JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

pubs.acs.org/jcim

pROC-Chemotype Plots Enhance the Interpretability of Benchmarking Results in Structure-Based Virtual Screening

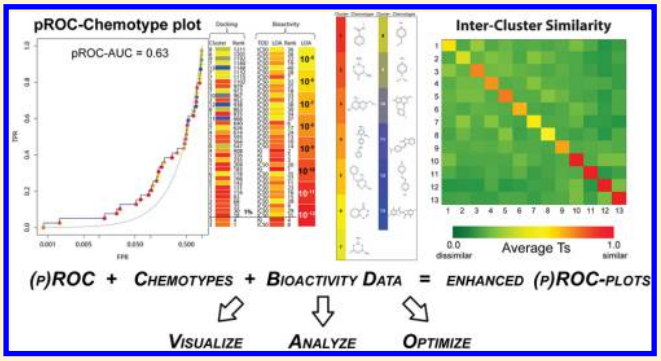
Tamer M. Ibrahim,[†,§] Matthias R. Bauer,[†,§] Alexander Dörr,[‡] Erdem Veyisoglu,[†] and Frank M. Boeckler*[,†,‡]

[†]Laboratory for Molecular Design and Pharmaceutical Biophysics, Department of Pharmaceutical and Medicinal Chemistry, Institute of Pharmaceutical Sciences, Eberhard Karls Universität Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany

[‡]Center for Bioinformatics Tübingen (ZBIT), Eberhard Karls University Tübingen, Sand 1, 72076 Tübingen, Germany

Supporting Information

ABSTRACT: Recently, we have reported a systematic comparison of molecular preparation protocols (using MOE or Maestro) in combination with two docking tools (GOLD or Glide), employing our DEKOIS 2.0 benchmark sets. Herein, we demonstrate how comparable settings of data preparation protocols can affect the profile and AUC of pROC curves based on variations in chemotype enrichment. We show how the recognition of different classes of chemotypes can affect the docking performance, particularly in the early enrichment, and monitor changes in this recognition behavior based on score normalization and rescoring strategies. For this, we have developed "pROC-Chemotype", which is an automated protocol that matches and visualizes ligand chemotype information together with potency classes in the pROC profiles obtained by docking. This tool enhances the understanding of the influence of chemotype recognition in early enrichment, but also reveals trends of impaired recognition of chemotype classes at the end of the score-ordered rank. Identifying such issues helps to devise score-normalization strategies to overcome this potential bias in an intuitive manner. Furthermore, strong perturbations in chemotype ranking between different methods can help to identify the underlying reasons (e.g., changes in the protonation/tautomerization state). It also assists in the selection of appropriate scoring functions that are capable to retrieve more potent and diverse hits. In summary, we demonstrate how this new tool can be utilized to identify and highlight chemotype-specific behavior, e.g., in dataset preparation. This can help to overcome some chemistry-related bias in virtual screening campaigns. pROC-Chemotype is made freely available at www.dekois.com.



INTRODUCTION

Structure-based virtual screening (SBVS) is among the most important approaches for drug discovery, because it is a knowledge-driven approach and has the advantage that, when relying on existing compound libraries, experimental verification can be achieved without synthetic efforts.[1] A pragmatic approach for preselecting a docking program prior to a SBVS effort is benchmarking. The benchmarking process evaluates the ability of a docking program to better score the bioactive molecules over the challenging, closely related, and presumably inactive molecules, also referred to as "decoys", at the respective target.[2−13] Many approaches are used to mathematically and graphically present the outcome from the docking/screening and benchmarking evaluation (e.g., EF,[14] ROC-AUC,[15] RIE,[16] BEDROC[17]). A good way to emphasize early enrichment while considering all docking performance is the widely used pROC-AUC protocol.[2,3,7,16,18,19]

There is no commonly agreed formal definition of a chemotype.[20] Chemotypes can be recognized as sets of molecules sharing certain molecular features that can be clustered into the same class.[20] Such classification of molecules by their chemotypes was previously done by both manual inspection[21,22] and automated approaches.[23−25]

Several approaches have tried to incorporate chemistry-related information into quality assessment metrics, such as ROC curves or enrichment factors.[18,20−22,26,27] For example, one approach calculated the percentage of the database required to find at least one member of each chemotype,[21] also referred to as "First Found" (FF) method.[20] Alternatively, another approach proposed setting the contribution of each active to the final score (true positive rate of the ROC curve) to be inversely proportional to the number of the other members in its cluster,[18] also referred to as "Cluster Average" method (CA).[20]

In this study, we use our recently published benchmarking results,[19] to explore how different protein/ligand preparation procedures, normalization and rescoring efforts affect chemo-

Received: July 10, 2015
Published: October 5, 2015

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

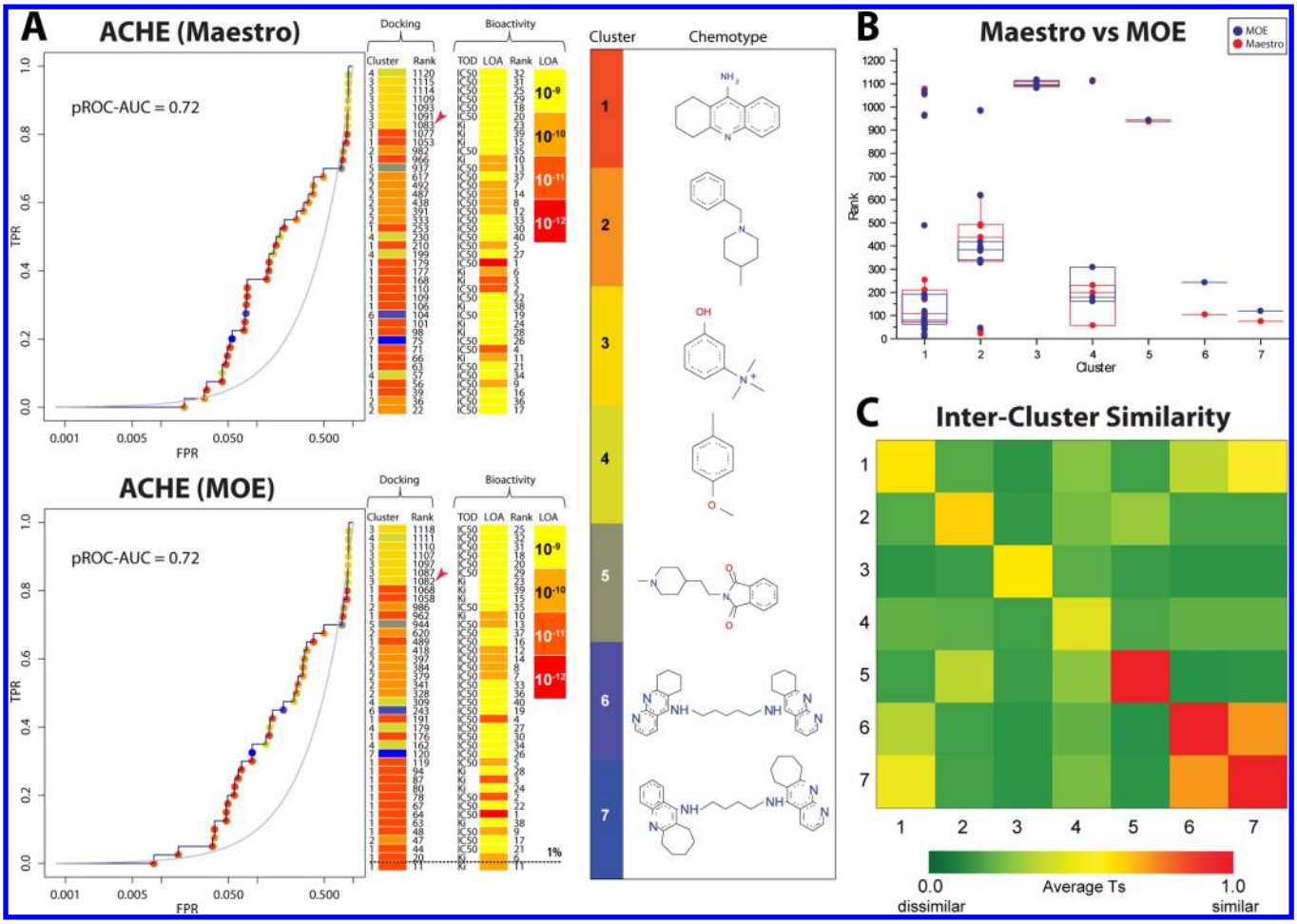


Figure 1. Chemotype profile of pROC curves for the ACHE dataset docked by GOLD after preparation with Maestro or MOE. (A) pROC-Chemotype plot including color-coded pROC curve, annotated by docking and bioactivity bar. FPR and TPR stand for False and True Positive Rate on the *X*-axis and *Y*-axis, respectively. The chemotype legend illustrates the most common substructures of each cluster and their assignment into the color-coded annotation scheme. The docking bar is composed of a cluster number, color code, and docking rank. The bioactivity bar is composed of the type of data (TOD), the level of activity (LOA), and rank information. Compounds ranked within 1% of the entire benchmark set are indicated by a dotted horizontal line. The red arrows indicate the First Found (FF) member of cluster 3 in each plot. (B) Box plot of the rank versus chemotype clusters. Red and blue dots/boxes indicate Maestro and MOE preparation schemes, respectively. (C) Heat map of the seven chemotype clusters of the ACHE benchmark set based on the average *Tanimoto similarity* (Ts) over all cross-cluster pairs (*FCFP_6* fingerprints). The color gradient represents changes in the average Ts. Green indicates maximum dissimilarity (Ts ≈ 0), and red indicates maximum similarity (Ts = 1); the yellow color is assigned to the cutoff Ts = 0.4.

type enrichment behavior. For this, we developed a new, automated protocol to match and visualize ligand chemotype information with the pROC profiles obtained by using our recently introduced DEKOIS 2.0 benchmark sets.[3] DEKOIS 2.0 sets have the advantage of being relatively small and homogeneous in size, featuring 40 bioactives and 1200 decoys per target. The bioactives have been chosen to represent the diversity of the bioactive chemical space for each target, trying to avoid unnecessary redundancies. Combining important biological and chemotype-related information with pROC plots, they can be used to identify chemistry-related issues in VS workflows, such as unfavorable protonation or tautomerization states of molecules or protein binding sites. In addition, they can help to detect potential bias in docking procedures that lead to a substantially impaired scoring of a respective chemotype class.

RESULTS AND DISCUSSION

Impact on Chemotype Behavior. Modifying the original ROC-AUC calculation and its plotted curve with different chemotype weighing factors (e.g., CA) is certainly useful to correct for some chemotype bias; however, it is less applicable for in-depth analysis of the problems or issues in the benchmark set. Comparison to the original ROC curve is required to track the differences, and the curve profile can change significantly, depending on the clustering method that has been applied. In addition, diverse chemotypes in the bioactive set can be related to the lead identification and optimization history. When diversity originates from early discovered chemotypes of low potency, comparison to optimized compounds of high potency that share the same chemotype can cause significant distortion of the resulting ROC curve. Therefore, we have created a pragmatic approach to integrate chemotype and bioactivity information in ROC plots, facilitating analysis of their influence on the comparison of VS performance of different docking tools. We provide an

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

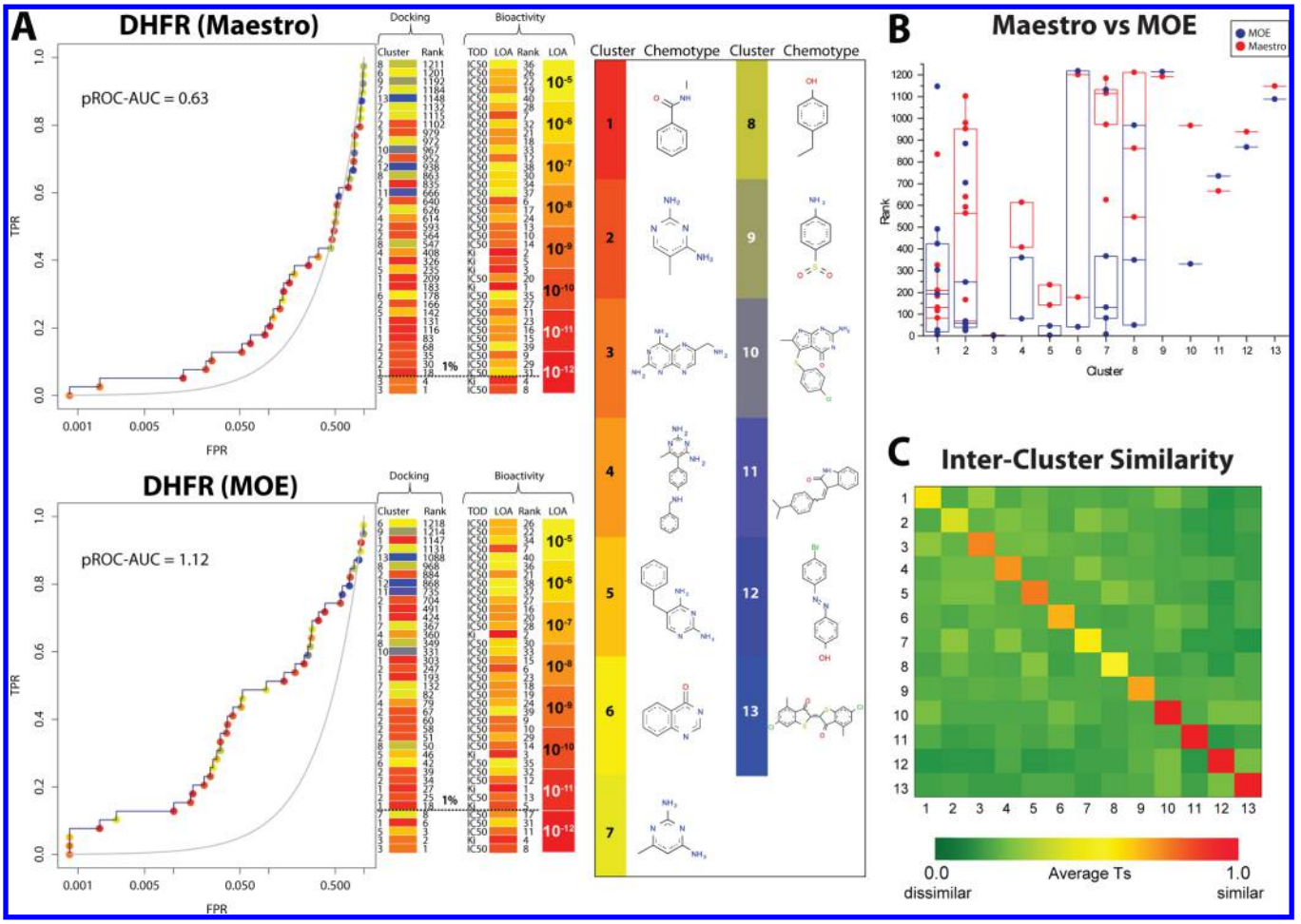


Figure 2. Chemotype profile of pROC curves for the DHFR dataset docked by GLIDE after preparation with Maestro or MOE. (A) pROC-Chemotype plot including color-coded pROC curve, annotated by a docking bar and a bioactivity bar. The chemotype legend illustrates the most common substructures of each cluster and their assignment into the color-coded annotation scheme. The docking bar is composed of a cluster number, color code, and docking rank. The bioactivity bar is composed of the type of data (TOD), the level of activity (LOA), and rank information. Compounds ranked within 1% of the entire benchmark set are indicated by a dotted horizontal line. When prepared by Maestro, 1 bioactive (from cluster 9) and 19 of the total 1200 decoys failed in the GLIDE docking protocol; these were, therefore, omitted from the visualization. When prepared by MOE, the same bioactive (from cluster 9) and 11 of the total 1200 decoys failed and were omitted. (B) Box plot of the rank versus chemotype clusters. Red and blue dots/boxes indicate Maestro and MOE preparation schemes, respectively. (C) Heat map of the 13 chemotype clusters of the DHFR benchmark set based on the average *Tanimoto similarity* (Ts) over all cross-cluster pairs (*FCFP_6* fingerprints). The color gradient represents changes in the average Ts. Green indicates maximum dissimilarity (Ts ≈ 0), and red indicates maximum similarity (Ts = 1); the yellow color is assigned to the cutoff Ts = 0.4.

automated protocol for studying chemotype enrichment, which we integrated into Pipeline Pilot[28] (employing a LibMCS component) and KNIME.[29] The detailed methodology is described in the Methods section below and in the Supporting Information (SI).

Visualizing the chemotype-dependent ranking can help to identify specific problems that potentially prevent a certain ligand class from being detected or enriched by the respective docking program. As a consequence, counteractive measures in the preparation or docking setup can be taken to correct or alleviate the problem. Based on the docking performance difference (ΔpROC-AUC$_{prep}$) between the two preparation schemes (Maestro vs MOE) reported in our previous study (see Tables S1 and S2 in the SI),[19] we selected two prominent cases to demonstrate the usefulness of our protocol: the ACHE dataset docked with GOLD and the DHFR dataset docked with Glide. Both examples showed the lowest and highest absolute values of ΔpROC-AUC$_{prep}$, respectively.

The "pROC chemotype plot" for the ACHE dataset (Figure 1) visualizes the GOLD screening performance. We obtained seven clusters based on the *Maximum Common Substructure* (MCS) concept for the bioactives, representing different chemotype classes with *level of activity* (LOA) values ranging from 10^{-9} to 10^{-12} M (recorded as IC$_{50}$ or K_i). It should be noted that IC$_{50}$ and K_i values are not necessarily comparable. However, there are other factors compromising data comparability, such as deviating assay conditions, varying assay types, and lack of reproducibility between different laboratories.[30] Thus, we consider reported IC$_{50}$ and K_i values to be similar enough, within the overall given limitation of assay comparability, to assign them to logarithmic affinity/activity classes. Although the screening performance was comparable between MOE and Maestro preparations with a pROC-AUC value of 0.72, we observed some variations in the chemotype enrichment. Such variations can be attributed to multiple factors, such as the difference in protonation/tautomerization states and input conformation between the two preparation

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

schemes.[19] Generally, we recommend the user to perform the experiment multiple times to identify any source of bias produced by the stochastic or heuristic nature of the docking algorithm. A member of chemotype cluster 1 was found within 1% of the database after MOE preparation. However, preparation in Maestro did not yield any bioactive hits for the same database cutoff. In addition, MOE preparation led to a better early enrichment of the bioactives with higher affinity/potency (typically belonging to cluster 1). Such a priorization of the recognition of highly potent ligands from large screening libraries is strongly desired. A rank comparison between both preparation schemes for all the ligands in each cluster (see Figure 1B) produced similar results for four clusters (1, 2, 3, and 5), while it shows perturbations for clusters 4, 6, and 7. Figure 1C highlights the relative intercluster (dis)similarity. The cluster pairs 6 and 7 and 1 and 7 share some substructures; therefore, their average *Tanimoto similarity* (Ts) indicates some similarity. Based on the dataset and the clustering technique, clusters 5−7 contained only one molecule per cluster. Thus, their average Ts was determined by using definition 1.

In addition to analyzing the early enrichment, the accumulation of certain chemotype clusters at the end of the ranked database can reveal some issues or biases in the dataset. Docking tools may systematically fail to accurately score and rank certain molecules based on problems resulting from ligand preparation, binding site preparation, intrinsic incompatibilities of certain key interactions with the scoring function, failure to detect the binding mode correctly due to ligand flexibility, and other factors. In the case of ACHE, the VS results showed that all ligands of cluster 3 were only found after 87% of the database. Our tool visualizes this type of chemotype enrichment bias in an intuitive way, and helps to identify and resolve issues prior to a VS effort. An example of how to make use of this information is described below in the Section "Score Normalization and Chemotype Behavior".

The DHFR dataset showed strong screening performance differences for GLIDE as a result of the different preparation schemes (see Figure 2). Clustering of the DHFR bioactives yielded 13 MCS clusters with LOA ranging from 10^{-5} M to 10^{-12} M. Since MOE preparation demonstrated better performance, molecules from chemotype clusters 3, 5, 1, and 7 were found in the early enrichment within 1% of the ranked database. On the other hand, only molecules from cluster 3 were found in the same segment of the ranked database after preparation with Maestro. Interestingly, both members of cluster 3 were ranked in the top four positions independent of the preparation scheme, despite the obvious perturbations in the recognition of other chemotype clusters. An important prerequisite for this result is that the protonation state of the bioactive ligands of cluster 3 did not differ between MOE and Maestro preparations (see Table 1, presented later in this work), while different protonation/tautomerization states were assigned to 50% of the bioactive ligands in the DHFR dataset (see Figure S1 in the SI). The correlation between perturbations in chemotype recognition and variations in protonation/tautomerization states (e.g., in clusters 1, 2, 4, 6, 7, and 8) is demonstrated in Table 1 (presented later in this work) and Figure 2B. As an illustrative example, the behavior of cluster 10 can be compared with that of clusters 11−13, since all of them possess only one bioactive ligand per cluster. Unlike clusters 11−13, the bioactive ligand in cluster 10 was affected by the preparation protocols, yielding a different tautomerization state, and, hence, suffered a loss of more than 600 positions in the overall ranking (as indicated in the box plot of Figure 2B).

Table 1. Overview of the Protonation/Tautomerization States of Chemotype Clusters of the DHFR Active Set

cluster	cluster size	$N_{\Delta prot}$[a]	average pairwise root-mean-square deviation, RMSD[c]
1	8	4	**2.5**
2	10	7	1.7
3	2	none	1.9
4	2	1	1.3
5	2	none	**2.4**
6	2	1[b]	1.8
7	5	4	1.7
8	3	2	1.9
9	2	none	**2.4**
10	1	1[b]	1.2
11	1	none	0.7
12	1	none	0.6
13	1	none	0.1

[a]Number of ligands in each cluster that differ in protonation state. [b]Difference in the tautomerization state. [c]RMSD is calculated by GOLD suite. Bold-formatted RMSD values indicate highest values.

To visualize how the change in the protonation/tautomerization states affects the prediction of the experimentally determined binding mode and, thus, the docking score and rank, Figure 3 shows the docked pose of the most potent bioactive ligand in the dataset of DHFR (K_i = 0.005 nM),[31] also noted as ligand **1** in this study. It is clear that the protonation state of the pyrimidine ring predicted by Maestro for **1** hindered the prediction of the correct binding mode (docking rank = 183, at ∼15% score-ranked database), while the protonation state predicted by MOE reproduced the correct pose of the pyrimidine ring with the key polar contacts of the surrounding residues (docking rank = 27, at ∼2% docking-ranked database) in 1S3V.[32]

However, detailed inspection of these key contacts reveals a severe problem. The automated preparation protocol in both programs, MOE and Maestro, assigned the proton between the carboxylate of GLU30 and the nitrogen in position 1 of the 5,6,7,8-tetrahydroquinazoline-2,4-diamine moiety to the carboxylate, producing a neutral interaction with two parallel hydrogen bonds. Instead, a shift of the proton to the nitrogen appears more plausible. This was confirmed by QM calculations of a representative subsystem (as shown in Figures 3E and 3F), including all key interactions on the MP2/TZVPP level of theory. The salt bridge between the protonated ligand and the carboxylate in GLU30 (Figure 3F) is strongly preferred over the uncharged hydrogen bonds (Figure 3E). Thus, the reason for the preference of the MOE preparation seems to be dependent on the protonation state of GLU30. However, when the ligand-free protein was subjected to the same preparation protocol, we always retrieved the deprotonated GLU30. As a consequence, we also performed docking of the DHFR benchmark set into the binding site bearing a deprotonated GLU30 to investigate the potential bias of the automated protocols. The results are shown in Figure 4. The preference for the MOE preparation scheme was still maintained, although the 5,6,7,8-tetrahydroquinazoline-2,4-diamine and similar substructures were not protonated by MOE. However, the pROC-AUC value for the MOE preparation was significantly decreased (from 1.12 to 0.61). The pROC-AUC value for the

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

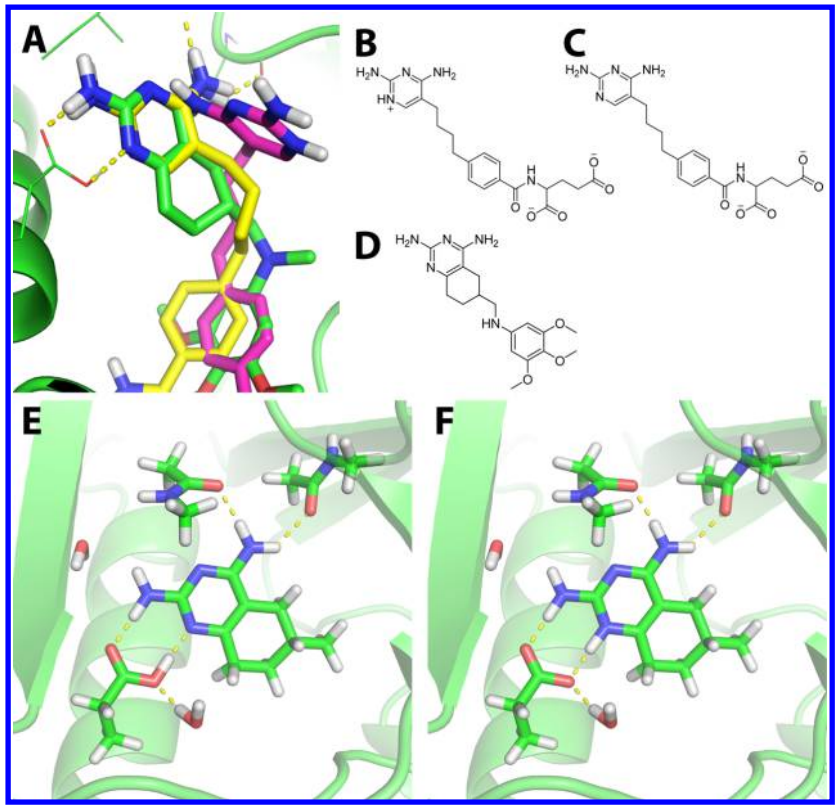


Figure 3. (A) Overlay of the co-crystallized ligand (green) of 1S3V with the poses of ligand **1**, the most potent bioactive in the DHFR dataset, retrieved from GLIDE-docking after MOE preparation (yellow) or Maestro preparation (purple). (B) Protonation state difference for ligand **1** based on Maestro preparation schemes. (C) Protonation state difference for ligand **1** based on MOE preparation schemes. (D) Structure of the co-crystallized ligand in 1S3V. (E, F) Subsystem of the ligand−protein complex in 1S3V that was subjected to QM calculations on a MP2/TZVPP level of theory: (E) parallel hydrogen bonds of the carboxylic acid of GLU30 with the 5,6,7,8-tetrahydroquinazoline-2,4-diamine moiety of the ligand in the neutral complex; (F) deprotonated GLU30 forming a salt bridge to the protonated ligand. Adduct formation energies were calculated as $\Delta E = E(\text{complex}) - E(\text{ligand}) - E(\text{binding site})$. The difference in adduct formation energy $\Delta\Delta E$ between panels F and E is huge, clearly favoring panel F.

Maestro preparation scheme remained constant and did not benefit from the deprotonation of GLU30. For the Maestro preparation, ligand **1** was now truthfully recognized as the most potent bioactive in the dataset, but the inferior recognition of other bioactives compensated this improvement. Although the deprotonation of GLU30 is reasonable from a biochemical and thermodynamic point of view, the docking results are surprising. This example demonstrates that our new tool for chemotype-dependent analysis of docking results facilitates the detection and closer investigation of intrinsic problems in benchmarking and VS strategies.

Another factor that can contribute to the perturbation of the docking rank—and, hence, the "chemotype behavior"—is the difference in the input conformation of the respective molecule when prepared differently. A minor change in the input structure can have a major impact on the docking outcome, especially for deterministic docking tools such as Glide.[33] For the initial docking using the automated preparation protocol (Figure 2), we evaluated the average pairwise root-mean-square deviation (RMSD) for all bioactives, comparing the conformer retrieved after preparation with MOE to the conformer retrieved after preparation with Maestro. Assuming that there is a correlation between the average pairwise RMSD per cluster and a perturbation in the chemotype enrichment behavior, we expected to find more pronounced perturbations in clusters 1, 5, and 9, because they have the highest average pairwise RMSDs among all chemotype clusters. However, only for cluster 5, there is evidence for a dependence on the ligand input structure. For cluster 1, the differences in ligand protonation states are more likely to affect the score-ordered rank than the variations of the input structure. Molecular recognition of the bioactive ligand from cluster 9 seems to be generally impaired. Therefore, the observed VS performance perturbations for cluster 5 can be attributed to different ligand input conformations. Compared to the impact of different protonation/tautomerization states in other clusters, the change in the ligand ranking was not as pronounced.

Score Normalization and Chemotype Behavior. Normalizing the docking score by molecular size parameters (e.g., number of heavy atoms) is a well-known strategy to minimize the size-related bias of the empirical scoring functions toward larger molecules being ranked at the top of a score-ranked list of molecules.[34,35] In our recently published study, we demonstrated the impact of the normalized docking score by the square root of the number of heavy atoms ($N^{1/2}$) or the cubic root of the squared number of heavy atoms ($N^{2/3}$) on the VS performance of the selected subset of the DEKOIS 2.0 datasets.[19]

In this part of the paper, we focus on investigating the score normalization strategies on the chemotype behavior employing the pROC-Chemotype protocol. A good example is the ACHE dataset docked by GOLD (e.g., Maestro preparation), as shown in Figure 5. The original docking performance of GOLD (Figure 5A) showed the accumulation of cluster 3 at the end of

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

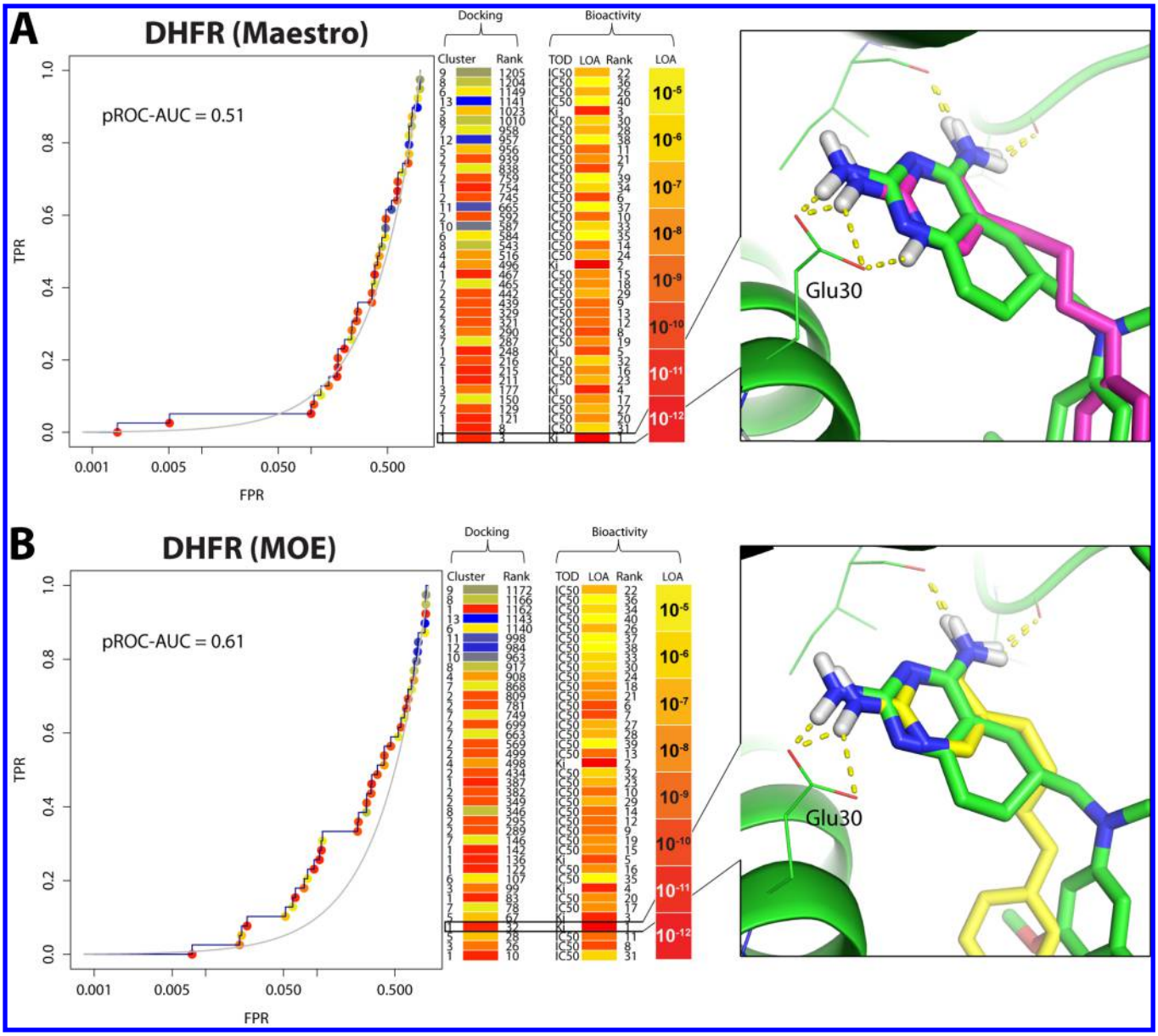


Figure 4. Chemotype profile of pROC curves for the DHFR dataset obtained by docking into the binding site after deprotonation of GLU30. Docking was performed by GLIDE after preparation with (A) Maestro or (B) MOE. The pROC-Chemotype plot shows the color-coded pROC curve, annotated by a docking bar and a bioactivity bar. Chemotype clusters and color codes are identical with Figure 2. The docking bar is composed of a cluster number, color code, and docking rank. The bioactivity bar is composed of the type of data (TOD), the level of activity (LOA), and rank information. The most active compound, ligand **1**, is highlighted in the ranking and the pose obtained in the docking is shown (in magenta for **1** prepared by Maestro and in yellow for **1** prepared by MOE) superposed onto the co-crystallized ligand in 1S3V (green).

the score-ranked database, although all the members of this cluster display relatively good potency at the nM (10^{-9} M) range of bioactivity. After normalizing the docking scores by ($N^{1/2}$) and ($N^{2/3}$), particularly chemotype cluster 3 displayed a more homogeneous distribution. The FF member of cluster 3 in the original docking was located at rank 1083 (~87% of the database), while normalization by $N^{1/2}$ or $N^{2/3}$ improved the rank of this ligand to 922 (~74% of the database) or 256 (~20% of the database), respectively. It demonstrates how emphasizing the number of the heavy atoms in the docking score can influence the recognition of an entire cluster of ligands. This is not surprising, since ligands in chemotype cluster 3 have the smallest average number of heavy atoms, as well as the smallest average molecular weight (see Table 2).

Expressing bioactivity data as ligand efficiency (LE)[36] is reasonable in this example, where the score is also normalized based on the number of heavy atoms. Score normalization by $N^{1/2}$ showed a small improvement, with respect to LE, compared to the original docking. However, the members of cluster 3 with the highest LE in the dataset were not significantly enhanced (see Figures 5A and 5B). Normalization by $N^{2/3}$, where the N is more emphasized, produced a more reasonable recognition of ligands with high LE rank including cluster 3 (Figure 5).

Thus, the pROC-Chemotype plot can help to explore and rationalize strategies (e.g., normalization) to overcome impaired ranking of certain chemotypes, which failed to be properly ranked, based predominantly on their average size.

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

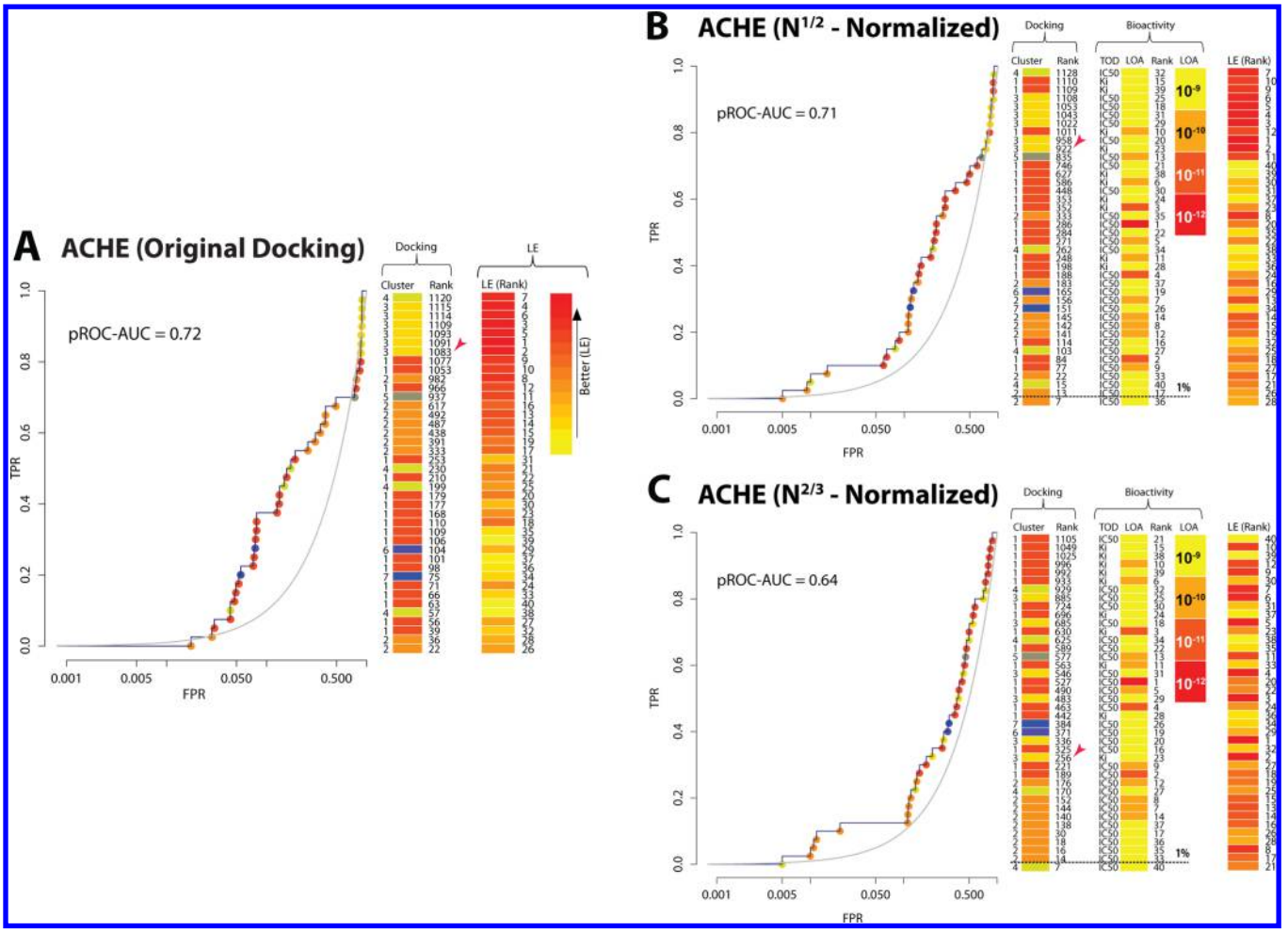


Figure 5. Chemotype profile of pROC curves for the ACHE dataset docked by GOLD and prepared with Maestro: (A) original docking performance without score normalization, (B) normalized scores by $N^{1/2}$, and (C) normalized scores by $N^{2/3}$. The pROC-Chemotype plot shows the color-coded pROC curve, annotated by a docking bar, a bioactivity bar, and a ligand efficiency (LE) bar. Chemotype clusters and their color codes are identical to those in Figure 1. The docking bar is composed of a cluster number, color code, and docking rank. The bioactivity bar is composed of the type of data (TOD), the level of activity (LOA), and rank information. A yellow-to-red color gradient indicates the LE rank. The red arrow in the docking legend indicates the first found (FF) member of cluster 3.

Table 2. Molecular Weight and Number of Heavy Atoms Distribution of the Chemotype Clusters in the ACHE Dataset

cluster	mean number of heavy atoms, N	mean molecular weight, MW	cluster size
1	35.61	491	18
2	27.78	374.1	9
3	12.33	170.9	6
4	28.75	392	4
5	20	273.4	1
6	37	495.7	1
7	38	508.7	1

Rescoring and Chemotype Behavior. Rescoring a docking solution with other scoring functions is a common methodology to compare the results and get information about the relative accuracy of these functions, as well as retrieving more hits for a VS effort.[37,38] We rescored the docked poses with various types of scoring functions provided by GOLD suite v5.2: ASP (a knowledge-based scoring function; see Table 3), Goldscore (a force-field-based scoring function), Chemscore (an empirical scoring function), and ChemPLP (for rescoring Glide data only; see Table S3 in the SI).

There is no clearly dominant performance of any scoring function in the rescoring evaluation. The rescoring performance varies in a target-dependent manner. To elucidate the effect of the rescoring on the chemotype perturbation behavior, we selected the rescoring outcome of the already herein discussed DHFR dataset (now for the GOLD poses), since it showed the highest ΔpROC-AUC$_{RESCORE}$. As mentioned previously, the clustering of the DHFR bioactives yielded 13 MCS clusters, representing different chemotype classes with LOA ranging from 10^{-5} M to 10^{-12} M. The chemotype plots in Figure 6 show that the rescoring with Chemscore failed to retrieve the same number of bioactives in the early enrichment (within 1% of the dataset), although it showed a comparable overall pROC-AUC (0.81 vs 0.80). Goldscore showed slightly improved overall performance (pROC-AUC = 0.89) and retrieved exactly the same bioactives in the early enrichment zone. Interestingly, ASP showed a substantially better overall performance (pROC-AUC = 1.53) and ranked a much higher number of the potent bioactives (with pM to nM affinities), including new chemotype classes, even within the top 0.5% of the data set. Extending

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

Table 3. Overview of the Rescoring Performance of the Docked Poses Generated by GOLD Using the Scoring Function ChemPLP and the Maestro Preparation Scheme

		pROC-AUC (Scoring Function)			
target	pROC-AUC (original)	ASP	Chemscore	Goldscore	positive ΔpROC-AUC$_{RESCORE}$ [a]
ACE	1.25	0.31	0.85	0.83	
ACHE	0.72	0.67	**0.76**	0.52	0.04
ADRB2	0.65	0.63	**0.80**	0.27	0.14
CATL	0.75	0.62	0.64	0.42	
DHFR	0.80	**1.53**	0.81	0.89	0.73
ERBB2	2.12	1.96	1.01	0.78	
HDAC2	1.20	0.89	1.17	0.93	
HIVPR	1.66	1.46	1.33	0.43	
HSP90	0.33	**0.48**	0.36	0.46	0.15
JAK3	0.91	0.91	0.81	0.81	
JNK2	0.74	**0.80**	0.61	0.73	0.06
MDM2	0.44	0.39	**0.50**	0.34	0.06
P38a	0.56	0.56	0.38	0.52	
PI3KG	0.95	**1.06**	0.77	0.99	0.11
PNP	1.04	1.19	0.60	**1.26**	0.22
PPARg	0.92	0.69	**1.07**	0.58	0.15
thrombin	1.15	**1.28**	0.78	0.71	0.14
TS	1.05	0.81	0.60	0.96	

[a]To highlight improvements by rescoring, only the highest value of pROC-AUC (indicated by bold numbers) was used to calculate ΔpROC-AUC$_{RESCORE}$, according to the following equation: ΔpROC-AUC$_{RESCORE}$ = (pROC-AUC (scoring function) − pROC-AUC (original)). pROC-AUC (original) values were used as previously reported.[19] The rest of the rescoring evaluation for Glide and GOLD with the two preparation schemes MOE and Maestro is available in the Supporting Information (Table S3).

the enrichment analysis to 1% of the database, additional bioactives with new chemotype classes were found. This example highlights the utility of our protocol to visualize within the same pROC-Chemotype plot, whether more potent and diverse hits were retrieved in the early enrichment. Thus, it can facilitate the selection of scoring/rescoring protocols that are more promising for a VS effort.

METHODS

Intercluster Similarity Assessment. To assess intercluster similarities, we calculated the average Ts values, based on FCFP_6 fingerprints (Scitegic) between all molecules in a dataset. For two molecule clusters I and J of n and m molecules, respectively, our protocol generates a list $T(i,J)$ of Tanimoto similarities of molecule i to all molecules in cluster J. In order to compute a general similarity Ts between clusters I and J, for each molecule i, we take the mean of $T(i,J)$, which we call $\dot{T}(i,J)$, and set Ts to be the median of $\dot{T}(i,J)$:

$$\mathrm{Ts} = \mathrm{Median}(\dot{T}(i, J),\ i = 1, ..., n)$$

Preparation Procedures and the Selected DEKOIS 2.0 Benchmark Sets. Preparation of structure and ligand data were done as described.[19] Briefly, we prepared the dataset (bioactive and decoy sets) using the LigPrep module in Maestro or the Wash and Minimize functions of MOE. Similarly, target structures were prepared by the ProtAssign function in Maestro and the Prot3D function in MOE. We have recently reported an extensive analysis of the impact of the previously mentioned settings on the overall VS performances.[19] To include a broad variety of protein−ligand binding situations, we use 18 diverse DEKOIS 2.0 datasets,[2,3] each representing a different enzyme class, including proteases, kinases, transferases, oxido-reductases, nuclear receptors, and hydrolases. This allows for a comprehensive analysis of performance differences for various VS experiments. A complete list of the datasets, PDB codes used, and the benchmarking performance of the selected subset can be found in the Supporting Information (Tables S1 and S2).

pROC-Chemotype Automated Protocol. Our protocol matches and visualizes ligand chemotype information in combination with the pROC profile obtained by docking, without biasing or modifying the original pROC graph. Cluster number and rank are annotated. For each bioactive molecule, we provide information about the type of data (TOD), the level of activity (LOA), and its bioactivity rank, which also serves as a molecular identifier. The clustering method employed in our protocol is based on *Maximum Common Substructures* (MCS), as implemented in ChemAxon.[28,39] In addition, we provide a comparable protocol based on a self-implemented MCS clustering of *Small Molecule Subgraph Detector* (SMSD).[40,41] MCS clustering reflects the shared scaffolds of the molecules in a clustered class. This type of clustering is intuitively useful for medicinal/computational chemists. We employ in the main article the ChemAxon-based MCS clustering, but show examples of the SMSD-based MCS protocol in Figures S2 and Figure S3 in the SI. The automatization of the protocol was achieved by Pipeline Pilot[28] and, alternatively, by KNIME.[29] A detailed description of the protocol is available in the SI.

Preparation of Targets and DEKOIS 2.0 Datasets. The VS performances for the 18 datasets (i.e., pROC−AUC values) were adopted from our previous publication.[19] The pairwise RMSD calculation for the comparison of the two prepared conformers (per bioactive ligand) retrieved by Maestro and MOE preparation schemes was done by employing an in-house *Python* script, applying the "smart_rms" function of GOLD.

Docking Experiments. Docking data employed by GOLD (version 5.1)[42−45] and Glide (version 5.6)[46−48] were adopted from our previous publication.[19] The rescoring was conducted by the scoring functions available in the GOLD suite, namely, ASP (knowledge-based scoring function), Goldscore (force field scoring function), ChemPLP (for rescoring Glide data only), and Chemscore (empirical scoring functions). The rescoring was performed with the default setting and the option "local optimization" was enabled.

CONCLUSION

Herein, we present a smart way of enhancing ROC plots by introducing chemotype, ranking, and bioactivity annotations. The plots can be generated using an automated protocol in Pipeline Pilot or KNIME. We highlight how this tool can be useful to visualize and understand "chemotype behavior" viz. trends, perturbations, and artifacts in chemotype enrichment, when using different docking or preparation protocols in benchmarking. Our tool provides a comprehensive graphical analysis of the resulting chemotype clusters based on the *Maximum Common Substructures* (MCS) method and their influence on docking performance, as well as their reported bioactivity profile. Without introducing any bias into the ranking, such "pROC-Chemotype plots" offer possibilities for optimizing virtual screening strategies based on benchmarking. We exemplify this use of our DEKOIS 2.0 benchmark sets by investigating optimization of early enrichment performance,

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

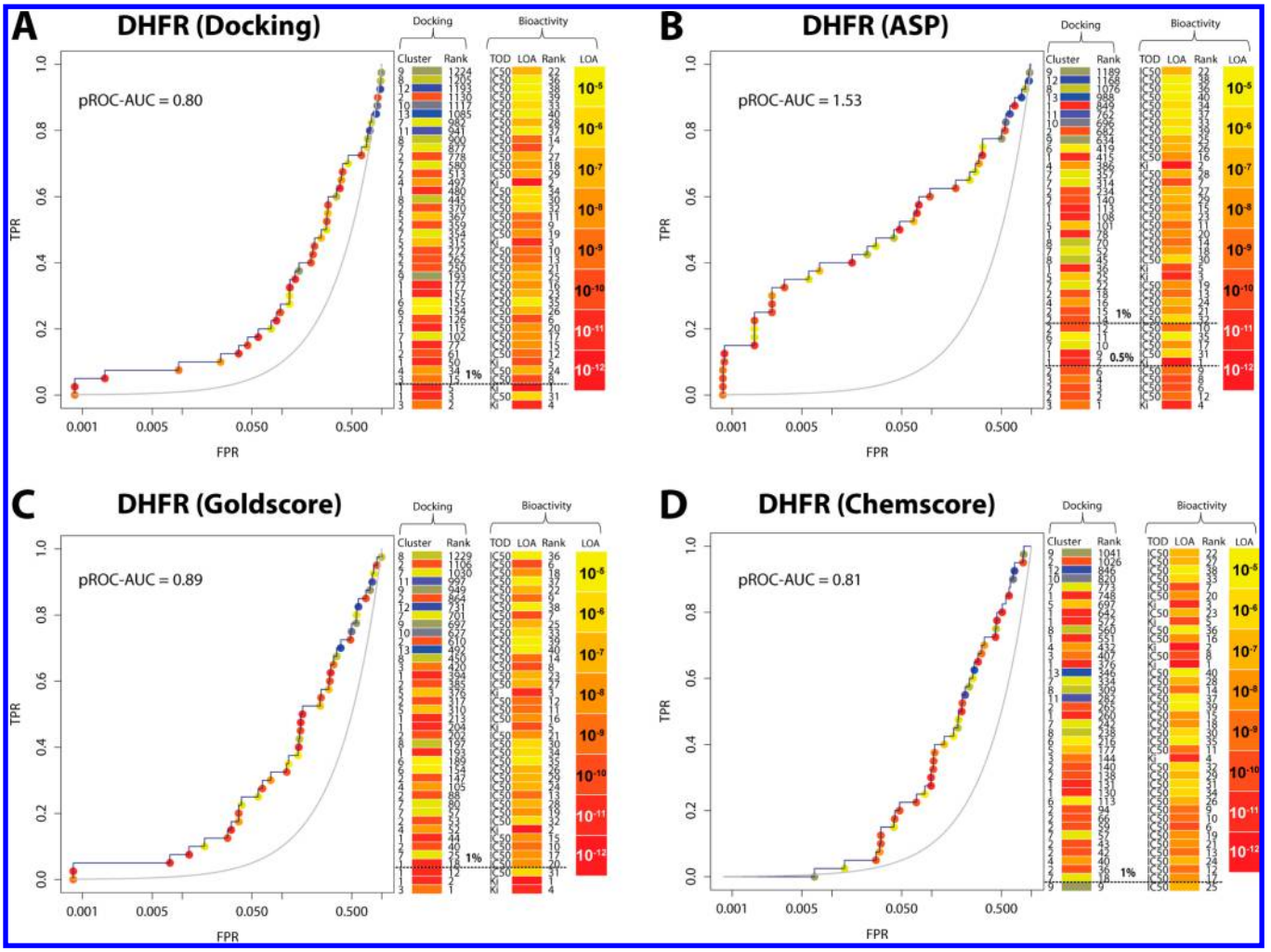


Figure 6. Rescoring performance of the DHFR dataset of the docked poses by GOLD (using ChemPLP and the Maestro preparation scheme): (A) pROC-Chemotype plot for the original GOLD (ChemPLP) docking performance; (B) rescoring performance for ASP; (C) rescoring performance for Goldscore; and (D) rescoring performance for Chemscore.

score normalization strategies, and rescoring strategies. We also demonstrate that pROC-Chemotype plots can help to intuitively explore the underlying reasons for variations in screening performance, e.g., changes in protonation/tautomerization states, molecular flexibility, and other parameters characteristic for some chemotype clusters. We conclude that this tool could be a real alternative for visualizing ROC curves with a high information density without additional efforts. Hence, we suggest that this depiction or similar enhanced ROC plots should be used as a standard in reporting benchmarking results. The automated Pipeline pilot and KNIME protocols are made available to the community free of charge at www.dekois.com.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00475.

Overview of the targets (Table S1), comparison of docking performances between MOE and Maestro preparation schemes for 18 diverse protein targets (Table S2), and overview of the rescoring performance of the docked poses by GOLD and Glide prepared by Maestro and MOE schemes (Table S3). Overview of the protonation/tautomerization states selected by MOE and Maestro preparations for the DHFR dataset (Figure S1), "chemotype behavior" of the TS dataset docked by GOLD and prepared by Maestro preparation schemes using the SMSD-based MCS method (Figure S2), and "chemotype behavior" of ACHE dataset docked by GOLD and prepared by Maestro preparation schemes using the SMSD-based MCS method (Figure S3). Detailed description of the "pROC-Chemotype plot" protocol and QM methods for the assessment of the preference of different Glu30/ligand protonation states in the binding site of DHFR (PDB: 1s3v) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: frank.boeckler@uni-tuebingen.de.

Author Contributions

§T.M.I. and M.R.B. contributed equally to this work.

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

■ ACKNOWLEDGMENTS

T.M.I. is grateful to the GERLS (German−Egyptian Long-Term Scholarship) Program of the German Academic Exchange Service (DAAD) for funding his Ph.D. fellowship. We thank Thomas Exner for his valuable comments regarding the manuscript.

■ REFERENCES

(1) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(2) Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: demanding evaluation kits for objective *in silico* screening—A versatile tool for benchmarking docking programs and scoring functions. *J. Chem. Inf. Model.* **2011**, *51*, 2650−2665.
(3) Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0—A Public Library of Challenging Docking Benchmark Sets. *J. Chem. Inf. Model.* **2013**, *53*, 1447−1462.
(4) Jain, A. N. Bias, reporting, and sharing: Computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201−212.
(5) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.
(6) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Model.* **2004**, *44*, 793−806.
(7) Mysinger, M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E) - Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582.
(8) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(9) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein−ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856−5868.
(10) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46*, 2895−2907.
(11) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.
(12) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(13) Lagarde, N.; Zagury, J. F.; Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *J. Chem. Inf. Model.* **2015**, *55*, 1297−1307.
(14) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **2002**, *322*, 339−355.
(15) 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.
(16) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Model.* **2001**, *41*, 1395−1406.
(17) 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.
(18) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141−146.
(19) Ibrahim, T. M.; Bauer, M. R.; Boeckler, F. M. Applying DEKOIS 2.0 in Structure-Based Virtual Screening to Probe the Impact of Preparation Procedures and Score Normalization. *J. Cheminf.* **2015**, *7*, 21.
(20) Mackey, M. D.; Melville, J. L. Better than random? The chemotype enrichment problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154−1162.
(21) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(22) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 529−536.
(23) 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.
(24) 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.
(25) Ertl, P. Intuitive ordering of scaffolds and scaffold similarity searching using scaffold keys. *J. Chem. Inf. Model.* **2014**, *54*, 1617−1622.
(26) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem.* **2004**, *47*, 6144−6159.
(27) 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.
(28) *Pipeline Pilot, V6.1.5.0*, Student Edition; Accelrys: San Diego, CA, 2007.
(29) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer−Verlag: Heidelberg, Berlin, 2007.
(30) Watzig, H.; Oltmann-Norden, I.; Steinicke, F.; Alhazmi, H. A.; Nachbar, M.; El-Hady, D. A.; Albishri, H. M.; Baumann, K.; Exner, T.; Bockler, F. M.; El Deeb, S. Data quality in drug discovery: the role of analytical performance in ligand binding assays. *J. Comput.-Aided Mol. Des.* **2015**, DOI: 10.1007/s10822-015-9851-6.
(31) Gossett, L. S.; Habeck, L. L.; Gates, S. B.; Andis, S. L.; Worzalla, J. F.; Schultz, R. M.; Mendelsonn, L. G.; Kohler, W.; Ratnam, M.; Grindey, G. B.; Shih, C. A. Synthesis and biological evaluation of a new series of dihydrofolate reductase inhibitors based on the 4-(2,6-diamino-5-pyrimidinyl)alkyl-L-glutamic acid structure. *Bioorg. Med. Chem. Lett.* **1996**, *6*, 473−476.
(32) Cody, V.; Luft, J. R.; Pangborn, W.; Gangjee, A.; Queener, S. F. Structure determination of tetrahydroquinazoline antifolates in complex with human and *Pneumocystis carinii* dihydrofolate reductase: correlations between enzyme selectivity and stereochemistry. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 646−655.
(33) Feher, M.; Williams, C. I. Numerical errors and chaotic behavior in docking simulations. *J. Chem. Inf. Model.* **2012**, *52*, 724−738.
(34) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Model.* **2003**, *43*, 267−272.
(35) Carta, G.; Knox, A. J.; Lloyd, D. G. Unbiasing scoring functions: A new normalization and rescoring strategy. *J. Chem. Inf. Model.* **2007**, *47*, 1564−1571.
(36) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: A useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430−431.
(37) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(38) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein−ligand docking programs is difficult. *Proteins: Struct., Funct., Genet.* **2005**, *60*, 325−332.

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307

(39) *JChem, Library MCS, V0.7*; Chemaxon: Budapest, Hungary, 2010.
(40) Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminf.* **2009**, *1*, 12.
(41) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the chemistry development kit (CDK)—An open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111−2120.
(42) 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.
(43) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(44) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.
(45) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(46) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(47) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(48) *Glide, V5.6*; Schrödinger, LLC: New York, 2010.

DOI: 10.1021/acs.jcim.5b00475
J. Chem. Inf. Model. 2015, 55, 2297−2307