JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

PLS-DA - Docking Optimized Combined Energetic Terms (PLSDA-DOCET) Protocol: A Brief Evaluation

Sorin Avram,[†] Liliana M. Pacureanu,[†] Edward Seclaman,[‡] Alina Bora,[†] and Ludovic Kurunczi*[,§]

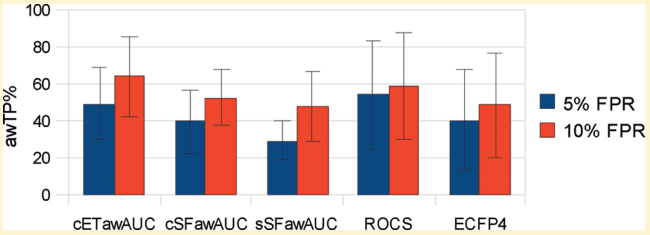
[†]Department of Computational Chemistry, Institute of Chemistry of Romanian Academy, Timisoara, Mihai Viteazul Avenue, 24, 300223 Timisoara, Romania

[‡]Department of Biochemistry, Faculty of Medicine, University of Medicine and Pharmacy "Victor Babes", Eftimie Murgu 2, Timisoara 300041, Romania

[§]Department of Chemical Physics, Faculty of Pharmacy, University of Medicine and Pharmacy "Victor Babes", Eftimie Murgu 2, Timisoara 300041, Romania

Supporting Information

ABSTRACT: Docking studies have become popular approaches in drug design, where the binding energy of the ligand in the active site of the protein is estimated by a scoring function. Many promising techniques were developed to enhance the performance of scoring functions including the fusion of multiple scoring functions outcomes into a so-called consensus scoring function. Hereby, we evaluated the target oriented consensus technique using the energetic terms of several scoring functions. The approach was denoted PLSDA-DOCET. Optimization strategies for consensus energetic terms and scoring functions based on ROC metric were compared to classical rigid docking and to ligand-based similarity search methods comprising 2D fingerprints and ROCS. The ROCS results indicate large performance variations depending on the biological target. The AUC-based strategy of PLSDA-DOCET outperformed the other docking approaches regarding simple retrieval and scaffold-hopping. The superior performance of PLSDA-DOCET protocol relative to single and combined scoring functions was validated on an external test set. We found a relative low mean correlation of the ranks of the chemotypes retrieved by the PLSDA-DOCET protocol and all the other methods employed here.



INTRODUCTION

In the field of drug discovery, virtual screening (VS) comprises ligand-based and structure-based methods for the retrieval of small-molecules from large databases. The applicability of these approaches in any particular circumstance depends upon the type of data available. Ligand-based methods presume the knowledge of an active molecule in order to retrieve similar molecules. A vast part of these methods are based on the *Similar Property Principle*.[1] If the crystal structure of the biological target is available, docking studies can be carried out to find molecules interacting with the binding site of the protein.[2] Nowadays, over 70,000 proteins and nucleic acids are available in RCSB Protein Data Bank,[3] providing large resources for structure-based virtual screening experiments such as protein−ligand docking. The scope of the VS is the identification of novel series for a given target. The retrospective studies are mainly used to analyze and calibrate novel scoring functions.

Protein−ligand docking searches the conformational space of the ligand, while the binding site residues are treated either rigid (rigid docking)[4] or flexible (searching the conformation of the side-chains in the binding site).[4,5] Thus, in rigid docking, the program employs algorithms to find the best pose (orientation and conformation) of the ligand and to compute the binding energy of the complex.[4]

Empirical scoring functions perform fast searches in large databases and are widely used in protein−ligand docking programs. The key concept of these functions is the assumption that the overall Gibbs free energy of binding is divided into singular contributions including hydrogen bond interactions, ionic interactions, hydrophobic interactions, and the loss of entropy as a function of the number of rotatable bonds.[6] Thus, the empirical scoring functions (SFs) contain independent energetic terms (ETs) accounting for more or less specific interaction types, and the sum of these terms estimate the binding energy between the ligand and the protein.[7] However, SFs differ in their ability to reflect certain interactions. The scoring function terms are weighted by fitting the SF to experimental data. The development of empirical SFs is based on various training sets and consequently even if two functions share the same type of ETs they provide various results, due to different procedures and data sets used for the calibration of the components.

The combination of multiple docking SFs into a consensus scoring function, using different fusing rules, was introduced by Charifson et al.[8] Many promising consensus techniques were

Received: May 22, 2011
Published: November 08, 2011

developed, during the past years, including machine learning algorithms to select and/or weight SFs.[9−13] For example, Teramoto and Fukunishi[10] tested a feature selection-based rescoring method that selected a combination of SFs that enhanced the enrichment and outperformed the individual SFs.

Based on this assumption Stahl and Rarey[7] derived ScreenScore, a scoring function combining the hydrogen-bond term and the lipophilic interaction term from two parent SFs. For the protein targets used in their evaluation, the new function was found to exhibit superior retrieval compared to the original SFs.[7] Teramoto and Fukunishi[10] suggested that a feature selection-based consensus scoring would be applicable also in the selection of terms of SFs providing a target-specific SF.

In this paper we employ a rescoring docking-protocol that identifies an optimal target- and objective-specific combination of ETs derived from several parent empirical SFs. We applied the Partial-Least-Square Discriminant-Analysis[14,15] (PLS-DA) approach to select the essential ETs which best discriminate the active molecules (actives) from the decoys for each target. We called the approach PLSDA-Docking Optimized Combined Energetic Terms (PLSDA-DOCET). The evaluation of three ROC-based strategies was conducted, in order to optimize the combination of ETs and SFs, and the results were compared to single SFs, 2D fingerprint-based, and 3D similarity search (SS) methods. The performance of the optimized DUD docking-models was assessed on an external validation set. Finally, we discuss the relative order of the ranked actives.

MATERIALS AND METHODS

Data Sets. The directory of useful decoys (DUD)[16] is a 40 target large database comprising 2950 ligands. For each active molecule 36 *drug-like* decoys were selected by matching physical but not topological properties. The DUD was specifically designed to address the weaknesses of docking methods[16] providing, in this sense, a platform to find better docking algorithms for VS applications. In order to solve the drawbacks like analogue bias and VS compatibility data set, Good and Oprea[17] created a clustered version of the active ligands in DUD. They filtered the ligand sets using the lead-like criteria of Oprea et al.[18] (molecular weight <450 and AlogP < 4.5 or 5.5 for nuclear hormone receptor targets) and grouped the ligands roughly according to chemotype using reduced graphs.[19]

In this study subsets of five protein targets were downloaded from the DUD Web-page (http://dud.docking.org/r2/ - accessed Jan 25, 2011): aldose reductase (ALR2 − an NADPH-dependent oxidoreductase that catalyzes the reduction of a variety of aldehydes and carbonyls, including monosaccharides); cyclin-dependent kinase 2 (CDK2 − a member of the cyclin-dependent kinase family of Ser/Thr protein kinases); cyclooxygenase 2 (COX2 − an enzyme responsible for formation of important biological mediators called prostanoids); epidermal growth factor receptor (EGFr − a cell-surface receptor for members of the epidermal growth factor family of extracellular protein ligands); and estrogen receptor alpha for agonists (ERagonist − a nuclear receptor activated by estrogens). The selected proteins correspond to various types of binding sites in terms of protein chain flexibility and intermolecular interactions. The size of the data sets ranges broadly from 26 actives (ALR2) to 365 (EGFr). All five proteins represent potential targets for cancer therapy:

ALR2: the inhibition of this protein is known to prevent human colon cancer cell growth;[20]

CDK2: it is possible to selectively interrupt the cell cycle regulation in breast cancer cells by interfering with cyclin-dependent kinase action;[21]

COX2: appears to be related to cancers in the gastrointestinal tract; COX inhibitors have been shown to reduce the occurrence of cancers and precancerous growths;[22]

EGFr: new drugs such as IRESSA and Tarceva directly target the epidermal growth factor receptor in lung cancer patients;[23]

$ER_{agonist}$: may have a tumor suppressor role in the prostate gland and loss of its expression may be an early event in prostatic disease.[24]

Table 1. DUD Data Sets Used in This Study

target	no. of actives	no. of decoys	no. of clusters	PDB code	Tanimoto[f]
ALR2[a]	26	918	14	1ah3	92.00
CDK2[b]	47	1779	32	1ckp	95.75
COX2[c]	212	12442	44	1cx2	67.92
EGFr[d]	365	14894	40	1m17	89.32
$ER_{agonist}$[e]	63	2355	10	1l2i	73.02

[a] Aldose reductase. [b] Cyclin-dependent kinase 2. [c] Cyclooxygenase 2. [d] Epidermal growth factor receptor. [e] Estrogen receptor for agonists. [f] Diversity of the ligand set and decoys: the percentage of actives for which the Tanimoto similarity to the nearest neighbor in the decoy set is less than or equal to 0.85 (see last paragraph Overall discriminative power section).

The actives filtered and clustered by Good and Oprea,[17] including also their ionization states and tautomers were used for 3D SS and for docking. In the case of 2D SS the actives and decoys were reduced to unique compounds. Table 1 shows for each subset the protein-target, the number of actives, clusters, and decoys used in this study. Prior to docking and 3D SS, the database was processed for generating conformers (OMEGA software).[25]

Scoring Functions and Docking. Empirical scoring functions were first developed by Böhm in the mid 1990s.[6] Based on this idea a large number of empirical SFs were developed.

We used the docking software FRED[26] to perform rigid protein−ligand docking on the pregenerated multiconformer database of molecules. Seven fast empirical SFs were employed to score the docking poses: Shapegauss,[27] PLP (Piecewise Linear Potential),[28] Chemgauss2,[26] Chemgauss3,[26] Chemscore,[27,29] OEChemScore,[26] and ScreenScore.[7] Only the best FRED consensus scored tautomer and/or ionization state was retained for each molecule for consensus scoring and evaluation.

PLSDA-DOCET Protocol. We describe PLSDA-DOCET as a two-steps method. In the first step we merged the individual ET values of the seven scoring functions (available in FRED) and applied Partial-Least-Square Discriminant-Analysis (PLS-DA) (available in SIMCA P 9.0 package)[30] in order to reduce the data and find the ETs that best discriminate the actives from the decoys.

The PLS-DA procedure was applied as follows: (a) for a given model the appropriate number of latent components was determined using a seven round cross-validation method, retaining only the components with Q^2 (the fraction of the Y's that can be predicted by the current component) greater than 0; (b) the overfitting in the final model was removed by variable selections for each consecutively constructed model for a target protein, retaining only the variables with VIP (variable influence on projection)[31] parameter statistically greater than 1, and PLS

coefficient values significantly different (t-Student test) from zero; (c) the statistical significance of the estimated predictive power and the absence of data overfit for the final model was certified by Y-permutation,[31] using 999 randomizations.

In order to identify the most successful combination, in the second step of the protocol, we adopted three strategies: we looked for the combination displaying the best early enrichment, for the best overall discriminative power and for both characteristics simultaneously. In the next section we will describe the metrics used in this study to approach the second step of the PLSDA-DOCET protocol and also to select the combinations of SFs.

Choosing Combinations of ETs (cET) and SFs (cSF). A largely used metric for the evaluation of prediction methods is the receiver-operator-characteristic (ROC) curve.[32] Details about the superior properties of ROC-based metrics relative to other evaluation metrics are highlighted by Jain.[33] The ROC curve plots the true positives rates (or *sensitivity* on the *Y*-axis) relative to the false positive rates (or *1-specificity* on the *X*-axis). However, the ROC plot relies on visual examination and is unsuited for our task because of the large number of combinations generated in the case of ETs and SFs. Jain and Nicholls[34] suggested to report the performance of VS methods in terms of ROC Enrichment denoting the ratio of true positive rates (TPR) to the following values of false positive rates (FPR): 0.5%, 1%, 2%, and 5%, and the area under the receiver operating curve[35] (ROC AUC further denoted simply AUC). The first parameter measures the early recognition capacity of the evaluated method and is proportional to the percentage of the true positives (TPs) found at the above-mentioned FPRs. In this study we chose the cET and cSF components based on ROC-derived metrics, and this allowed us to use the *arithmetic weighting* scheme (aw)[36] recommended by Mackey and Melville[37] in order to account for scaffold-hopping.

In the attempt to identify the combination exhibiting the best early enrichment we simplified the task and focused mainly on the very top of the ranking list. We found the addition of early TPR/FPR ratios helpful: the later a TP is found in the ranked list, the lesser is its contribution to the sum (the corresponding FPR value being higher). In this study we selected discontinued FPR values thereby creating multigrade relevance intervals for the TPs (e.g., the TPs found before the 5% FPR are more relevant in VS compared to the TPs found between the 5% and 10% FPR). The TPs of a relevance interval (between two successive FPRs) were weighted equally and separately from those contained in a different interval. Because the terms are proportional to the TPR/FPR ratios we shall call the sum AROCE, reflecting the Addition of ROC Enrichment

$$\mathrm{AROCE} = \frac{\sum_{i=1}^{n} \frac{TP_i - TP_{i-1}}{A} \frac{1}{p_i}}{\frac{1}{p_1}} \quad (1)$$

where $p_1 > 0$ and $p_1 < p_2 < p_3 \ldots < p_n$ are percentages of false positives (FPR·100) found along the ranked list. The total number of true positives is denoted by A, TP_i represents the number of true positives at p_i, and TP_0 is zero. The TPs found between p_{i-1} and p_i are weighted with $1/p_i$. The presence of $1/p_1$ term in the denominator assures AROCE takes values only between zero and one.

We chose AROCE with the following values for p_i: 0.5, 1, 2, 5, and 10, as a criterion to identify combinations of ETs and SFs exhibiting the best enrichment in the very top of the database (cET_{AROCE} and cSF_{AROCE}). The range between the first two percentages is very small (compared to the proportionally increase of the other percentages), and the TPs found until 0.5% FPR are the most heavily weighted.

The AUC parameter indicates the probability that a randomly selected active is higher scored by a method than a randomly selected decoy. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test (random selection). The AUC measures the performance throughout the entire data set and is insensitive to the enrichment in the top percentage. Therefore, we used AUC as a second criterion for the selection of ETs (cET_{AUC}) and SFs (cSF_{AUC}).

To account for both early recognition and overall discriminative power, we weighted AROCE with the AUC value. Thus, we shall consider the mean of the parameters (AROCE+AUC)/2 and designate this as AROCE50 (the AROCE has a 50% contribution). We denoted the most successful AROCE50 combinations as $cET_{AROCE50}$ and $cSF_{AROCE50}$.

The cET and cSF were identified also based on the *arithmetic-weighted* form of the ROC metric described above: awAROCE ($cET_{awAROCE}$ and $cSF_{awAROCE}$), awAUC (cET_{awAUC} and cSF_{awAUC}), and awAROCE50 ($cET_{awAROCE50}$ and $cSF_{awAROCE50}$). Equations 2 and 3 describe awAROCE and awAUC, while awAROCE50 is the mean of the two parameters. If the actives are clustered relative to the same chemotype, the effect of this weighting scheme can be described as follows: every k^{th} active will contribute to the total number of actives found in the ranked list with a weight of $w_{jk} = 1/N_j$, where N_j is the size of the j^{th} cluster containing the k^{th} molecule. If the number of clusters equals the number of structures w_{jk} equals one

$$\mathrm{awAROCE} = \frac{\sum_{i=1}^{n} \left(\frac{1}{N_{Clusters}} \sum_{j}^{N_{Clusters}} \sum_{k}^{N_j} w_{jk} \alpha_{jk}^{TP_i - TP_{i-1}} \right) \frac{1}{p_i}}{\frac{1}{p_1}} \quad (2)$$

with $p_i = \{0.5, 1, 2, 5, 10\}$, and

$$\mathrm{awAUC} = 1 - \frac{1}{N_{Clusters}} \sum_{j}^{N_{Clusters}} \sum_{k}^{N_j} w_{jk} \frac{N_{decoys}^{jk}}{N_{total\_decoys}} \quad (3)$$

where $N_{Clusters}$ and N_{total_decoys} represent the total number of clusters and decoys, N_{decoys}^{jk} represents the number of decoys ranked above the k^{th} active contained in the j^{th} cluster; $\alpha_{jk}^{TP_i - TP_{i-1}}$ is 1 if the k^{th} structure of the j^{th} cluster is found between TP_i and TP_{i-1}, otherwise $\alpha_{jk}^{TP_i - TP_{i-1}}$ is 0. Equation 3 was already used by Jahn et al.[38] in their paper from 2009.

Ligand-Based Methods. 2D SS methods provide a fast approach to search for molecules (resembling the query) throughout very large databases. Molecular fingerprints describe the substructural features of the molecule, assigning 1 to the bits of each present feature and 0 to the absent ones. In this study, the molecules were described by the 1024bit SciTegic ECFP4 (Extended Connectivity Fingerprint) and FCFP4 (Functional Connectivity Fingerprint) fingerprints. For every target the native ligand was used as the query, and the Tanimoto similarity coefficient was computed against the DUD target-subset. Descriptors

generation and coefficient calculations were carried out using PipeLine Pilot Student Edition.[39]

3D SS was performed using ROCS[40] (Rapid Overlay of Chemical Structures). The program requires a database of conformers and a query molecule. Each conformer is overlaid rigidly on the query molecule in order to maximize the shape and chemical overlap.[41] The similarity is further assessed by various scoring functions. We used the cocrystallized conformation of the bound ligand as the query to perform 3D ligand-based similarity search. The ranking was performed according to the ComboScore function since it accounts for both shape and functional group overlay.

External Validation. The reliability of the docking-based protocols was tested using an external test set. The ChEMBLdb database[42] was explored in order to retrieve bioactivity data for the five protein-targets used in this study. The IC50 values (ED50 for $ER_{agonist}$) of several ChEMBL IDs were downloaded (the details are specified in the Supporting Information). The human related isoforms were assured to exhibit a minimum sequence identity of 85% (according to blast/uniprot).[43,44] The data were filtered to keep only expert curated data expressed in nM. For molecules with more than one activity entry the mean value was computed. Only molecules with activity values less than 100 nM were considered active on the targets.

The corresponding decoy sets were generated using a similar protocol to that employed to create DUD. The drug-like subset from ZINC[45] containing 1.67 million compounds, at neutral pH, was downloaded. Duplicate molecules were removed after standardization using PipeLinePilot Student Edition components. The SciTegic MDL Public Keys[46] fingerprints were computed, and the compounds with a Tanimoto coefficient higher than 0.9 (relative to the actives extracted from ChEMBLdb) were removed from the ZINC subset. Further, the decoy sets for each protein were created by performing an additional similarity analysis based on the following drug-like descriptors: molecular weight, number of hydrogen bond acceptors, number of hydrogen bond donors, number of rotatable bonds, and ALogP. The number of decoys have been selected such as the decoys/active ratio to be higher than 36 (specific to DUD). For each target, the new sets were filtered against the corresponding DUD ligands and decoys. The size of the external validation sets are shown in Table 2.

Table 2. Number of Actives and Decoys Comprised in the External Validation Sets for Each Target

target[a]	no. of external actives	no. of external decoys	decoys/actives ratio
ALR2	40	6294	157.4
CDK2	49	12428	229.7
COX2	48	7349	320
EGFr	32	10232	253.7
$ER_{agonist}$	30	1348	45.1

[a] See Table 1.

RESULTS AND DISCUSSION

Before discussing the PLSDA-DOCET protocol performance a test for the predictive validation and overfit of the PLS-DA procedure was performed. For this purpose we applied the response permutation method implemented in SIMCA package,[30] as mentioned also in the PLSDA-DOCET Protocol section. It requires the analysis of the plot showing on the *Y*-axis the R^2Y_{cum} (cumulative sum of squares of all the Y's explained by all extracted components) and the Q^2_{cum} (cumulative Q^2 for all the Y's for the extracted components) values of all PLS models (the Y permuted models and also the "tested" one) and on the *X*-axis the correlation coefficients between permuted and original response variables. One regression line is fitted among the R^2Y_{cum} points, and another line is fitted among the Q^2_{cum} points. If the *Y*-axis intercepts of the regression lines do not exceed 0.3−0.4 for R^2Y_{cum}, and 0.05 for Q^2_{cum} the model is considered valid.[31] The values obtained after 999 permutations for the five targets implied in this work were between 0.0213 ÷ 0.0003 for the R^2Y_{cum} line and between 0.0180 ÷ −0.0112 for the Q^2_{cum} line (detailed results in the Supporting Information). Thus, the model complexity is adequate, and the obtained PLS-DA models were not overfitted.

Table 3. DUD Model and External Validation Results

	model (DUD set)						external validation					
	AROCE		AUC		AROCE50		AROCE		AUC		AROCE50	
case[a]	a5[b]	a4[c]	a5[b]	a4[c]	a5[b]	a4[c]	a5[b]	a4[c]	a5[b]	a4[c]	a5[b]	a4[c]
sSF_{AROCE}	0.27	0.26	0.70	0.74	0.48	0.49	0.13	0.16	0.53	0.59	0.33	0.37
sSF_{AUC}	0.17	0.18	0.80	0.83	0.48	0.50	0.17	0.22	0.72	0.80	0.45	0.51
$sSF_{AROCE50}$	0.24	0.23	0.76	0.80	0.50	0.51	0.14	0.17	0.64	0.72	0.38	0.44
cSF_{AROCE}	0.29	0.29	0.73	0.76	0.51	0.53	0.08	0.10	0.53	0.60	0.31	0.35
cSF_{AUC}	0.22	0.22	0.79	0.82	0.50	0.52	0.17	0.22	0.68	0.77	0.43	0.49
$cSF_{AROCE50}$	0.27	0.26	0.76	0.81	0.52	0.53	0.09	0.11	0.63	0.73	0.36	0.42
cET_{AROCE}	0.33	0.36	0.85	0.88	0.59	0.62	0.22	0.27	0.76	0.88	0.49	0.58
cET_{AUC}	0.29	0.33	0.87	0.90	0.58	0.61	0.22	0.27	0.78	0.88	0.50	0.58
$cET_{AROCE50}$	0.32	0.32	0.86	0.90	0.59	0.62	0.24	0.30	0.77	0.89	0.51	0.60

[a] sSF_{AROCE} − best AROCE performing single scoring function (SF); sSF_{AUC} − best AUC performing single scoring function; $sSF_{AROCE50}$ − best AROCE50 performing single scoring function; cSF_{AROCE} − best AROCE performing combination of SFs; sSF_{AUC} − best AUC performing combination of SFs; $cSF_{AROCE50}$ − best AROCE50 performing combination of SFs; cET_{AROCE} − best AROCE performing combination of ETs (energetic terms from those chosen by the PLS-DA analysis); cET_{AUC} − best AUC performing combination of ETs; $cET_{AROCE50}$ − best AROCE50 performing combination of ETs. The values represent the averages over the five (four) targets. [b] Average values for the five targets. [c] Average over COX2, EGFr, CDK2, $Er_{agonist}$ (see text).

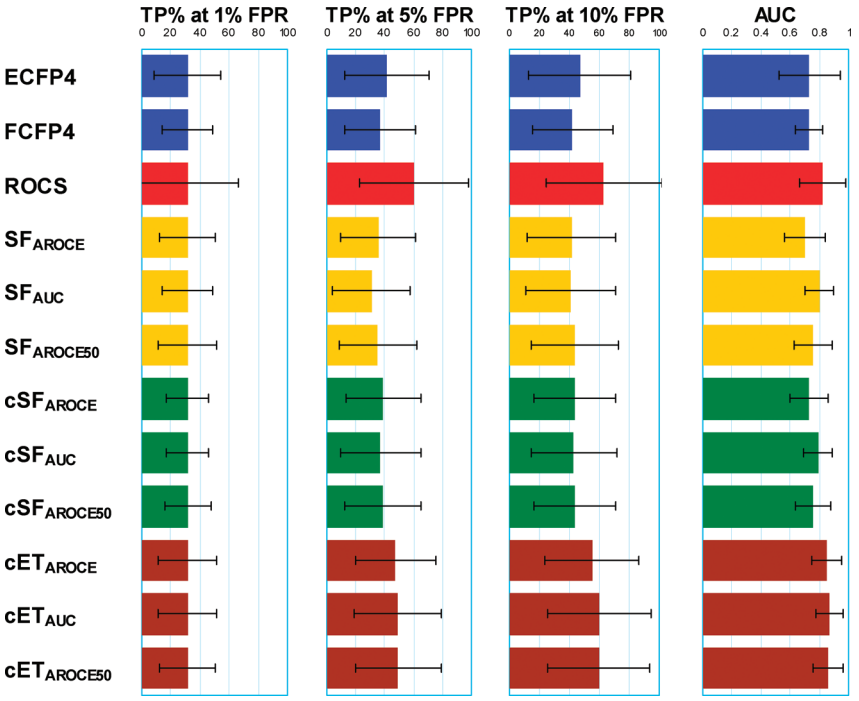


Figure 1. Average (and the corresponding standard deviation bars) TP% (targets presented in Table 1) at 1%, 5%, and 10% FPR and AUC for ligand-based 2D fingerprint (ECFP4 and FCFP4; blue) and 3D ROCS similarity search (red) methods and docking-based best single scoring function (sSF; gold) and rescoring protocols (cSF and cET; green and brown) considering three selection criteria: AROCE, AUC, and AROCE50.

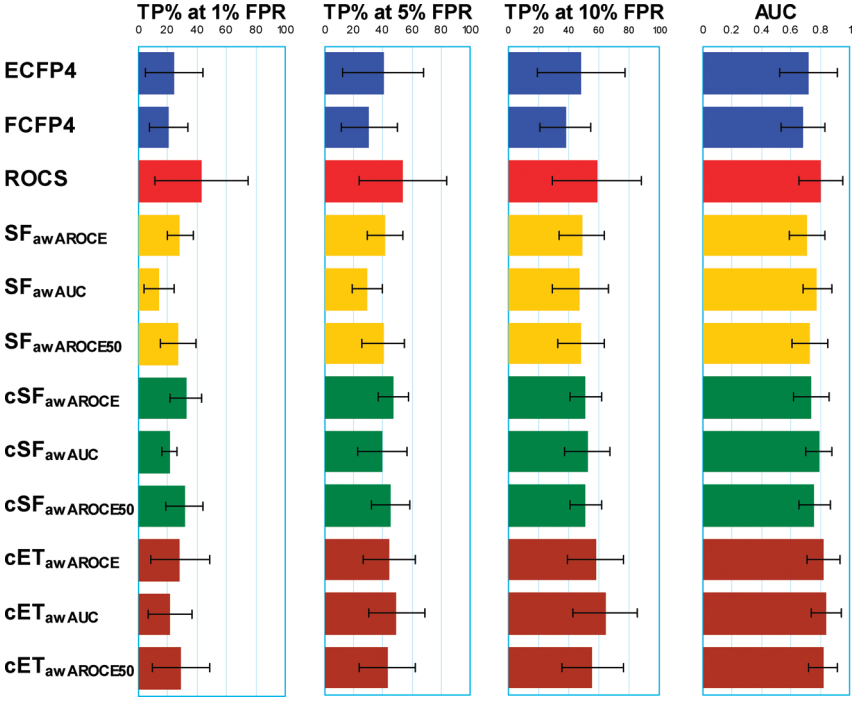


Figure 2. Average (and the corresponding standard deviation bars) awTP% (targets presented in Table 1) at 1%, 5%, and 10% FPR and awAUC (the arithmetic weighting scheme — aw — accounts for the retrieval of chemotypes) for ligand-based 2D fingerprint (ECFP4 and FCFP4; blue) and 3D ROCS similarity search (red) methods and docking-based best single scoring function (sSF; gold) and rescoring protocols (cSF and cET; green and brown) considering three selection criteria: awAROCE, awAUC, and awAROCE50.

In this study we evaluated the PLSDA-DOCET protocol using three selection criteria for simple retrieval and their homologue version reflecting the scaffold-hopping ability (arithmetic-weighting scheme): AROCE (awAROCE), AUC (awAUC), and AROCE50

dx.doi.org/10.1021/ci2002268 |J. Chem. Inf. Model. 2011, 51, 3169–3179

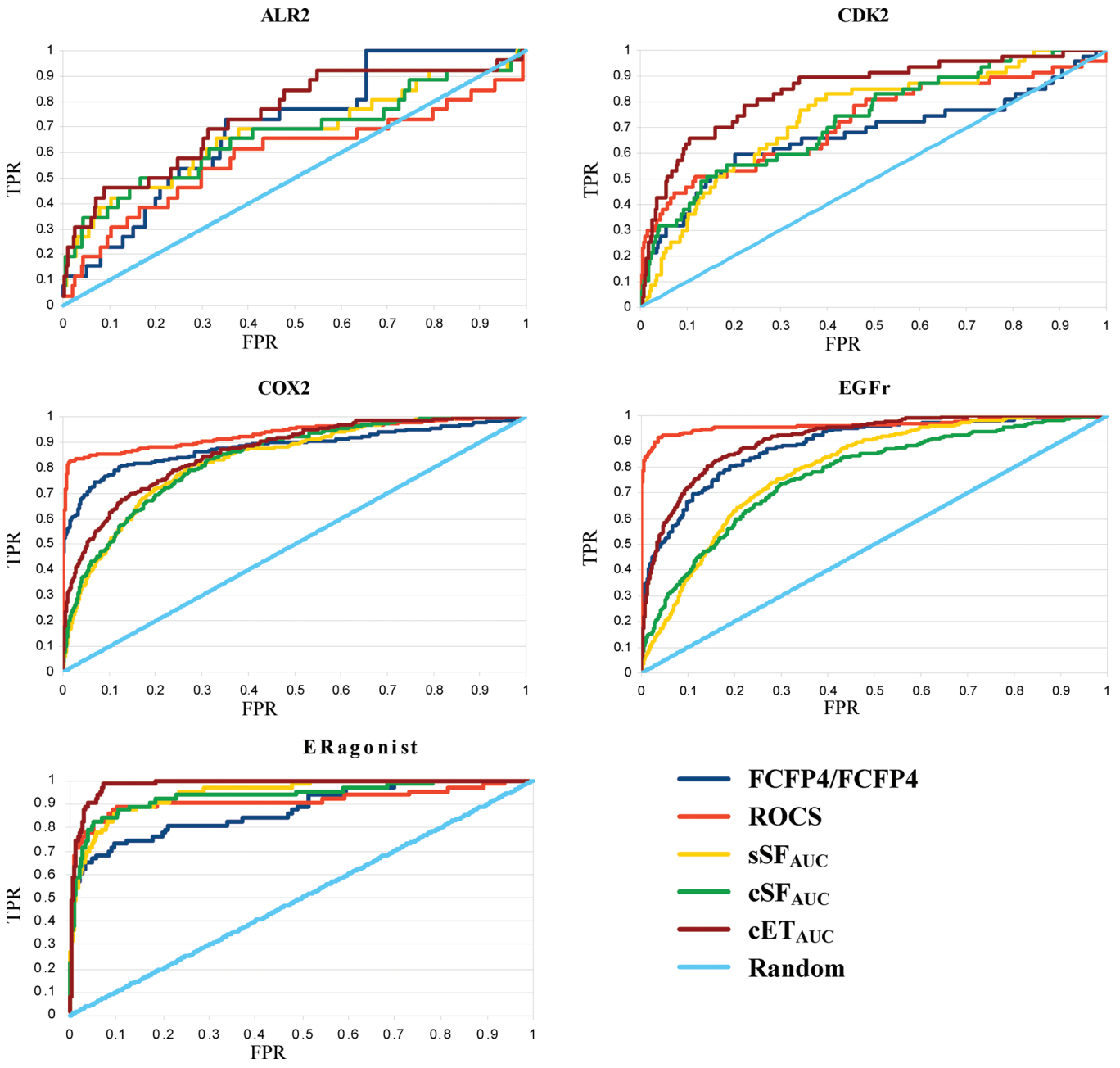


Figure 3. ROC plots of the five protein targets used in this study showing the performance of ligand-based 2D fingerprint (ECFP4 or FCFP4) and 3D ROCS similarity search methods and docking-based best single scoring function (sSF) and rescoring protocols (cSF and cET) considering the AUC selection criteria.

(awAROCE50). The same parameters were employed for the selection of combined SFs. The resulted cET and cSF were compared to single SFs, to ECFP4 and FCFP4-based 2D SS, and to 3D ROCS. The data set used consists of five DUD targets described in Table 1.

Along with the internal cross-validation described in the PLSDA-Docet Protocol section, an external test set (see External Validation section) was used in order to validate our protocol together with the other docking-based methods (i.e., sSF and cSF). In Table 3 the average performance (over the five targets) of the evaluated approaches are shown. The weakest retrieval rates for all docking-based methods (see the Supporting Information) were obtained in the case of ALR2. We will point out later that very often the results obtained for this target are poorer than those achieved for the others. For this reason Table 3 contains also the parameters calculated as mean values for COX2, EGFr, CDK2, and ERagonist.

Although the decoys/actives ratio for the external test set is generally much higher (see Table 2) compared to that of the DUD set used in our modeling, some satisfactory conclusions can be drawn. The smallest differences between DUD set and external set performances for the five target averages were found for the cET scoring functions. These findings are enforced by the four target averages (in brackets in Table 3), especially for AUC and AROCE50 results, for which the DUD set and external set efficiency are very similar. Also, in terms of the evaluation

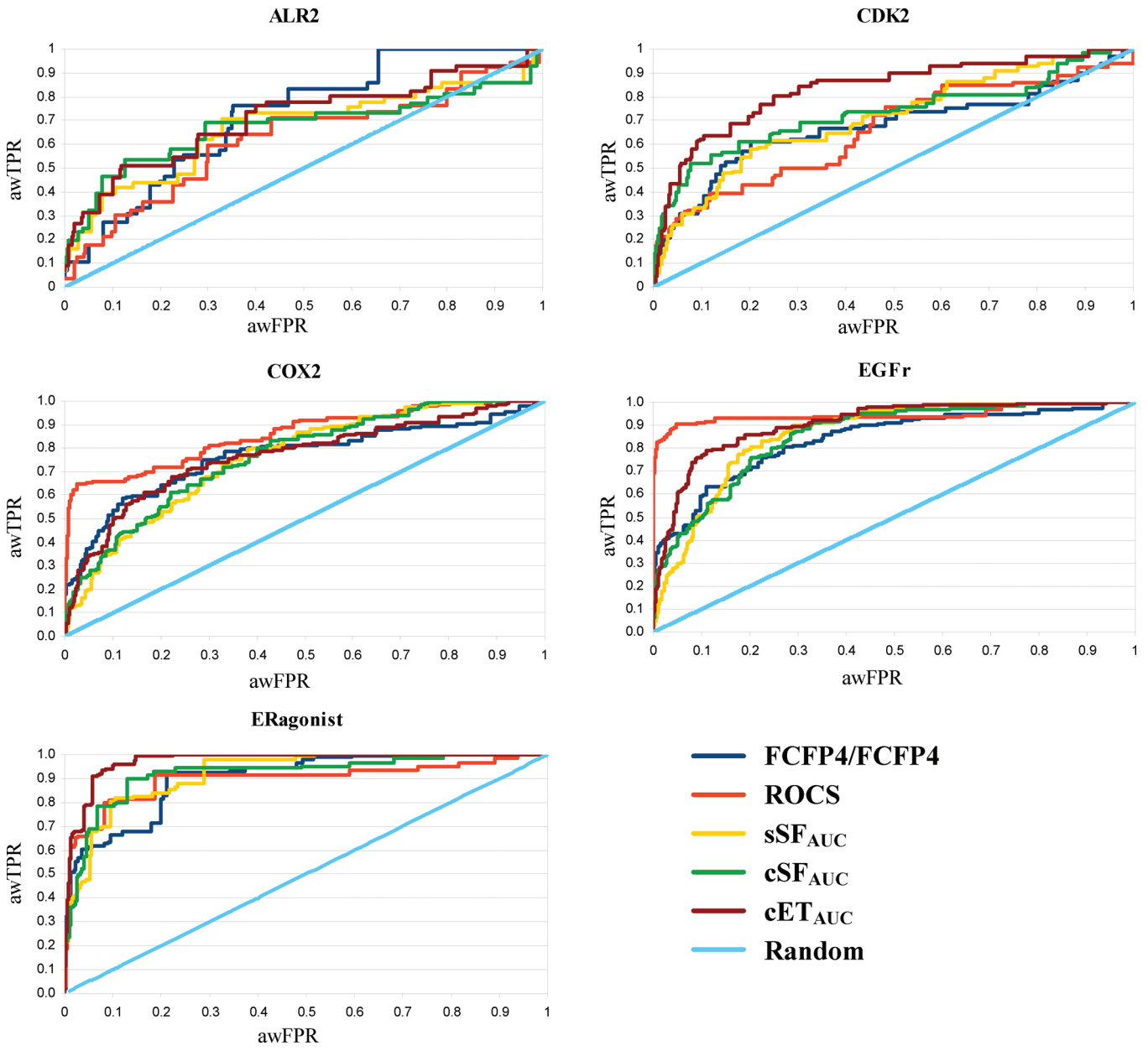


Figure 4. awROC plots (the arithmetic weighting scheme — aw — accounts for the retrieval of chemotypes) of the five protein targets used in this study showing the performance of ligand-based 2D fingerprint (ECFP4 or FCFP4) and 3D ROCS similarity search methods and docking-based best single scoring function (sSF) and rescoring protocols (cSF and cET) considering the awAUC selection criteria.

parameters represented in Table 3, in comparison to the other docking methods, the PLSDA-DOCET procedure demonstrates superior retrieval performance.

The averaged percentages (over the five DUD target sets) of TP and awTP in the top 1%, 5%, and 10% FPR (and respectively awFPR) are reported in Figure 1 and Figure 2 together with the average AUC and awAUC. To account for statistically significant differences between the average results paired-samples *t* tests were conducted for each of the two methods (detailed results are available in the Supporting Information).

Early Enrichment. The results reflected by the TP% at 1%, 5%, and 10% FPR show often statistically significant differences between the mean values of the docking methods. In the case of ligand-based methods the highest average early enrichment values were obtained with ROCS. However, for ROCS, statistically superior mean values of the TP% at 5% and 10% FPR were obtained only compared to ECFP4 (Figure 1). In Figure 2 FCFP4 is outperformed by almost all docking techniques in terms of mean awTP%. Lower scaffold-hopping abilities of simple fingerprint-based similarity methods compared to 3D SS and docking were also reported in earlier studies.[47,48] Regarding the docking approaches at 1% FPR, the average cSF$_{AROCE}$ values are slightly smaller in Figure 1 and slightly higher in Figure 2 compared to cET$_{AROCE}$. Among the docking techniques in the first 1% and 5% FPR the AUC and awAUC-based approaches show poorer performance. In Figure 1, at 10% FPR the average results indicate all cET methods to outperform significantly all sSF and cSF-derived methods.

dx.doi.org/10.1021/ci2002268 |*J. Chem. Inf. Model.* 2011, 51, 3169–3179

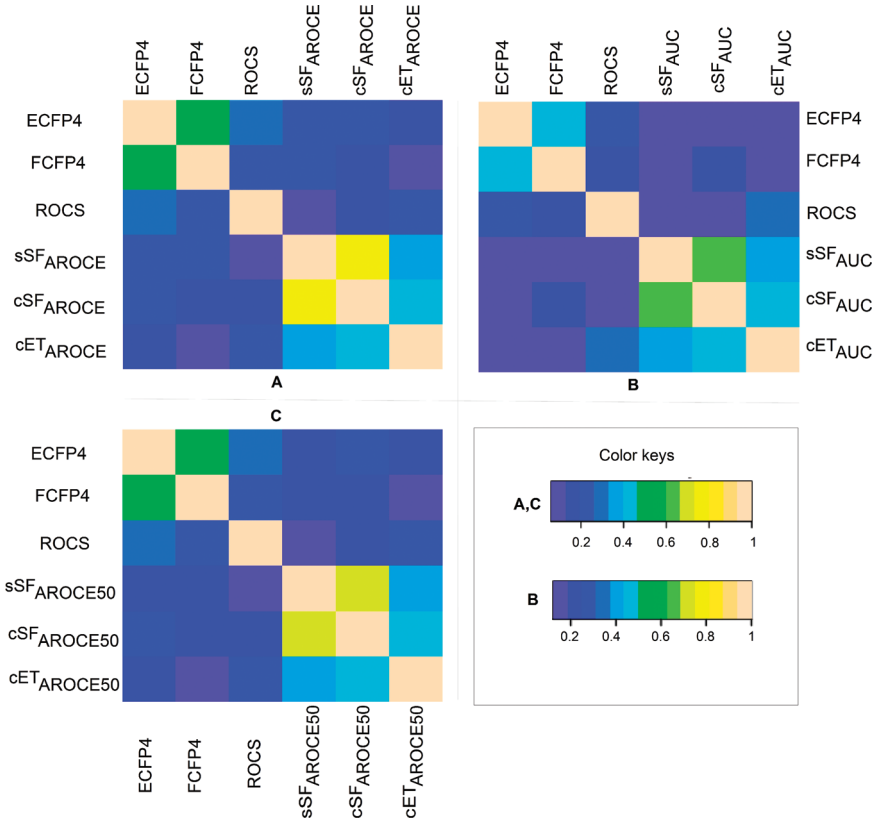


Figure 5. The averaged Kendall's tau correlation coefficient of the ranked lists of ligand-based methods and postdocking scoring protocols (best AROCE - **A**, AUC — **B**, and AROCE50 - **C**).

The average percentages of the awTPs in Figure 2 show ROCS and cET to generally exhibit superior early enrichment. At 10% FPR the awAUC-based selected cET show the highest average percentage of almost 65% found TP. The other methods indicate TP percentages around 50% at 10% FPR. It can be seen from Figure 2 that the awAROCE50-based selected sSF, cSFs, and cETs commonly exhibit TP% between the awAROCE and the awAUC-based selected docking percentages at the 1% and 5% FPR.

The general trend in Figures 1 and 2 shows ROCS to exhibit large variations in performance (indicated also by the ROC and awROC-plots in Figure 3 and Figure 4) and designate the AUC/awAUC-based selected cET to be a promising approach especially in situations where the enrichment in the top 5% or 10% of the database is important. The AROCE and AROCE50-selected docking methods (and also their aw-forms) have shown valuable enrichment abilities.

Overall discriminative power of the methods is computed by the five-target average AUC and awAUC and presented in Figures 1 and 2. The mean values are constantly higher for cET_{AUC}, cET_{AROCE}, and $cET_{AROCE50}$ (and their equivalent aw-form) relative to all the other approaches (docking and ligand-based). PLS-DA postdocking protocol (all cET variants) offers highly confident overall classification reflected in an AUC generally higher than 0.8. A possible explanation lies in the effectiveness of the multivariate method: the PLS-DA approach selected essential ETs separating most of the actives from most of the decoys and this ability is measured exactly by the AUC parameter (disregarding enrichment in the top of the database).

In Figure 3 and Figure 4 the ROC and awROC plots show the overall retrieval of ROCS, the best 2D SS and the AUC and awAUC-based selected docking methods. A quick overview reveals major variations of enrichment among the five targets: very low enrichment can be observed in the case of ALR2 and very high enrichment for $ER_{agonist}$. Figures 3 and 4 suggest large target-specific variations in performance among different approaches. No single ligand or docking method exhibited superior retrieval for all targets. However, cET_{AUC} or cET_{awAUC} can be considered the best or second best choice in the vast majority of the cases.

The PLSDA-DOCET procedure (i.e., cET_{AUC} and cET_{awAUC}) shows higher enrichment in the cases of $ER_{agonist}$ and CDK2, compared to other methods, and outperforms the sSF and cSF for EGFr and COX2. Moreover, it can be observed that sSF and cSF present similar ROC curves, cSF exhibiting only marginally better performance. Jacobsson et al.[9] applied three multivariate statistical methods (including PLS-DA) to seven SFs. The results indicate relative similar performance of cSF compared to the individual SFs.[9] Only slightly better early retrieval of the consensus-selected SFs compared to the parent single SFs can be observed also in the paper of Teramoto et al.[10] Our methodology differs from the one employed by Jacobsson et al.,[9] as follows: (i) our PLS-DA analysis has used the individual ET values resulting from each calculated SF, instead of the entire SF values; (ii) the cET components are selected from the ETs furnished by the PLS-DA technique, by optimizing several parameters (high enrichment, maximal AUC, or both). Moreover, the PLSDA-DOCET protocol is more complex, and, for the most of the evaluation parameters employed here, the cET variants outperformed both the sSF and the cSF (somewhat similar with Jacobsson's PLS-DA approach) versions.

The ligand-based methods are weaker compared to docking for ALR2 in both Figures 3 and 4. The ROC curves in Figure 3 indicate ROCS to perform excellent for EGFr and COX2 but disappointing in the case of ALR2 and CDK2. Kirchmair et al.[49] reported recently that among all the targets available in DUD, in the case of EGFr, ROCS provides one of the best retrieval rates at the top 1% of the database, while docking performance is the poorest. These results are in agreement with our findings related to the docking methods applied here. We computed the percentage of actives for which the Tanimoto similarity to the nearest neighbor in the decoy set is less than or equal to 0.85 presented in Table 1 (we employed the SciTegic MDL Public Keys fingerprints). The results show the highest percentages (poorest diversity of the ligand set and decoys) for ALR2 (92%) and CDK2 (95.75%), while for COX2, EGFr, and ERagonist lower percentages of 67.92%, 89.32%, and 73.02%. The explanation for the poor performance (in the case of ALR2 and CDK2) obtained with ligand-based methods (especially 2D SS) could lie in the high number of actives similar to the decoys.

Chemotype Affinities. A large number of VS approaches have been developed in the last years and challenges regarding a better assessment of these methods are discussed extensively in several papers.[33,34,38,50,51] Besides the retrieving rates and the scaffold-hopping abilities, comparing the ranks of the actives (or even better the chemotypes) retrieved by the evaluated methods would add useful information regarding practical aspects of data fusion in drug discovery (e.g., if two methods retrieve complementary chemotypes their results can be fused enhancing the lead-hopping in the top percentage of the database).

To evaluate the chemotype affinities of the investigated approaches we computed the ranks of the actives. For all the non-singleton clusters we averaged the ranks within each cluster so that a list of the size of the number of chemotypes resulted. The Kendall correlation coefficient[52] was computed using the "Kendall" package available in the R statistical packages.[53] Kendall's tau (τ) is a linear function of the number of pairs of items which are in different orders in two ranking lists. In other words τ is a nonparametric statistic used to measure the association between two measured quantities. If the two ranking lists are identical $\tau = 1$, if the order is exactly reversed $\tau = -1$, and if the lists are independent $\tau = 0$.

The calculations were performed on every target set for each pair of approaches. The Kendall coefficients were averaged over the proteins to compare the affinities among the investigated protocols. We computed different heatmaps based on the correlation coefficient values for each parameter: Figure 5 (A) shows the averaged τ Kendall values between the ligand-based methods and the AROCE best performing docking protocols (sSF_{AROCE}, cSF_{AROCE}, cET_{AROCE}), the heatmaps (B) and (C) are similar but containing the AUC and AROCE50 specific docking approaches. It can be easily observed that ligand-based and docking-based methods are separately clustered showing low (near to zero) averaged correlation values. This observation is consistent with the results of other studies,[37,48] showing the practical advantages of fusing the results of ligand and structure-based methods to enhance the success of a VS campaign.

Examining Figure 5 more carefully, relatively low mean-correlation (less than 0.4) can be observed between ROCS and the two 2D SS methods. Moreover, averaged τ values between the cETs and other docking approaches are about 0.45. We should mention here the low number of clusters for two of the proteins (ALR2 14 clusters and $ER_{agonist}$ 10 clusters; see Table 1) which increase the chance correlations and elevate the average value of the coefficients.

We are aware that DUD is not an ideal database for the optimization of methods, and given the small number of targets used in this study, we cannot make pertinent appreciations regarding the actual SFs and ETs behind the results obtained. We would like to mention, at this point, that we neither checked the goodness of the poses nor applied constraints to the docked conformers to ensure a valid pose of the ligand in the binding site, and this could be a possible source of error. Here, we applied docking in the purpose of virtual screening and measured only this competence. Regarding the ETs selected by the PLS-DA approach and further picked up in the cET, we like to point out that these terms do not necessarily reflect the specific interactions to the binding site of a protein, but the critical terms that would separate the actives (or chemotypes) from the decoys in the data set used. However, it can be stated that the components of the cETs are also, to some extent, target specific. Generally the ETs with the highest PLS-DA VIP values are present in the optimal cETs. The differences between the corresponding cET_{AROCE} ($cET_{awAROCE}$), cET_{AUC} (cET_{awAUC}), and $cET_{AROCE50}$ ($cET_{awAROCE50}$) functions consist usually in the presence or absence of a supplementary ET. To examine in detail the PLS-DA VIP values, the SFs and ETs present in this study the reader is advised to consult the Supporting Information.

CONCLUSIONS

In this paper we evaluated a target-oriented docking rescoring method named PLSDA-DOCET. It is based on the PLS-DA selection of scoring functions terms followed by the optimization of these terms for virtual screening purposes. Three ROC based parameters, namely the just introduced AROCE, the AUC, and the mean of the first two, AROCE50, were used for the selection of combinations of ETs and SFs. The results were compared to single SF and ligand-based 2D and 3D SS methods. The study was carried out on a subset of five protein-targets from the clustered version of DUD. We assessed the relative performances of the VS approaches in terms of early enrichment, the overall discriminative power, and the scaffold-hopping ability. ALR2 stands out as a difficult target for docking.[16] The binding site of this protein is in part exposed to the solvent but also includes an anionic and a hydrophobic pocket. The latter is highly flexible depending on the size of the bound ligand. The X-ray structure 1ah3 (present in DUD) contains the "tolrestat-conformation" specific only to one compound (i.e., tolrestat), while almost all the other known ligands induce the "sorbinil-" or the "zopolrestat-conformation" of ALR2 (two other conformation were observed but only for two compounds).[54] Together, these complex requirements designate ALR2 as a fine candidate for ensemble docking[55] rather than docking in a single protein conformation (applied in this work).

The five or four (without ALR2) target averaged values obtained by us for the early enrichment and AUC performances of our PLSDA-DOCET protocol outperform the other docking methods. However in terms of early enrichment we were not able to identify a method outperforming the others for all targets. The best overall performance is demonstrated by the AUC calibrated cETs. These facts suggest that a user is clearly advised to decide after a retrospective cET calibration (PLSDA-DOCET) study, in which way (implying cET_{AROCE}, cET_{AUC}, or $cET_{AROCE50}$) he will use profitably the results in a real-life virtual screening campaign.

The target-specific variation of the performances achieved by the methods is especially evident for the ligand-based methods. The scaffold-hopping abilities of the docking approaches were superior to 2D fingerprint-based SS. Among the docking methods, the results indicate the AUC-based PLSDA-DOCET protocol to perform best in enrichment and scaffold-hopping at 5% and 10% FPR. The combinations of SFs performed only slightly better than single SFs.

The chemotype affinities for ligand-based methods compared to docking are very different. Quasi-complementary ranking lists were found for PLSDA-DOCET compared to single or combined SFs, but further studies need to investigate the complementary chemotype affinities among the docking rescoring techniques. Fusing the results obtained by ligand-based and docking methods can enhance the probability to find new chemotypes in real-life VS campaign.

Since this paper provides only a brief evaluation of the DOCET protocol, applying the PLS-DA method in the first step, the performance of other promising machine learning techniques (e.g., SVM) will be evaluated by us possibly in the future.

■ ASSOCIATED CONTENT

Supporting Information. Tables containing the ChEMBL IDs used to build the external validation set, the results of permutation tests for PLS-DA, the detailed results for the external test set, the data related to the pair-*t* tests conducted to account for statistically significant differences between the averaged results, figures showing early and complete ROC curves, the calculated values for a number of evaluation parameters, the ranking lists corresponding to each method for every protein-target used in this study, the actual scoring functions and energetic terms behind the sSF, cSF, and cET labels, the values of the averaged Kendall coefficients, and also the PLS-DA VIP values for the cET's corresponding to the five targets. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kurunczi@umft.ro.

■ ACKNOWLEDGMENT

This work was supported by a grant from CNCSIS—UEFISCSU, project number PNII — IDEI 2390/2008. The authors thank to Accelrys, Inc. for providing the free access to Pipeline Pilot Student Edition package, Open Eye Scientific Software for academic license, and Dr. Erik Johansson (Umetrics, Sweden) for providing the SIMCA program package. Sorin Avram gratefully acknowledges Professors Peter Willett and Tudor I. Oprea for an openhanded training in chemoinformatics.

■ REFERENCES

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons. Inc.: New York, 1990; p 393.

(2) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovovery* **2004**, *3*, 935–949.

(3) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535–542.

(4) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 409–443.

(5) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377–395.

(6) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.

(7) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.

(8) 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.

(9) Jacobsson, M.; Lidén, P.; Stjernschantz, E.; Boström, H.; Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46*, 5781–5789.

(10) Teramoto, R.; Fukunishi, H. Consensus scoring with feature selection for structure-based virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 288–295.

(11) Teramoto, R.; Fukunishi, H. Supervised Consensus Scoring for Docking and Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 526–534.

(12) Fukunishi, H.; Teramoto, R.; Takada, T.; Shimada, J. Bootstrap-based consensus scoring method for protein-ligand docking. *J. Chem. Inf. Model.* **2008**, *48*, 988–996.

(13) Betzi, S.; Suhre, K.; Chétrit, B.; Guerlesquin, F.; Morelli, X. GFscore: a general nonlinear consensus scoring function for high-throughput docking. *J. Chem. Inf. Model.* **2006**, *46*, 1704–1712.

(14) Ståhle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1987**, *1*, 185–196.

(15) Wold, S.; Sjöström, M.; Eriksson, L. PLS in chemistry. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; Wiley: Chichester, UK, 1999; pp 2006−2020.

(16) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(17) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.

(18) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs ? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(19) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.

(20) Tammali, R.; Saxena, A.; Srivastava, S. K.; Ramana, K. V. Aldose reductase inhibition prevents hypoxia-induced increase in hypoxia-inducible factor-1alpha (HIF-1alpha) and vascular endothelial growth factor (VEGF) by regulating 26 S proteasome-mediated protein degradation in human colon cancer cells. *J. Biol. Chem.* **2011**, *286*, 24089–24100.

(21) Akli, S.; Van Pelt, C. S.; Bui, T.; Meijer, L.; Keyomarsi, K. Cdk2 is required for breast cancer mediated by the low-molecular-weight isoform of cyclin E. *Cancer Res.* **2011**, *71*, 3377–3386.

(22) Jiménez, P.; García, A.; Santander, S.; Piazuelo, E. Prevention of cancer in the upper gastrointestinal tract with COX-inhibition. Still an option? *Curr. Pharm. Des.* **2007**, *13*, 2261–2273.

(23) Jackman, D. M.; Miller, V. A.; Cioffredi, L. A.; Yeap, B. Y.; Jänne, P. A.; Riely, G. J.; Ruiz, M. G.; Giaccone, G.; Sequist, L. V.; Johnson, B. E. Impact of epidermal growth factor receptor and KRAS mutations on clinical outcomes in previously untreated non-small cell lung cancer patients: results of an online tumor registry of clinical trials. *Clin. Cancer Res.* **2009**, *15*, 5267–5273.

(24) Prins, G. S.; Korach, K. S. The role of estrogens and estrogen receptors in normal prostate growth and disease. *Steroids* **2008**, *73*, 233–244.

(25) *OMEGA*, version 2.0; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2009.
(26) *FRED*, version 2.2.5; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2009.
(27) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76–90.
(28) 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 AC-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol. (Cambridge, MA, U. S.)* **1995**, *2*, 317–324.
(29) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
(30) *SIMCA P*, version 9.0; Umetrics AB: Umea, Sweden, 2001.
(31) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Woold, S. *Multi- and Megavariate Data Analysis. Principles and Applications*; Umetrics AB: Umeå, 2001; pp 92−97; 489−491.
(32) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
(33) Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201–212.
(34) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.
(35) Hanley, J. A.; McNeil, B. J. The Meaning and Use of the Area under a Receiver Operating Characteristics (ROC) Curve. *Radiology* **1982**, *143*, 29–36.
(36) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141–146.
(37) Mackey, M. D.; Melville, J. L. Better than random? The chemotype enrichment problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154–1162.
(38) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *J. Cheminf.* **2009**, *1*, 14.
(39) *Pipeline Pilot Student Edition*, version 6.1.5; SciTegic, Accelrys Inc.: San Diego, CA, 2007.
(40) *ROCS*, version 3.1; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2010.
(41) Grant, J. A; Gallardo, M. A; Pickup, B. J. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653–1666.
(42) Overington, J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 195–198.
(43) The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **2011**, *39*, D214-D219.
(44) Jain, E.; Bairoch, A.; Duvaud, S.; Phan, I.; Redaschi, N.; Suzek, B. E.; Martin, M. J.; McGarvey, P.; Gasteiger, E. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinf.* **2009**, *10*, 136.
(45) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screenin. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(46) Durant, J. L; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
(47) 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.
(48) Krüger, D. M.; Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem* **2010**, *5*, 148–158.
(49) Kirchmair, J; Distinto, S; Markt, P; Schuster, D; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to optimize shape-based virtual screening: choosing the right query and including chemical information. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.
(50) 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.
(51) 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.
(52) Kendall, M. G. *Rank Correlation Methods*, 4th ed.; Griffin: London, 1976; p 55.
(53) R Development Core Team. *R: A language and environment for statistical computing*, version 2.11.1; R Foundation for Statistical Computing: Vienna, Austria, 2010. http://www.R-project.org (accessed Apr 07, 2011).
(54) Steuber, H.; Zentgraf, M.; La Motta, C.; Sartini, S.; Heine, A.; Klebe, G. Evidence for a novel binding site conformer of aldose reductase in ligand-bound state. *J. Mol. Biol.* **2007**, *369*, 186–197.
(55) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511–524.