	2.9	1.5
omega_highestE	0.71	0.22	19.4	12.0	8.9	6.2	5.1	3.1	2.9	1.5
omega_lowestE	0.72	0.21	20.2	12.9	8.8	6.0	5.1	3.0	2.9	1.4
Sybyl	0.72	0.20	18.2	11.9	8.1	5.8	4.8	2.9	2.9	1.4
Xtal	**0.73**	0.20	**19.4**	**12.9**	8.4	6.2	5.1	3.0	**3.0**	1.5

[a] av, average; sd, standard deviation.

and decoys of all different targets is almost identical (i.e., 36 decoys per known active compounds). Hence, direct comparisons between targets are possible—as well as comparisons to evaluations of any virtual screening technology based on the DUD test set. Another disadvantage of the EF is that all actives contribute equally to the value. On that account, the EF does not distinguish highest-ranked active molecules from actives ranked at the end of a rank-ordered list. In other words, two virtual screening methods that differ in the ability of ranking the highest scored active molecules at the beginning of such a rank-ordered list but show the same enrichment for active molecules would be assessed to perform equally.[47] We conquer this issue by calculating EF values for every percent of the database screened. This procedure allows insight on the success or failure of a screening campaign at any stage of the virtual screening process. Depending on the technological resources for biological testing, researchers may prefer to biologically validate only a small selection of compounds or to test up to 20% of the database. Performance snapshots such as EFs are particularly useful for characterizing the reliability of virtual screening protocols, highlighting how many actives are found within in a hit list at a certain stage of virtual screening. Popular EFs for the characterization of virtual screening performance are 1%,[10,34] 2%,[48] or 5%[10] of the database for early enrichment rates and 10%[48] or 20%[34] for late enrichment values. In this study we report EF1%, EF5%, EF10%, and EF20% values in order to render a comprehensive performance spectrum.

ROC curve analysis is considered the best approach for performance characterization of virtual screening protocols so far. These graphs describe sensitivity (Se) for any possible change of n as a function of the specificity (1-Sp). For more information on ROC curve analysis the reader is referred to the work of Triballeau et al.[43] The area under the ROC curve (AUC) represents another descriptor for virtual screening performance and can be calculated as the sum of all rectangles formed by the Se and 1-Sp values for the different thresholds.[43] For ideal distributions of actives and decoys an AUC value of 1 is obtained; random distributions cause an AUC value of 0.5. Virtual screening workflows that perform better than a random discrimination of actives and decoys retrieve an AUC value between 0.5 and 1, whereas an AUC value lower than 0.5 represents the unfavorable case of a method that has a higher probability to assign higher scores to decoys than to actives.

The ROC curve and the AUC value are useful and easily manageable evaluation techniques for determining the discriminatory power of virtual screening methods. In contrast to the EF the AUC does not depend on the ratio of actives to decoys in a database. However, the AUC value itself suffers from the disadvantage that two virtual screening methods cannot be discriminated according to their ability of recognizing actives at the beginning of an ordered list. For example, an identical AUC value for two different virtual screening workflows does not mean that both workflows are equal in scoring the actives of a database. Therefore it is necessary to consider both AUC values and EFs for early enrichment in order to allow quantitative conclusions on the performance of virtual screening protocols. More information on performance descriptors for virtual screening is provided in a recent review.[33]

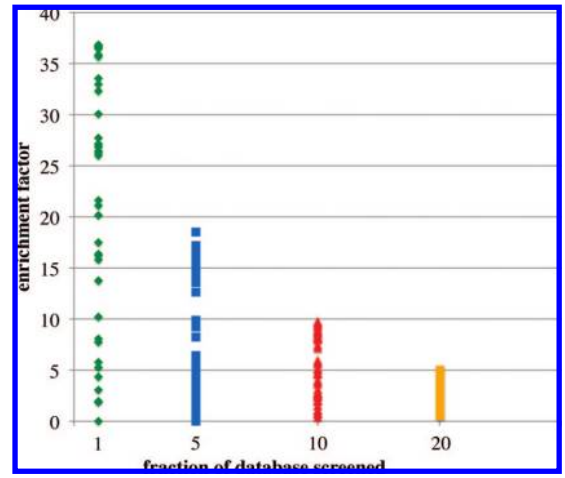

Figure 1. Spread of enrichment factors calculated at different stages (percent of database screened) of virtual screening. The spread of the enrichment factors obtained for the DUD targets narrows down as more compounds of the data sets are screened. The highest-ranked compounds are of particular importance for the success of a virtual screening tool, since these molecules are the ones being selected for biological validation experiments and further development.

RESULTS AND DISCUSSION

In this study, the screening performance of ROCS was investigated considering implications of the query conformation, possible benefits by using multiconformer and multicompound queries, and refined scoring functions for optimum performance. So far, computational chemists have preferred to use an experimentally determined query conformation, if available. As there are experimental structural data on protein−ligand complexes for 39 of the 40 DUD targets available (a modeled structure has been used for PDGFRB kinase) we deliberately chose the X-ray conformation of the ligand selected by Huang et al. as the starting point for our investigations. That way, we avoid any bias during compound selection from our side, and we are able to render the performance of ROCS in the desirable case, that the bioactive conformation of the ligand is known (which is supposed to be the best case scenario for structure-based methods,

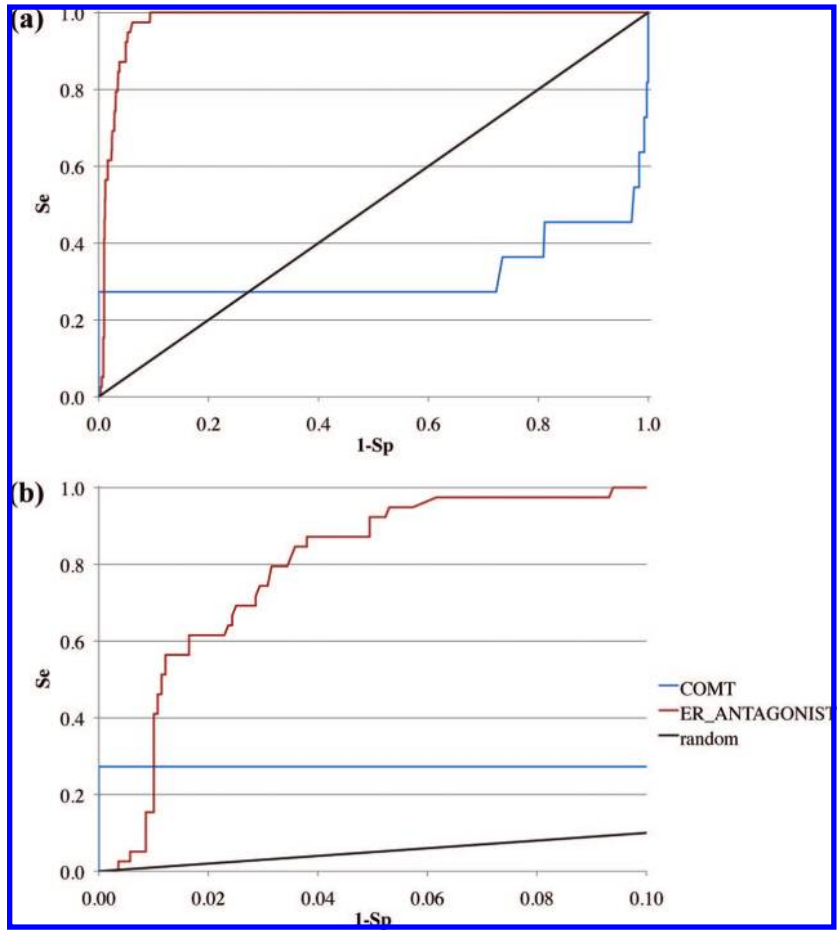

Figure 2. Limits of AUC and EF: The AUC value obtained for ER_ANTAGONIST (0.98) suggests excellent performance of ROCS. A closer look on the enrichment factor at 1% of the database screened, however, reveals that the top-scoring compounds are mostly inactives. The other case is found for the COMT test set, where a low AUC value (0.32) indicates screening failure even though the highest ranked compounds are active. (a) AUC of COMT and ER_ANTAGONIST and (b) ROC curve detail of the highest-ranked fraction of the screening database.

representing the best data basis). Furthermore, this allows us to put the performance of shape-based screening using ROCS and docking using DOCK (by Huang et al.) into direct relation.

Performance of ROCS Using the Bioactive Conformation as Query Structure. Overall, ROCS obtains good virtual screening performance (Table 1): the average AUC obtained for the DUD targets is 0.73, which exactly matches the value reported by Nicholls,[37] who provides an overview of the performance of ROCS compared to fingerprint-based screening and docking. Average initial enrichment is high (EF1% 19.4), meaning that a large portion of actives is ranked as the top virtual screening hits, together with only a few decoys. EF1% shows significant spread over all DUD targets tough (sd ± 12.9, Figure 1). However, high initial enrichment rates do not necessarily imply high AUC values: in the case of the ER_ANTAGONIST set, for example, despite the excellent AUC value of 0.98, EF1% is only 5.3 (Figure 2). This shows that the overall hit ranking is good, however, not the scoring of the foremost compounds of the hit list, which are certainly the most interesting ones. On the other hand, we detect a high EF1% of 30.1 for COMT, with AUC below 0.40. In this case, the performance of ROCS over the whole test set is insufficient, while biological testing results for the first percent of top-ranked compounds will delight every computational chemist. Nevertheless, a correlation between EF and AUC can be detected at later stages of virtual screening, at EF10%, and—more developed—at EF20% (Figure 3). Hence, snapshots on late enrichment rates allow estimates on the AUC and vice versa. While early enrichment rates spread widely over the DUD, late enrichment rates show significantly decreased variation (Figure 1). Overall, these observations highlight once more the limitations of both AUC and EF when interpreted segregated and ask for consideration of both benchmarks in a complementary way in particular in the early stages of virtual screening.

Data sets obtaining very good AUC values (i.e., AUC > 0.90) are ER_ANTAGONIST, SAHH, NA, RXRα, EGFR, ER_AGONIST, COX2, GART, DHFR, PPAR_GAMMA, HMGA, GBP, and PNP (in order of decreasing AUC). ROCS fails (i.e., AUC < 0.50) in the case of FGFR1, VEGFR2, FXA, SRC, PDGFRB, and COMT. Unfortunately, there are no AUC values reported in the work of Huang et al. for a direct comparison to DOCK.

Excellent early enrichment rates (i.e., EF1% > 35) are obtained for HSP90, INHA, MR, GPB, HMGA, AMPC, and COX2; ROCS fails (i.e., EF1% < 2.5) in the case of GART, TRYPSIN, VEGFR2, FGFR1, THROMBIN, and SRC (compared to DOCK, which fails in the case of INHA, COMT, PPARG, PR, PDGFRB, FGFR1, SRC, VEGFR2, ACHE, EGFR, and P38). All benchmark data are provided as Supporting Information.

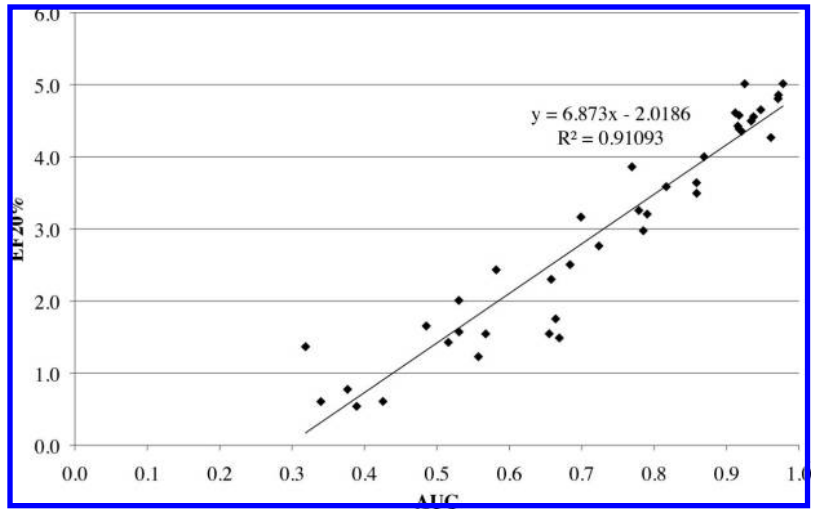

Figure 3. Correlation between the AUC and EF20%. The graph demonstrates that late enrichment factors (here represented by EF20%) allow estimating the global performance of a virtual screening protocol (represented by the AUC).

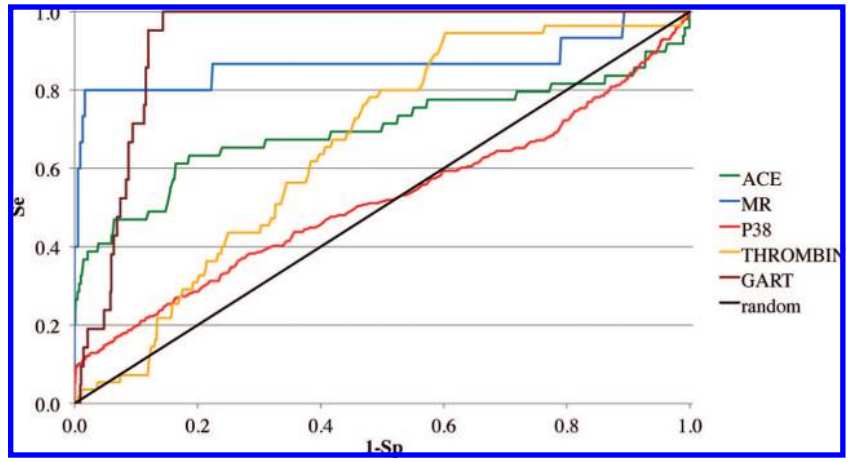

Figure 4. ROC curves obtained for representative examples of each target family. Due to the small sample of targets per protein class statistically significant conclusions on the performance of ROCS are not legitimate; however, trends appear to exist that folate enzymes perform very well, while kinases and serine proteases are challenging.

Both methods, ROCS and DOCK, show some parallels on which target they perform well and on which proteins they fail to enrich active compounds. Serine proteases and protein kinases seem to be the most challenging for ROCS. The latter protein family caused high failure rates also with the DOCK approach (i.e., in six out of nine protein kinases). One reason for failure could be structural issues with the bioactive conformation. In this case, however, such problems can rather be excluded since most of the calculated conformations do obtain inferior performance. Differences in the characteristics of the DUD test sets could be another explanation for failure; however, the design of the DUD aims to minimize differences by considering topological dissimilarity and physicochemical properties. On this respect, we have investigated the structural (dis)similarity between each query structure and the respective DUD test set as well as the distances within the DUD compound sets. The number of clusters of actives and inactives was calculated applying a Tanimoto coefficient similarity cutoff of 0.7 based on ECFP4 fingerprints. We found that the average ratio between the clusters formed by the actives and the inactives is 26—with significant spread over the DUD data sets (sd ± 17). So we can detect (dis)similarity discrepancies between actives and inactives of different test sets, however, they do show no noticeable correlation with the screening success or failure of ROCS. In other words, these data suggest that the performance of ROCS is not necessarily connected to the variety of chemical scaffolds represented in the actives compared to the inactives sets. The presumption that the query ligand—the compound cocrystallized with the protein—may be distinctively represented by the actives set is legitimate. A closer look on this aspect, however, reveals no significant irregularities: the average distances (measured by the same Tanimoto coefficient as described above) between the query compound and the actives and decoys sets, respectively, are almost identical—with the maximum deviation found to be only 0.04. Hence, considerable blurring of the evaluation results due to distinct representations of the cocrystallized ligand in the DUD sets seems to be rather unlikely.

Target Family Affiliation and Characteristics of the Individual Test Sets. Dependence of virtual screening performance on the target family is evident but difficult to quantify since only a few targets per target family are represented in the DUD (see ref 34 for the target classification). Our results (Figure 4; more data are provided as Supporting Information) indicate that folate enzymes perform well (average AUC 0.92); lowest performance on average

is obtained for kinase targets (average AUC 0.59, average EF1% 14.4, respectively), which is in agreement with the observations reported by Huang et al. for the DOCK algorithm (see below). An explanation for the difficulties observed for this protein family may be the characteristics of the ATP binding sites of protein kinases. These interaction sites are in general structurally highly conserved and very flexible; size and shape of the active sites are largely depending on the presence and properties of ligands and allosteric agents. Therefore, developing ligands interacting with the ATP binding pocket by computational methods has been shown to be highly challenging, and attempts to design specific kinase inhibitors have given way to the development of agents with favorable kinase activity spectra.

Comparison of Shape-Based Screening and Docking. Docking is considered the ruling technology in structure-based drug design. After decades of development, however, docking still faces severe issues that could not be solved so far, and thus the usefulness of docking in virtual screening being discussed controversial. Leaking estimation of entropic effects, protein flexibility, and the thereby linked issues of scoring are among the major points of critique. In our opinion, docking is one of the most interesting and perhaps most promising way of rational drug discovery. Docking not only is a virtual screening technique but also allows deducing possible binding modes, to analyze interactions between the protein and the ligand and to gain new ideas on how an optimum ligand should look like. On this respect we are aware that it is quite daring to put ligand-based virtual screening and docking into direct relation. Hence, in the following section we are not intending to promote a certain approach as being superior compared to another one; we would rather like to point out the relations between both approaches and are convinced that best computational modeling relies on the application and combination of an ensemble of different techniques.

Huang et al.[34] have published the design of the DUD accompanied by an examination of the docking program DOCK.[49] The benchmark defined for the evaluation of the performance of docking was EFmax, EF1%, and EF20%. From our experience, we think that EF1% and EF20% are most relevant for the characterization of the performance of virtual screening protocols: EF1% as a benchmark for virtual screening campaigns where only a very few compounds can be forwarded to biological testing facilities; EF20% as an estimate for the performance in industrial HTS environment. Average enrichment rates for ROCS are EF1% 19.4 and EF20% 3.0, respectively (Table 1); DOCK obtains EF1% 17.3 and EF20% 2.6, respectively. Poor performance of DOCK (i.e., EF1% < 2.5) is obtained for eleven of the DUD targets—six of those being kinases: EGFR, FGFR1, P38, PDGFRB, SRC, and VEGFR2. In this respect, also ROCS fails in the case of FGFR1, SRC, and VEGFR2, however, provides superior enrichment rates for EGFR (EF1% ROCS 33.0 vs DOCK 2.1), P38 (EF1% ROCS 10.2 vs DOCK 2.1), and PDGFRB (EF1% ROCS 7.7 vs DOCK 0.0). Nevertheless, the global ROCS performance for these problematic targets is low for P38 (AUC 0.51) and PDGFRB (AUC 0.34) and excellent only in the case of EGFR (AUC 0.95). Contrary to that, ROCS fails in enriching compounds in cases where DOCK shows good enrichment rates: GART (EF1% ROCS 0.0 vs DOCK 42.4), TRYPSIN (EF1% ROCS 0.0 vs DOCK 22.5), and THROMBIN (EF1% ROCS 1.8 vs DOCK 13.7). Overall, differences in performance between both approaches appear to be rather minor and depend on the target (the standard deviation of the AUCs among all DUD targets with ROCS is 0.20). However, it is interesting to observe that currently the consideration of the protein structure (as in the case of docking) does not lead to better enrichment rates than ligand-centered methods, which do not rely on the availability of structural data on the target. The data corroborate once more that the careful selection, validation, and combination of virtual screening approaches allows enhanced hit retrieval rates for a particular target.

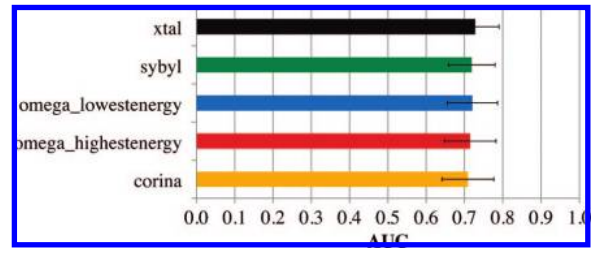

Figure 5. Average AUC values for all 40 DUD targets using the conformation of the structure model (xtal) and 3D structure generators. The performance of ROCS with any reasonable input conformation is comparable. The error bars indicate the 95% confidence level.

Different Query Conformations and their Impact on the Performance of ROCS. In order to render a global image on the impact of the query conformation on the performance of ROCS we have analyzed its success in enriching active compounds for all DUD targets using the bioactive conformation in direct comparison to calculated conformations from different sources. We have generated 3D queries with CORINA and OMEGA from scratch. Using OMEGA, conformational models have been generated with default settings, and the performance of the lowest and highest energetic conformer has been investigated, respectively. 3D CORINA structures were generated using default settings. Furthermore, the bioactive conformation derived from the 40 DUD protein−ligand complexes has been minimized using the Powell conjugate gradient[50] minimization method in combination with the Tripos force field in SYBYL.

Our data show that there is basically no difference in the performance of ROCS with respect to the query conformation (considering both AUC as well as EFs). ROCS shows robust enrichment with all different kinds of query conformations, even with the highest energy conformation generated with OMEGA (Figure 5). There is no significant performance loss detectable when using calculated query conformations instead of the experimentally determined one. The reason for this homogeneous performance is obviously the fact that OMEGA allows to enumerate all relevant conformations of the low-energy conformational space of the database molecules and is able to represent the bioactive conformation.[42] As long as the query conformation is represented in the conformational model of the database compound ROCS is able to identify related compounds correctly.

The average AUC values for all DUD targets are located around 0.72 (sd ± 0.21) and early enrichment factors (EF1%) around 19 (sd ± 13). The high standard deviation of EF1% indicates significant performance differences among the DUD

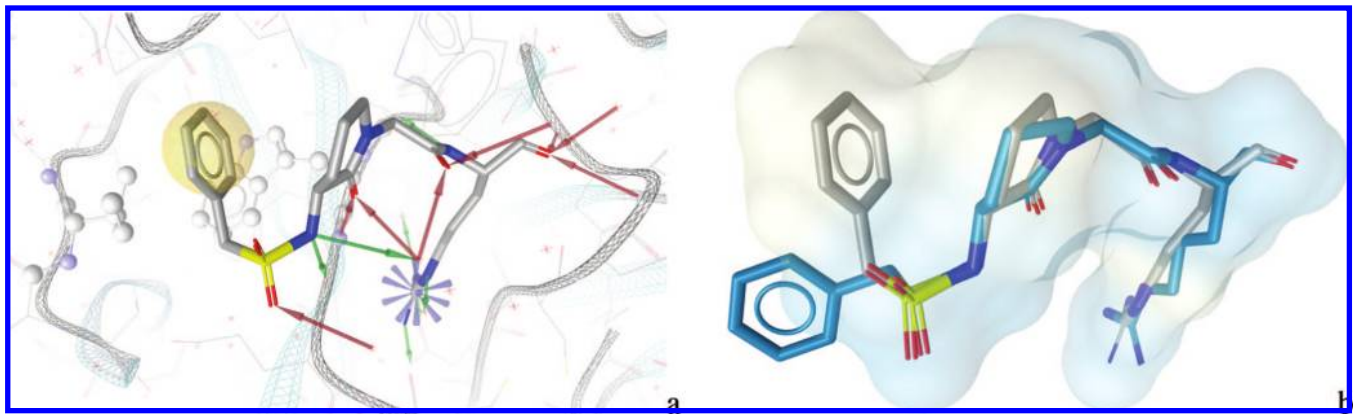

Figure 6. Screening for thrombin inhibitors: (a) Pharmacophore model of PDB complex 1ba8 showing interactions of the ligand and the environment in LigandScout. The convoluted conformation is stabilized by the phenyl ring forming hydrophobic interactions with ILE174 and pi-pi interactions with TRP215 (both visualized in ball-and-stick mode). (b) Alignment of the best-matching OMEGA conformation (blue color) with the query (gray color) in the conformation of the structural model. The conformational ensemble calculated with OMEGA does not include the higher-energetic conformation, which is why a calculated low-energy conformation as ROCS query obtains superior recognition rates among databases consisting of calculated conformational ensembles.

targets (Figure 1). Looking at the individual performances of each target, only in a few cases we are able to detect considerable discrepancies between conformations from different sources. AUCs smaller than 70% of the AUC obtained by the bioactive conformation are reported for CORINA on HIVPR, OMEGA's highest energy conformation for PARP, and OMEGA's lowest energy conformation for PDGFRB. In the case of THROMBIN, the conformations generated with both CORINA and OMEGA reach AUCs values below 70% of the AUC obtained when using the bioactive conformation. For THROMBIN the reason seems to be the convoluted bioactive conformation, which is not present in the calculated query conformations (Figure 6). Another possibility would be that the large conformational space of thrombin inhibitors, which in part can be represented by the molecular weight (467 in case of the query structure) and the number of rotatable bonds (14 in case of the query structure), is responsible for this. However, we were not able to detect a general correlation between the accessible conformational space of active compounds and screening performance. Moreover, there is no obvious correlation between the accessible conformational space of the query compound and its performance during screening detectable. It seems plausible that the dimensions of the accessible conformational space do have an influence in virtual screening performance, but these effects may be masked by other interfering influences.

Our data suggest also that ROCS does not benefit from using energy-minimized query conformations derived from the experimentally determined one. This does not necessarily indicate force field issues but is likely to be caused by a bias of OMEGA to matching bioactive conformations (i.e., conformations as published in PDB structural models). It must be assured that the conformational model generator is capable of representing the query conformation so that compounds in the database similar to the query conformation can be identified correctly during virtual screening. As a guideline, it seems advisable to stick to the experimentally determined conformation without postprocessing unless structural issues are encountered or to use a calculated low-energy conformation.

In order to estimate the robustness of the method, we also did experiments using problematic conformations of query compounds (e.g., steric clashes within the query structure) and found that severe loss of performance is rather unlikely even in such unfavorable cases. Overall, these data prove that ROCS is a ligand-based method, which is capable of high performance virtual screening without dependence on the availability of the bioactive ligand conformation.

Multiconformer Queries and Their Impact on the Performance of ROCS. ROCS has implemented a feature for handling multiconformer queries for virtual screening. In this mode, ROCS loops over all conformers of each molecule of the screening database and compares them to all query conformers. By this a data matrix is generated and by default ROCS reports the best overlay. Shape-based screening represents a rather slow virtual screening technique—faster than docking—but slower than fingerprint-based similarity measures or 3D pharmacophore models.[25] In this multiconformer query mode, screening time increases linearly with the number of conformers. In order to keep a balance between efficiency and reliability, a significant increase in prediction accuracy would be desirable. In general the assumption, that multiconformer queries may obtain better virtual screening performance, may seem plausible. We analyzed the performance of ROCS using conformational ensembles generated based on the ligand of the PDB structure used by Huang et al. in their evaluation of the DOCK program. The conformational ensembles were generated without providing a 3D seed structure and consist of one, three, five, and ten conformations, respectively, using the −maxconfs flag in OMEGA. Our data indicate that none of the investigated conformational ensembles allow for increasing the performance of ROCS. On average, none of the performance benchmarks—neither AUC nor EF snapshots—show superior performance of conformational ensembles compared to the single conformation query (Table 2). Also the values for the individual targets obtained with multiconformational queries comply largely with the respective single conformation queries (Figure 7). This is another indication for a smooth representation of the low-energy conformational space by OMEGA conformational ensembles, since any reasonable conformation seems to be adequately and comparably well represented by the calculated models. On that account it is plausible that using several well-embedded query conformations instead of a single one does not alter the

Table 2. Implication of Using Multiple Confomers per Query for Virtual Screening[a]

number_of_conformers	AUC	EF1%	EF5%	EF10%	EF20%
1	0.72	20.2	8.8	5.1	2.9
3	0.73	20.3	9.0	5.2	3.0
5	0.73	20.4	8.9	5.2	3.0
10	0.74	20.2	9.1	5.3	3.0

[a] Different 3D coordinates show practically no effect on screening performance.

outcome. Even more, shape is depending on conformation to a considerably lower extent than atom positions.

Multicompound Queries and Their Impact on the Performance of ROCS. Data fusion is a common approach applied in a widespread field of disciplines for decision support. Thereby, the usage of a set of sensors instead of a single one is expected to increase success rates by introducing additional information. One of the most prominent applications of data fusion methods in virtual screening is consensus scoring, where multiple scoring functions are applied in order to increase the predictive power of docking approaches.[51,52] Other examples include the application of data fusion for similarity-based virtual screening approaches.[53,54] Based on the insight and ideas gained from preceding analyses and the encounter of problematic false negative rates during virtual screening with ROCS, we tried to increase screening performance by using multicompound queries employing data fusion. Thereby, we aimed at the elucidation of guidelines on how to tackle virtual screening campaigns in cases where several active molecules are already known. It would be interesting to see if knowledge on several active molecules helps to increase hit retrieval rates with ROCS.

We clustered all actives sets by maximum dissimilarity using ECFP4 fingerprints in Pipeline Pilot. These fingerprints allow abstracting certain compound properties, considering basic pharmacophoric functions such as hydrogen bond donors and acceptors. We generated one, three, and five clusters per actives set and selected the respective cluster centers as query molecules for ROCS screening. In other words, we selected the representative compounds that depict the chemical space covered by the actives sets consistently. All compounds used as queries were certainly removed from the screening library. The lowest energetic conformation generated using OMEGA defaults was selected as query structure.

ROCS screening using the center molecule of the actives set of each target obtains an average AUC of 0.82, which is 14% higher than using the ligand present in the respective X-ray structures (Table 3). Contemporarily, EF1% increases from 20.2 to 34.4 by 70%. Less dramatic improvements are achieved for late enrichment rates, leading to an increase of EF20% by 26%, compared to the X-ray ligand query. This gain in performance seems plausible, since the X-ray ligand is not as well embedded in the actives set as the center molecule selected based on pharmacophoric fingerprints. In this way it appears that using the center molecule of the actives sets is a suitable way to optimize screening accuracy. Individual AUC values within multiquery runs on a particular target spread considerably: in the case of three-compounds queries, the average AUC of the best performing query structure is 22% higher compared to the average of all queries (Figure 8). The maximum distance between the best performing query and the average of all queries was found for COX2 (AUC 75% higher AUC for a particular compound compared to the average of the three-compounds query). The reason for this is that one of the three query compounds—the query obtaining inferior results—represents a significantly smaller fraction of the actives set than the other two compounds. For five-compounds queries, these gaps between the average values and the ones of the best performing query are 30% for AUC, with the maximum deviation being 79% (again with COX2 and again because of the different cluster sizes).

So far we have seen that picking the right query compound(s) (i.e., the center molecule of the compound set or the compound representing the largest scaffold families of the compound collection, respectively) leads to considerable gain of performance using ROCS. Nevertheless, we were convinced that collecting the structural information provided in form of several different active compounds could be used to increase hit retrieval rates even further. Hence, instead of looking at the screening success of the individual compounds, we analyzed the performance of the compound ensembles by considering the highest ComboScore obtained for any of the database compounds with any of the query compounds. In other words, we created a data matrix containing all score values for all NxM overlays and considered the high-score values for hit ranking exclusively. Indeed, we found significant improvement of virtual screening performance using the high-score values (Figure 9): applying this ranking procedure for the three-compounds queries, the average AUC increases to 0.90 (compared to screening with the center molecule, this is an improvement of 10%) and also the average EFs increase by approximately 10% (compared to screening with the center molecule). Based on the high-scores of five-compounds queries, the average AUC improves even more, reaching the excellent average of 0.94 over all DUD targets and also the EFs continue the uptrend. Contemporarily, standard deviations halve in value, indicating a gain of robustness over all targets. Thirteen out of 40 DUD targets reach AUC 1.00 with five-compounds high-score ranking; the lowest level is obtained for HIVRT (AUC 0.64), which is considerably better than with the single, center molecule (AUC 0.32). This trend is continued when using more than five active compounds as query structures but would end up with the not surprising observation that the most-comprehensive multicompound queries achieve perfect performance.

Overall, the high-score ranking procedure, considering only the highest score value obtained for the multicompound query, overlays with the database molecules and allows for increasing the discriminatory power and robustness of ROCS considerably. This effect, the lowering of the signal-to-noise ratio by a set of queries or models, is consistent with results found for pharmacophore-based parallel screening, where multiple models applied in parallel allow superior hit retrieval rates.[55,56]

Refinement of the ComboScore for Optimum Performance of ROCS. ComboScore combines ShapeTanimoto and ScaledColor. The assignment of exactly the same weights to ShapeTanimoto and scaled ColorScore in ComboScore is a straightforward assumption yet leaves space for optimization, without losing universality and robustness by overparameterization. The first question arising on this

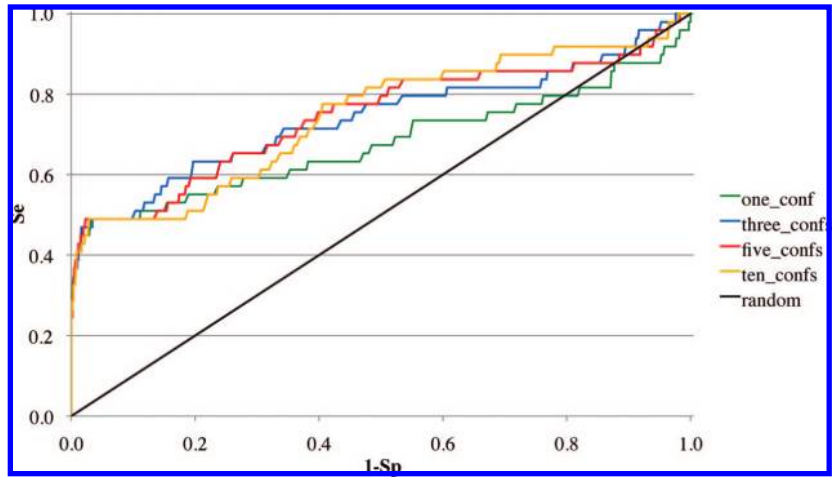

Figure 7. Representative example of a screening using a multiconformer query. As shown here for the ACE data set, no significant benefit from using several conformers per query compound for screening should be expected: enrichment rates are comparable, while demands in computational power rise linearly.

Table 3. Implication of Using Multiple Compounds as Queries for Virtual Screening[a]

target	AUC 1 | 3 | 5			EF1% 1 | 3 | 5			EF5% 1 | 3 | 5			EF10% 1 | 3 | 5			EF20% 1 | 3 | 5		
ACE	0.61	0.73	0.90	27.3	19.8	29.7	7.7	10.1	8.6	4.5	5.7	6.0	2.5	3.1	4.1
ACHE	0.81	0.90	0.93	33.1	36.2	35.2	11.8	13.0	14.3	7.8	8.6	8.8	4.0	4.5	4.6
ADA	0.82	1.00	1.00	46.6	46.6	46.6	12.5	19.3	19.3	6.2	10.1	10.1	3.6	5.0	5.0
ALR2	0.58	0.61	0.91	29.8	29.8	39.7	7.8	7.8	15.6	4.3	4.3	8.2	2.6	2.4	4.3
AMPC	0.96	0.95	0.88	49.8	42.7	0.0	16.2	14.8	2.7	8.7	8.7	6.1	4.7	4.3	4.0
AR	0.61	0.93	0.95	30.1	37.6	37.6	7.6	14.9	15.2	3.8	8.6	8.9	1.9	4.5	4.6
CDK2	0.66	0.75	0.68	19.3	33.7	28.9	8.6	10.0	10.5	4.3	6.0	5.5	2.5	3.2	3.0
COMT	0.86	0.96	0.92	72.7	18.2	36.3	13.8	6.9	13.8	8.4	10.1	8.4	4.2	5.0	4.2
COX1	0.56	0.38	0.75	0.0	5.7	28.6	1.1	1.1	7.4	1.6	1.6	4.8	1.1	1.1	3.2
COX2	0.95	0.93	0.96	37.4	37.2	36.9	17.4	17.3	17.6	8.9	8.9	9.1	4.5	4.5	4.6
DHFR	0.98	1.00	1.00	35.2	36.8	37.3	18.2	19.6	19.5	9.5	10.0	10.0	4.9	5.0	5.0
EGFR	0.97	0.97	1.00	34.7	34.7	34.5	18.2	18.5	19.4	9.5	9.3	9.9	4.8	4.8	5.0
ER_AGONIST	0.93	0.94	0.97	33.0	33.0	28.1	14.9	17.2	15.2	7.9	9.0	9.2	4.5	4.6	4.7
ER_ANTAGONIST	0.88	0.98	1.00	27.0	42.0	39.0	11.2	16.0	20.1	6.8	9.2	10.1	4.0	5.0	5.0
FGFR1	0.66	0.99	1.00	38.0	38.0	38.0	10.9	19.2	19.5	5.7	9.7	9.7	3.0	4.9	5.0
FXA	0.93	0.95	0.95	39.0	39.0	29.2	15.1	16.7	15.8	8.2	9.0	8.7	4.3	4.7	4.7
GART	0.92	0.94	0.95	0.0	21.9	21.9	6.7	10.8	9.4	6.0	6.0	8.1	5.0	5.0	5.0
GPB	0.91	0.98	0.97	41.2	41.2	41.2	16.2	18.9	17.5	8.5	9.8	9.6	4.4	4.9	4.9
GR	0.79	0.94	0.93	38.4	39.8	39.8	9.5	15.3	14.5	6.3	7.8	7.7	3.7	4.7	4.2
HIVPR	0.92	0.98	0.98	38.2	38.2	38.2	13.0	18.5	18.0	7.7	9.4	9.6	4.2	4.8	4.9
HIVRT	0.32	0.58	0.64	9.8	26.2	39.3	1.9	5.7	8.8	1.3	2.8	4.4	0.6	2.0	2.5
HMGA	0.93	0.91	0.97	45.3	34.0	41.5	17.3	14.4	15.8	9.0	8.3	8.6	4.5	4.1	4.7
HSP90	0.89	1.00	1.00	48.8	48.8	48.8	10.2	20.4	20.4	5.0	10.1	10.1	3.6	5.0	5.0
INHA	0.75	0.83	0.89	38.9	35.2	30.1	11.1	11.8	10.3	5.6	6.3	6.4	2.9	3.6	4.1
MR	0.97	0.96	0.96	40.7	54.3	54.3	15.1	17.6	15.1	8.8	8.8	8.8	4.4	4.4	4.4
NA	0.96	0.91	1.00	43.6	43.6	43.6	16.6	14.7	20.1	8.8	7.8	10.0	4.5	4.3	5.0
P38	0.89	0.99	0.99	30.1	32.9	34.5	15.1	18.0	18.8	8.0	9.4	9.6	4.1	4.9	5.0
PARP	0.96	0.97	1.00	43.0	43.0	43.0	14.3	15.1	19.4	8.2	8.6	10.0	4.5	4.7	5.0
PDE5	0.46	0.85	0.88	14.0	35.1	39.8	2.7	10.1	15.1	1.4	5.5	7.5	1.1	3.5	4.0
PDGFRB	0.69	0.86	0.91	31.7	37.7	37.7	7.1	14.8	16.4	4.2	7.5	8.4	2.4	3.8	4.3
PNP	0.98	1.00	1.00	42.2	47.4	47.4	17.9	20.0	20.0	9.0	10.0	10.0	4.7	5.0	5.0
PPAR_GAMMA	0.98	0.99	1.00	39.2	39.2	39.2	19.0	19.2	19.5	9.6	9.6	9.9	4.8	5.0	5.0
PR	0.91	0.98	1.00	45.0	45.0	45.0	15.6	18.3	20.2	7.8	9.6	10.1	4.6	4.8	5.0
RXRα	0.95	1.00	1.00	51.6	51.6	51.6	17.2	20.1	20.1	8.6	10.0	10.0	4.3	5.0	5.0
SAHH	0.99	0.99	0.99	27.0	34.7	42.4	18.0	18.7	18.7	9.3	9.3	9.7	5.0	5.0	5.0
SRC	0.71	0.90	0.93	37.2	39.2	39.2	12.0	15.3	17.2	6.1	7.7	8.7	3.2	4.1	4.5
THROMBIN	0.91	0.98	0.98	46.4	46.4	46.4	16.8	17.6	17.6	8.6	9.2	9.2	4.3	4.8	4.8
TK	0.88	0.94	0.94	11.8	11.8	11.8	10.6	10.6	9.4	6.5	7.7	7.7	3.8	4.4	4.7
TRYPSIN	0.93	0.97	1.00	42.4	42.4	42.4	17.4	18.0	19.6	8.9	9.2	10.0	4.6	4.6	5.0
VEGFR2	0.55	0.69	0.75	17.7	22.1	23.6	6.2	6.5	7.7	3.2	4.0	5.0	1.6	2.7	3.2
average	**0.82**	**0.90**	**0.94**	**34.4**	**36.1**	**36.7**	**12.5**	**14.8**	**15.6**	**6.8**	**8.1**	**8.6**	**3.7**	**4.3**	**4.5**

[a] Using multiple compounds for screening simultaneously may increase the performance of ROCS dramatically. This table reports the benchmark values obtained for screening using the center molecule and the three and five representative compounds of each actives set, respectively.

respect is whether or not ComboScore performs better than ShapeTanimoto or ColorScore. Therefore, we have investigated the performance of ROCS on all 40 DUD targets using the bioactive conformation. Using ShapeTanimoto as exclu-

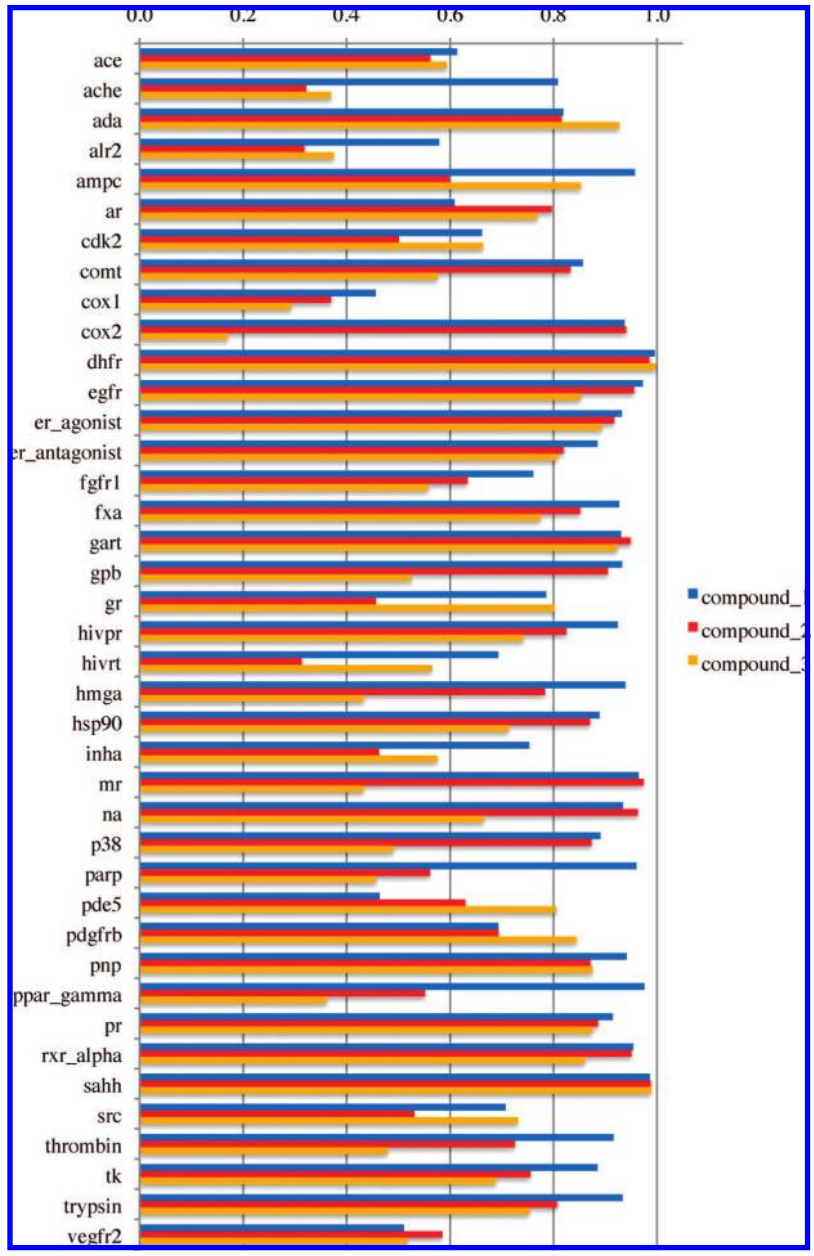

Figure 8. AUC values obtained by screening with three representative compounds (i.e., the center molecules obtained by clustering for maximum dissimilarity) of the actives. The individual compounds partially show widespread differences in terms of ROCS screening performance. Queries representing less prominent scaffolds of the actives set are likely to produce inferior hit lists than queries that are based on the most common actives scaffolds.

sive scoring functions results in a drop of the AUC by 12% compared to the ComboScore ranking (Figure 10). Also the EFs decrease: −2.6 for EF1% on average. Most affected by the lack of any kind of pharmacophoric ranking are TRYPSIN, THROMBIN, ALR2, FGFR1, and DHFR—with all but ALR2 and DHFR previously found to be challenging for virtual screening, and—even more interesting—with all but FGFR1 comprising ligands that form ionic interactions with the environment. There are only a few chemical groups that are positively or negatively ionizable, and, thus, chemical pattern matching on such rather infrequent function bears a decisive advantage during virtual screening. In this way, pharmacophoric filtering seems to play a key role particularly for successful screening of problematic targets.

After demonstrating that ShapeTanimoto scoring does not achieve the best virtual screening performance, we analyzed the importance of the contribution of the ColorScore in more detail. To our surprise, ColorScore performs at least as well as ComboScore, achieving an average AUC of 0.76. HIVRT represents the only test set obtaining considerably inferior results with ColorScore than with ShapeTanimoto; superior performance with ColorScore is achieved for ACE, ALR2, FGFR1, HSP90, PARP, SRC, and THROMBIN (Figure 10). These observations caused us to further investigate the impact of the ratio between ShapeTanimoto and ColorScore in ComboScore: we optimized ComboScore (the weighting of ShapeTanimoto and ColorScore) in order to obtain maximum AUCs considering weighting factors for each DUD target individually. We found that—on average—a ratio of 1:4 between ShapeTanimoto and ColorScore appears to be the optimum weighting. However, there is considerable variation among the DUD targets, including ratios of 7.5:−1 for HIVRT, 3.8:1 for RXRα, and 1:10 for AR. Using the individual optimum weightings for each target allows for

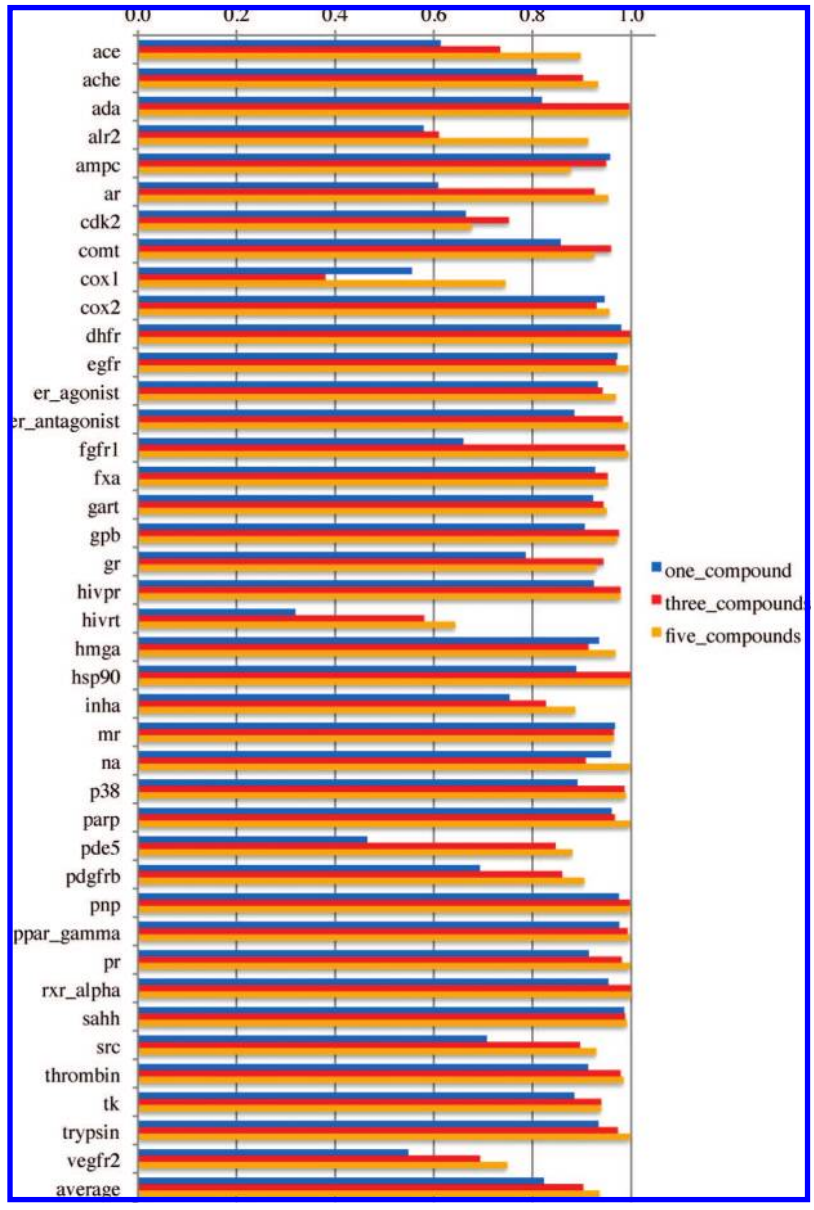

Figure 9. AUC values obtained with ROCS considering the highest scores obtained by any query compound for all compounds screened. The diagram illustrates the increase in virtual screening accuracy by using several representative compounds contemporarily for screening. Only in the case of three compounds used for screening with COX1 we find the exception, that score fusion obtains significantly inferior results compared to the single-compound run: The AUC of the three-compounds run is significantly smaller than the single-compound run. The reason for this performance bend seems to be that the single compound query is based on a highly popular scaffold, while the three-compounds query is based on one popular but two infrequent chemical scaffolds.

increasing the average AUC to 0.81. Instead, applying the optimum average weighting of 1:4 on the whole DUD database does not lead to significantly improved results: while 1:1 (ComboScore) obtains AUC 0.73, 1: 4 obtains a just 0.02 higher AUC. Optimized factors may therefore be able to increase the discriminatory power of ROCS, however, estimating these weights for a particular data set is difficult. We have calculated a collection of 1D, 2D, and 3D descriptors for each compound of the DUD, including chemical pattern counts, size, volume, shape, topological polar surface area, logP, etc., and analyzed these values for a possible correlation with the score weights. However, due to the numerous different influences within the individual test sets (discrepancies in the distances between actives and inactives, the number and kind of different interaction modes represented by the test sets, etc.) we were not able to elucidate characteristic parameters that allow for predicting the approximate factors for ShapeTanimoto and ScaledColor score. Nevertheless, a prospective screening campaign can benefit considerably from individually optimized score weights derived from screening experiments using known active and inactive compounds. Moreover, these results confirm the importance of pharmacophoric features for bioactivity and suggest that the contribution of pharmacophore-based scoring to the performance of virtual screening may have been underestimated so far. There is no doubt that the shape of ligands (complementary to the shape of the target interaction site) is of fundamental importance for bioactivity. However, focusing too tightly on shape properties that are derived from a single bioactive molecule is likely to lead to a considerable rate of false negatives, which is the major bottleneck of shape-based virtual screening approach. Scoring by shape causes compounds of dissimilar dimensions to be down-ranked during virtual screening. Such

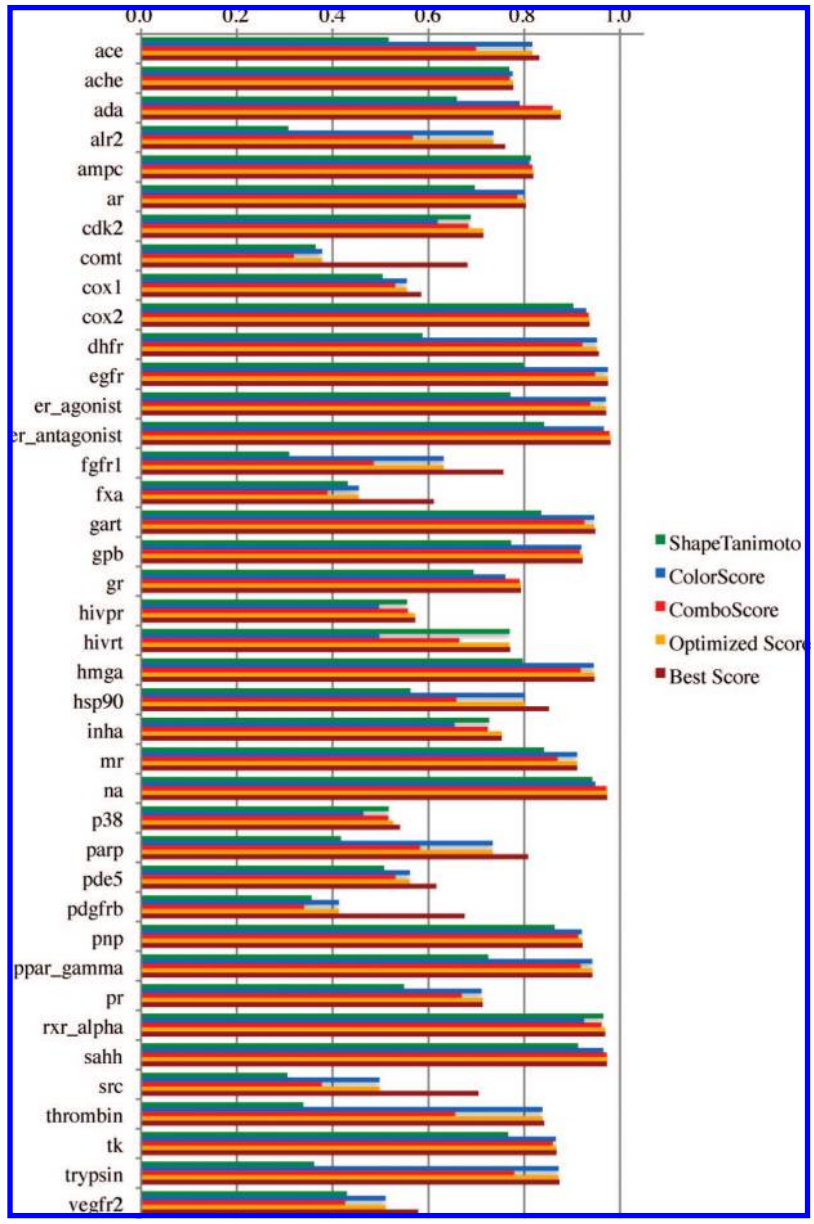

Figure 10. Impact of score weightings of ShapeTanimoto and ScaledColor score on the AUC. ShapeTanimoto and ColorScore represent the AUCs obtained by using these scoring components exclusively; ComboScore represents ROCS factory defaults; OptimizedScore illustrates the AUC values obtained for ShapeTanimoto and ScaledColor 1: 4; and BestScore illustrates the best-performing ratio of ShapeTanimoto and ScaledColor individually for each target.

compounds, however, may have high affinity to a target as they may form additional interactions with the target—provided that there are no clashes with the protein residues caused by the steric bulk. Nevertheless, the essential pharmacophoric features may be present, besides additional chemical features that may interact with the protein. A pharmacophore-based scoring function is able to discover such bioactive compounds, while a scoring function focusing on shape exclusively may down-rate it. Hence, it is important to find the optimum balance between shape-based and pharmacophore-based scoring functions in order to ensure maximum performance. Our results indicate that both parameters (ShapeTanimoto and ScaledColor score) may not be completely unrelated to each other, and it is plausible to expect a certain degree of overlap.

In the case of docking we see the opposite scenario of that observed with ROCS: docking suffers mostly from high false positives rates, as docking algorithms manage to place compounds in the protein−ligand interaction site (since the shape of the binding site—with bound ligand—is less restrictive than the shape of a single bioactive molecule) but fail to correctly estimate the interactions between the ligand and the protein (i.e., the pharmacophoric interactions). For both approaches the correct identification of the pharmacophoric interactions is an essential precondition for success in virtual screening.

CONCLUSIONS

We have investigated the performance of ROCS on the DUD, comprising actives and inactives sets for 40 pharmaceutically relevant targets. These data sets have been designed to show a reasonable degree of topological dissimilarity between actives and decoys and similar physicochemical properties, including basic descriptors also known by their importance in Lipinski's 'rule-of-five'. Therefore, these

compounds are considered a reasonable challenge for discrimination by virtual screening programs. The DUD has been designed as a benchmark set for molecular docking approaches, which—considering the design of the DUD—may favor docking programs, as they do have a better starting position with molecules that are—to a certain degree—diverse in terms of shape but comparable in terms of physicochemical, pharmacophoric properties. Docking is in the favorable position to have data present on the dimensions of the active site, while ligand-based screening is forced to render the properties of the active site based on known active compounds. Considering these circumstances, the performance of ROCS on the DUD targets can be stated good or very good and is not in an inferior position than docking programs.

In this study, we were able to show that the ColorScore is of particular importance for the reliable ranking of virtual compounds using ROCS. This shows that chemical or pharmacophoric information is of particular importance for shape-based screening, additionally to a shape description. On the other hand, we showed that optimized weighting factors for ComboScore do not lead to considerably enhanced screening performance unless parametrized for each data set individually, which points to the conclusion that the ColorScore implementation still bears room for improvement. Another important insight is that using multiple well-embedded actives as query molecules can lead to dramatic increases of virtual screening performance.

The selection of the query conformation proved to be less important, rendering shape-based screening suitable for ligand-based modeling: The availability of a bioactive conformation for the query seems not to be the limiting factor for screening—it is more the selection of query compound(s) that is decisive for screening performance.

Abbreviations. ROCS, rapid overlay of chemical structures; USR, ultrafast shape recognition; CXCR4, CXC chemokine receptor 4; CCR5, chemokine (C−C motif) receptor 5; DUD, directory of useful decoys; EF, enrichment factor; ROC, receiver operating characteristic; AUC, area under the curve; HTS, high throughput screening; Se, sensitivity; Sp, specificity; ACE, angiotensin-converting enzyme; AChE, acetylcholinesterase; ADA, adenosine deaminase; ALR2, aldose reductase; AmpC, AmpC β-lactamase; AR, androgen receptor; CDK2, cyclindependent kinase 2; COMT, catechol O-methyltransferase; COX-1, cyclooxygenase-1; COX-2, cyclooxygenase-2; DHFR, dihydrofolate reductase; EGFr, epidermal growth factor receptor; ER, estrogen receptor; FGFr1, fibroblast growth factor receptor kinase; FXa, factor Xa; GART, glycinamide ribonucleotide transformylase; GPB, glycogen phosphorylase β; GR, glucocorticoid receptor; HIVPR, HIV protease; HIVRT, HIV reverse transcriptase; HMGR, hydroxymethylglutaryl-CoA reductase; HSP90, human heat shock protein 90; InhA, enoyl ACP reductase; MR, mineralocorticoid receptor; NA, neuraminidase; P38 MAP, P38 mitogen activated protein; PARP, poly(ADP-ribose) polymerase; PDE5, phosphodiesterase 5; PDGFrb, platelet derived growth factor receptor kinase; PNP, purine nucleoside phosphorylase; PPARg, peroxisome proliferator activated receptor γ; PR, progesterone receptor; RXRa, retinoid X receptor α; SAHH, S-adenosyl-homocysteine hydrolase; SRC, tyrosine kinase SRC; TK, thymidine kinase; VEGFr2, vascular endothelial growth factor receptor 2; ATP, adenosine-5′-triphosphate; β-GAR, β-glycinamide ribonucleotide; NAD(P)-(H), nicotinamide adenine dinucleotide (phosphate)-(reduced); PLP, pyridoxal-5′-phosphate.

ACKNOWLEDGMENT

We thank OpenEye for providing us an academic license for the OpenEye software package.

Supporting Information Available: Performance benchmarks for ROCS on all 40 targets using different query conformations. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Grant, J. A.; Gallard, M. A.; Pickup, B. T. A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653–1666.
(2) Nicholls, A.; Grant, J. A. Molecular shape and electrostatics in the encoding of relevant chemical information. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 661–686.
(3) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.
(4) *Catalyst, 4.11*; Accelrys: San Diego, CA, 2005.
(5) Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P. A.; Lee, W.-C.; Pontz, T.; Corbley, M. J.; Cheung, H. K.; Arduini, R. M.; Mead, J. N.; Newman, M. N.; Papadatos, J. L.; Bowes, S.; Josiah, S.; Ling, L. E. Successful shape-based virtual screening: the discovery of a potent inhibitor of the type I TGFβ receptor kinase (TβRI). *Bioorg. Med. Chem. Lett.* **2003**, *13*, 4355–4359.
(6) *Phase, 3.0207*; Schrödinger, L. L. C.: New York, 2008.
(7) Livingstone, D. J.; Clark, T.; Ford, M. G.; Hudson, B. D.; Whitley, D. C. QSAR studies using the parashift system. *SAR QSAR Environ. Res.* **2008**, *19*, 285–302.
(8) Ritchie, D. W.; Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins: Struct., Funct., Bioinf., Genet.* **2000**, *39*, 178–194.
(9) Perez-Nueno, V. I.; Violet, A. I.; Ritchie, D. W.; Borrell, J. I.; Teixido, J. Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape-based virtual screening approach: further evidence for multiple binding sites within the CCR5 extracellular pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.
(10) Perez-Nueno, V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixido, J. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. *J. Chem. Inf. Model.* **2008**, *48*, 509–533.
(11) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: molecular alignment based on shape and electrostatics In *17th European Symposium on QSAR in "omics" and Systems biology*; Uppsala: Sweden, 2008.
(12) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723.
(13) Putta, S.; Beroza, P. Shapes of things: computer modeling of molecular shape in drug discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1514–1524.
(14) Bostrom, J.; Berggren, K.; Elebring, T.; Greasley Peter, J.; Wilstermann, M. Scaffold hopping, synthesis and structure-activity relationships of 5,6-diaryl-pyrazine-2-amide derivatives: a novel series of CB1 receptor antagonists. *Bioorg. Med. Chem.* **2007**, *15*, 4077–84.
(15) Freitas, R. F.; Oprea, T. I.; Montanari, C. A. Two-dimensional QSAR and similarity studies on cruzain inhibitors aimed at improving selectivity over cathepsin L. *Bioorg. Med. Chem.* **2008**, *16*, 838–853.
(16) Bologa, C. G.; Revankar, C. M.; Young, S. M.; Edwards, B. S.; Arterburn, J. B.; Kiselyov, A. S.; Parker, M. A.; Tkachenko, S. E.; Savchuck, N. P.; Sklar, L. A.; Oprea, T. I.; Prossnitz, E. R. Virtual and biomolecular screening converge on a selective agonist for GPR30. *Nat. Chem. Biol.* **2006**, *2*, 207–212.
(17) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941–948.
(18) Venhorst, J.; Nunez, S.; Terpstra, J. W.; Kruse, C. G. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *J. Med. Chem.* **2008**, *51*, 3222–3229.
(19) *GlideXp, 5.0207*; Schrodinger, L. L. C.: New York, 2008.

(20) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 235–249.
(21) 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.
(22) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.
(23) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293–2302.
(24) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
(25) Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **2007**, *47*, 2182–2196.
(26) Sykes, M. J.; Sorich, M. J.; Miners, J. O. Molecular modeling approaches for the prediction of the nonspecific binding of drugs to hepatic microsomes. *J. Chem. Inf. Model.* **2006**, *46*, 2661–2673.
(27) Wolber, G.; Dornhofer, A.; Langer, T. Efficient overlay of small molecules using 3-D pharmacophores. *J. Comput-Aided. Mol. Des.* **2006**, *20*, 773–788.
(28) Wolber, G.; Langer, T. LigandScout: 3D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160–169.
(29) Ebalunode, J. O.; Ouyang, Z.; Liang, J.; Zheng, W. Novel approach to structure-based pharmacophore search using computational geometry and shape matching techniques. *J. Chem. Inf. Model.* **2008**, *48*, 889–901.
(30) Markt, P.; Petersen, R. K.; Flindt, E. N.; Kristiansen, K.; Kirchmair, J.; Spitzer, G.; Distinto, S.; Schuster, D.; Wolber, G.; Laggner, C.; Langer, T. Discovery of novel PPAR ligands by a virtual screening approach based on pharmacophore modeling, 3D shape, and electrostatic similarity screening. *J. Med. Chem.* **2008**, *51*, 6303–6317.
(31) Lee, H. S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.; Yoon, S. Optimization of high throughput virtual screening by combining shape-matching and docking methods. *J. Chem. Inf. Model.* **2008**, *48*, 489–497.
(32) *OMEGA, 2.3.2*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2008.
(33) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.
(34) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(35) *ROCS, 2.3.1*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2007.
(36) Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G.; Langer, T.; Wolber, G. Enhancing drug discovery through in silico screening: strategies to increase true positives retrieval rates. *Curr. Med. Chem.* **2008**, *15*, 2040–2053.
(37) Nicholls, A. What do we know and when do we know it. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
(38) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron, Comp. Method.* **1990**, *3*, 537–47.
(39) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–8.
(40) Sadowski, J.; Rudolph, C.; Gasteiger, J. The generation of 3D models of host-guest complexes. *Anal. Chim. Acta* **1992**, *265*, 233–41.
(41) *Sybyl, 8.1*; Tripos: St. Louis, MO, 2008.
(42) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative performance assessment of the conformational model generators Omega and Catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848–1861.
(43) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.
(44) Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46*, 5781–5789.
(45) Hecker, E. A.; Duraiswami, C.; Andrea, T. A.; Diller, D. J. Use of Catalyst pharmacophore models for screening of large combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1204–1211.
(46) Diller, D. J.; Li, R. Kinases, homology models, and high throughput docking. *J. Med. Chem.* **2003**, *46*, 4638–4647.
(47) Truchon, J.-F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
(48) Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. *J. Chem. Inf. Model.* **2008**, *48*, 719–729.
(49) 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.
(50) Powell, M. J. D. Restart procedures for the conjugate gradient method. *Math. Prog.* **1977**, *12*, 241–254.
(51) 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.
(52) 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.
(53) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1–16.
(54) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(55) Steindl, T.; Schuster, D.; Wolber, G.; Laggner, C.; Langer, T. High-throughput structure-based pharmacophore modelling as a basis for successful parallel virtual screening. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 703–715.
(56) Steindl, T. M.; Schuster, D.; Laggner, C.; Chuang, K.; Hoffmann, R. D.; Langer, T. Parallel screening and activity profiling with HIV protease inhibitor pharmacophore models. *J. Chem. Inf. Model.* **2007**, *47*, 563–571.

CI8004226