JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

An Unbiased Method To Build Benchmarking Sets for Ligand-Based Virtual Screening and its Application To GPCRs

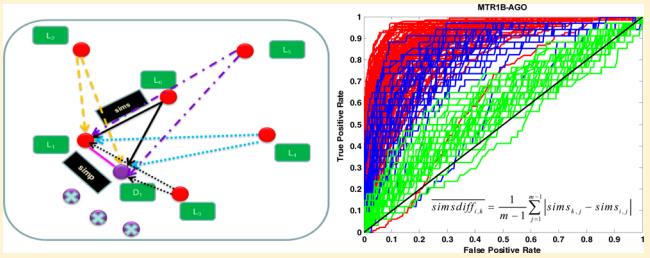
Jie Xia,[†,‡] Hongwei Jin,[†] Zhenming Liu,[†] Liangren Zhang,*[,†] and Xiang Simon Wang*[,‡]

[†]State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing 100191, China

[‡]Molecular Modeling and Drug Discovery Core for District of Columbia Developmental Center for AIDS Research (DC D-CFAR), Laboratory of Cheminfomatics and Drug Design, Department of Pharmaceutical Sciences, College of Pharmacy, Howard University, Washington, DC 20059, United States

Supporting Information

ABSTRACT: Benchmarking data sets have become common in recent years for the purpose of virtual screening, though the main focus had been placed on the structure-based virtual screening (SBVS) approaches. Due to the lack of crystal structures, there is great need for unbiased benchmarking sets to evaluate various ligand-based virtual screening (LBVS) methods for important drug targets such as G protein-coupled receptors (GPCRs). To date these ready-to-apply data sets for LBVS are fairly limited, and the direct usage of benchmarking sets designed for SBVS could bring the biases to the evaluation of LBVS. Herein, we propose an unbiased method to build benchmarking sets for LBVS and validate it on a multitude of GPCRs targets. To be more specific, our methods can (1) ensure chemical diversity of ligands, (2) maintain the physicochemical similarity between ligands and decoys, (3) make the decoys dissimilar in chemical topology to all ligands to avoid false negatives, and (4) maximize spatial random distribution of ligands and decoys. We evaluated the quality of our Unbiased Ligand Set (ULS) and Unbiased Decoy Set (UDS) using three common LBVS approaches, with Leave-One-Out (LOO) Cross-Validation (CV) and a metric of average AUC of the ROC curves. Our method has greatly reduced the "artificial enrichment" and "analogue bias" of a published GPCRs benchmarking set, i.e., GPCR Ligand Library (GLL)/GPCR Decoy Database (GDD). In addition, we addressed an important issue about the ratio of decoys per ligand and found that for a range of 30 to 100 it does not affect the quality of the benchmarking set, so we kept the original ratio of 39 from the GLL/GDD.



INTRODUCTION

G protein-coupled receptors (GPCRs) are a class of important proteins in cellular signal transduction and involved in many physiological functions and diseases.[1,2] They are thus considered to be promising targets for modern drug discovery[3] and have been targeted by ∼30−40% of marketed drugs.[4] In recent decades, huge efforts have been invested in understanding the structure and functions of GPCRs,[5−8] which facilitate the development of structure-based drug design (SBDD) on this type of target.[9] Although crystal structures of a limited number of GPCRs have been resolved,[10] those receptors only account for a notably small percent of over 800 GPCR members because it is challenging to conduct X-ray crystallographic studies of such membrane proteins.[3,11] Therefore, much of the efforts have to rely on ligand-based drug design (LBDD) approaches including 2D similarity searching,[12−14] pharmacophore modeling,[15−18] and predictive QSAR modeling.[19,20] Specifically, LBDD exploits the knowledge of the known ligands that bind to or act on the target rather than the structural information on macromolecular targets. It has been applied widely in GPCR-based drug discovery.[21−25]

Up to now, a variety of methods for LBDD have been developed while new methods are still emerging.[26−28] The objective evaluation of these methods becomes an important issue, since such an assessment can not only assist users to choose the reliable methods in their studies but also inspire developers to improve their methods as well.[29] In fact, this kind of benchmarking study has become common for *in silico* screening, especially in structure-based virtual screening (SBVS).[30−33] In those cases, the authors normally conducted retrospective small-scale virtual screening (VS) using the public or in-house benchmarking sets. In order to evaluate different methods in an accurate and impartial way, the quality of benchmarking sets proves to be rather crucial. In recent years, there have been a growing number of benchmarking sets developed by multiple research groups worldwide. Among them, the Directory of Useful Decoys (DUD) benchmarking sets provided by the Shoichet Laboratory (http://shoichetlab.compbio.ucsf.edu/) were widely used for validating novel methods or comparing different methods as they provide

Received: January 31, 2014
Published: April 21, 2014

   dx.doi.org/10.1021/ci500062f | J. Chem. Inf. Model. 2014, 54, 1433−1450

Table 1. Summary of GPCRs Ligand Data Sets Collected from GLL for This Study

GPCRs family	subclass	target	ligand type	label	no. of ligands
amine	serotonin	5HT1F	agonists	5HT1F-AGO	131
		5HT1F	antagonists	5HT1F-ANTA	11
	dopamine	DRD5	agonists	DRD5-AGO	11
		DRD5	antagonists	DRD5-ANTA	12
	histamine	HRH4	agonists	HRH4-AGO	11
		HRH4	antagonists	HRH4-ANTA	15
	muscarinic acetylcholine	ACM4	agonists	ACM4-AGO	15
		ACM4	antagonists	ACM4-ANTA	51
peptide	opioid	OPRM	agonists	OPRM-AGO	140
		OPRM	antagonists	OPRM-ANTA	27
	bombesin	BRS3	antagonists	BRS3-ANTA	17
	somatostatin	SSR2	antagonists	SSR2-ANTA	25
	angiotensin	AG22	antagonists	AG22-ANTA	32
prostanoid	prostaglandin	PE2R3	agonists	PE2R3-AGO	16
		PE2R3	antagonists	PE2R3-ANTA	125
melatonin	melatonin	MTR1B	agonists	MTR1B-AGO	135
		MTR1B	antagonists	MTR1B-ANTA	24

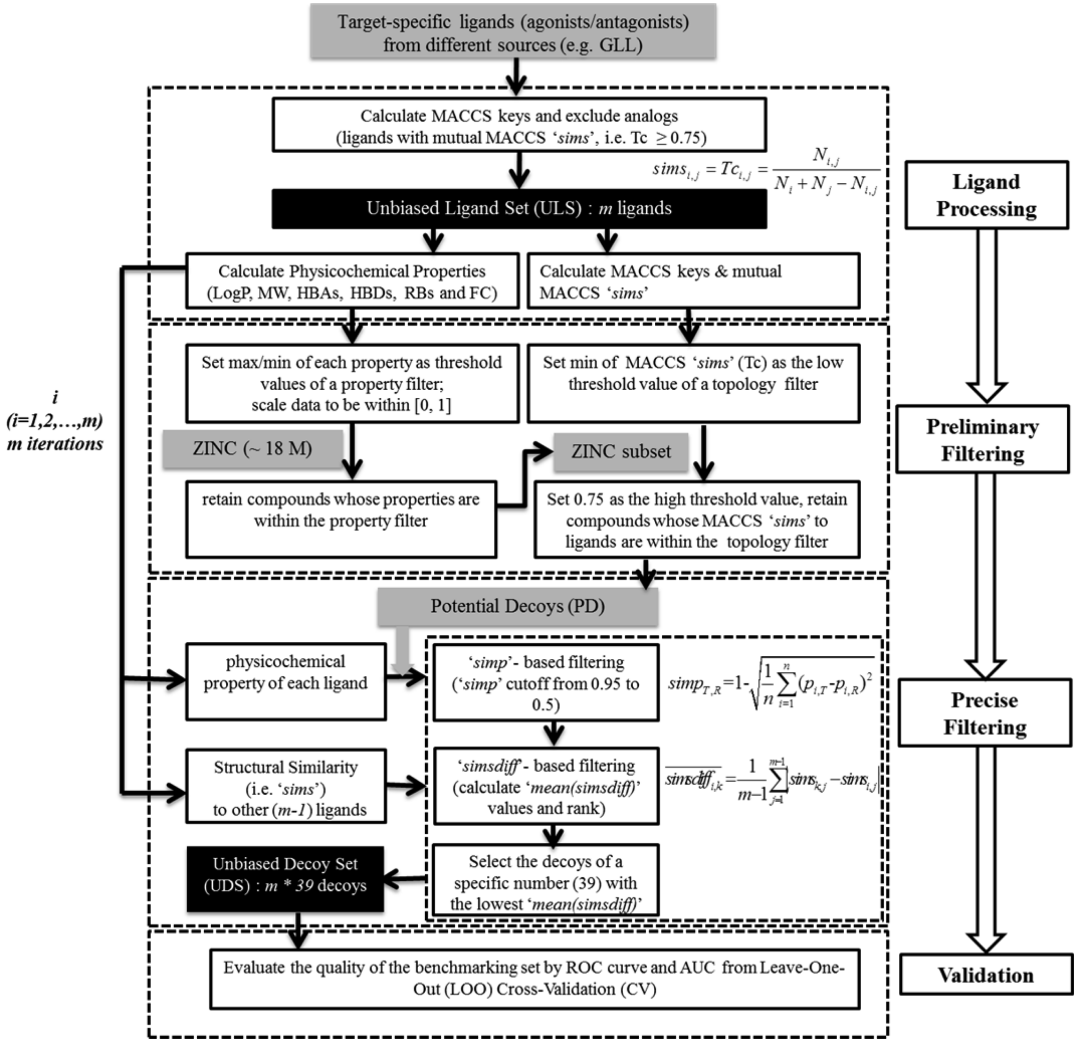
challenging but fair data sets.[31,33−35] Its first version was released by Huang et al.[36] in 2006, and its enhanced version DUD-E was released in 2012.[29] In addition to DUD/DUD-E, the maximum unbiased validation (MUV) data sets were recently developed based on PubChem Bioactivity data[37] using the refined nearest neighbor analysis originated from spatial statistics.[38] In 2011, Wallach and Lilien developed an algorithm to compile benchmarking virtual decoy sets (VDS) to enlarge the chemical space. They proved that VDS displays a similar quality to DUD,[39] though there exist concerns about the synthetic feasibility. The GPCR ligand library (GLL) and GPCR Decoy Database (GDD) were recently compiled with the focus on evaluating molecular docking methods for GPCR drug discovery.[40] The demanding evaluation kits for objective *in silico* screening (DEKOIS) was designed for benchmarking docking programs and scoring functions.[41] More recently, Cereto-Massague et al.[42] developed DecoyFinder for building target-specific decoy sets, which used the same algorithm as for DUD.

Depending on the initial purpose, e.g., SBVS or LBVS, the benchmarking sets are normally developed by relevant methods and can only be used for that purpose. From the beginning of the above-mentioned benchmarking efforts, the main focus has been on the evaluation of SBVS approaches, in particular molecular docking. Unfortunately, the application of these ready-to-apply data sets to ligand-based virtual screening (LBVS) is restricted because they normally include limited targets whose crystal structures are available. Until now there are only three benchmarking sets that can be directly employed for LBVS, i.e., MUV, REPROVIS-DB, and DUD LIB VS 1.0. The database of reproducible virtual screens, i.e., REPROVIS-DB, was compiled with data from prior LBVS applications including reference compounds, screening databases, compound selection criteria, and experimentally confirmed hits.[43] Although there are general tools to build decoy sets for those targets not included in the benchmarking sets above, they are not suitable in nature for LBVS, especially for GPCRs targets. As reported, the DUD-E decoy generating tools and DecoyFinder are specially designed for the evaluation of docking methods. The MUV can generate decoys for LBVS, but this method had not been validated on biological targets outside of the PubChem database. Therefore, there is a great need to design novel algorithms to build benchmarking sets for LBVS and validating them on important targets such as GPCRs.

As discussed in the prior studies, there are three critical issues to address in evaluating the quality of benchmarking sets, i.e., "artificial enrichment," "false negative," and "analogue bias."[41,44−46] Artificial enrichment is caused when the ligands differ significantly from the decoys in low-dimension vector space of physicochemical properties or molecular topologies.[44] As Rohrer and Baumann pointed out, in this case, to differentiate ligands from decoys actually relies on the obvious dissimilarities between them rather than the performance of VS methods.[38] "False Negative" means the decoys that are supposed to be inactive against the target are proved later to be active by bioassay. This situation did appear in DUD.[45] In order to reduce this type of error, a strict criterion with a preset cutoff for structural dissimilarity is normally introduced. "Analogue bias" is another important issue, especially for LBVS, which can make the performance of LBVS over-optimistic and cause large bias.[38,45,47]The first method to address this bias was proposed by Clark and Webster-Clark, a weighting scheme based on the ROC metric following ligand clustering.[48]The second one, which was applied in DUD LIB V1.0[49] and DUD-E,[29] is the clustering of actives to enlarge chemical diversity. Rohrer and Baumann proposed the third one to utilize two cumulative distribution functions of distances, $G(t)$ for active−active distance and $F(t)$ for active-decoy distance, to make actives and decoys exhibit spatially random distribution, which was finally proved to be effective in lowering analogue bias and artificial enrichment.[38,46]

In this paper, we introduce our novel method to address the above issues which was mainly composed of three main strategies: (1) analogues excluding, (2) physicochemical properties-based strategy, including a preliminary target-specific property filter and '*similarity in properties*' ("*simp*")-based filtering, and (3) topology-based strategy, including a preliminary target-specific topology filter and our unique "*similarity in structure difference*" ("*simsdiff*")-based filtering. We applied this workflow to build Unbiased Ligand Set (ULS)/Unbiased Decoy Set (UDS) for 17 agonists/antagonists sets of 10 representative GPCR targets and carried out Leave-One-Out (LOO) Cross-Validation (CV) to evaluate the performance of ULS/UDS compared with GLL/GDD based on the metrics of

Scheme 1. Workflow for Construction of Unbiased Benchmarking Sets for LBVS



mean(ROC AUCs). To make a fair comparison, we employed "*simp*"-based VS validation, MACCS[50] "*similarity in structure*" ("*sims*")-based VS validation, as well as the topological similarity search using function class fingerprints of maximum diameter 6 (FCFP_6) fingerprint. We also explored the underlying mechanisms of reducing enrichment bias among our three strategies, i.e., analogues excluding, physicochemical properties-based filtering (mainly "*simp*"), and topology-based filtering (mainly "*simsdiff*"). In addition, we investigate the effect of the decoys/ligands ratio on the quality of ULS/UDS which is an important question that has not been addressed before. We anticipate that the benchmarking sets built by our workflow can be utilized for performance evaluation of different LBVS approaches in an unbiased manner.

METHODS

Source of Ligand Sets. All GPCR ligand sets were downloaded from the GLL/GDD Web site (http://cavasotto-lab.net/Databases/GDD/).[40] In GLL, there are 25 145 ligands (agonists and antagonists) for 147 human Class A Rhodopsin-like GPCRs targets. In fact, they were initially taken from the GLIDA database which collects data from the literature and various public Web sites.[51] Notably, those ligands in GLL had already been prepared with an appropriate protonation state at pH 7.0, the most probable tautomer and correct stereochemical forms. In our study, since our purpose is to prove the efficacy of our methodology, it is unnecessary to build decoy sets for all GPCRs targets. We chose 17 ligand sets for 10 representative GPCRs targets from GLL, and each set contains various numbers of agonists or antagonists ranging from 11 to 140. More targets were selected for major subclasses of amine GPCRs and peptide GPCRs, e.g., 5HT1F, DRD5, HRH4, and ACM4 to amine GPCRs while OPRM, BRS3, SSR2, and AG22 to represent peptide GPCRs. For minor subclasses, prostanoid (PE2R3) and melatonin (MTR1B) receptors were included. The detailed information about the ligand data sets is shown in Table 1. As for the source of decoys, they were taken from the ZINC database (http://zinc.docking.org/), which is a free database of commercially available compounds for virtual screening.[52] In our case, we downloaded all purchasable molecules (~18 million) from ZINC. The decoys in GDD for those targets were also downloaded for the purpose of comparison.

General Workflow to Construct ULS/UDS. The workflow of building the benchmarking ligand/decoy sets for a specific target is shown in Scheme 1. It is written based on Matlab (version 7.6.0.324) and Pipeline Pilot (version 7.5, Accelrys Software, Inc.) and consists of four consecutive steps, including ligand processing, preliminary filtering, precise filtering, and validation. The purpose of ligand processing is

to ensure chemical diversity of ULS, where the analogue excluding strategy is applied. Preliminary filtering is used to build target-specific Potential Decoys (PDs) in a fast way using two preliminary target-specific filters, i.e., property filter and topology filter. Precise filtering is the most critical component, which consists of "*simp*"-based filtering and "*simsdiff*"-based filtering. Specifically, the former is applied to reduce the "artificial enrichment," a common problem found in benchmarking sets for SBVS, while the latter is designed to reduce the "analogue bias" in LBVS. The validation as of the last step is to prove the efficacy of those strategies applied.

Ligand Processing. This step is to (1) collect all the agonists or antagonists for a specific target; (2) exclude those ligands with mutual MACCS "*sims*", i.e. ,Tanimoto coefficient[53] (Tc) $\geqslant$ 0.75, also called analogues, to build ULS; (3) calculate physicochemical properties of the ligands in ULS by using Pipeline Pilot, including LogP, Molecular Weight (MW), Number of Hydrogen Bond Acceptors (HBAs), Number of Hydrogen Bond Donors (HBDs), Number of Rotatable Bonds (RBs), and Formal Charge (FC); and (4) calculate MACCS structural keys for each ligand and mutual MACCS "*sims*" between ligands. The formula of "*sims*" is shown in eq 1:

$$sims_{i,j} = \mathrm{Tc}_{i,j} = \frac{N_{i,j}}{N_i + N_j - N_{i,j}} \tag{1}$$

In fact, its formula was directly taken from Tc, a common metric for topological similarity between two chemical compounds. In the formula, i is for the target compound and j is for the reference compound. N means the number of bits in the fingerprint. Therefore, $N_{i,j}$ represents the number of the bits in the fingerprints shared by compounds i and j, while N_i and N_j indicate the number of bits for compounds i and j, respectively.

Since the ligands are collected from the literature and various public databases, they normally contain too many analogues of the similar chemical scaffold, which results in low structural diversity. In our method, mutual MACCS "*sims*" values (Tc) for all the ligands are calculated to build a similarity matrix, followed by our customized scripts in Matlab to exclude analogues. The reasons to set the cutoff value to be 0.75 are as follows: Tc = 0.75 was defined as a cutoff to differentiate actives from inactives in GDD. To be more specific, compounds in ZINC that were topologically similar to the query ligand with the Tc $\geqslant$ 0.75 were regarded as actives. When GDD was built, those "active" compounds were excluded in order to reduce the false negative rate during screening.[40] For the same reason, Tc = 0.75 is applied as the maximum threshold when our potential decoy (PD) set is being built in the next preliminary filtering step. Because of these, the topological similarity values between all PDs and every ligand are less than 0.75. Under this situation, if (1) one ligand in the ligand set is left out as a query for similarity search and (2) many other ligands in the ligand set are similar to that query with Tc $\geqslant$ 0.75, it is obvious that those similar ligands are easy to retrieve against the background of PDs due to their high similarity to the query. To reduce this type of screening bias, those analogues are excluded.

Preliminary Filtering. At this step, PDs are obtained by our two preliminary filters based on the range of physicochemical properties and mutual topological similarity (MACCS "*sims*") to the ligands. First, the maximum and minimum values of each physicochemical property for all the ligands are set as a target-specific property filter. Next, all data of each physicochemical property are scaled linearly so that the minimum value becomes 0 and the maximum value is 1.0. Ensuingly, we set the minimum value of mutual MACCS "*sims*" from all compounds in ULS and the maximum of 0.75 to be a target-specific, topology filter. After physicochemical properties and MACCS "*sims*" to ligands in ULS for each ZINC compound are calculated, the original ZINC database is filtered by these two preliminary filters in order to enrich PDs effectively (reduce the size of PDs largely for the next step) while ensuring the physicochemical and topological similarity between PDs and all the ligands.

Precise Filtering. Here, we design two formulas for precise filtering to obtain the FDs for UDS. To ensure the good quality of final decoys, we generate decoys for each ligand individually. One precise filtering criterion is referred to as "*simp*," defined to describe the physicochemical difference between each ligand and its PDs as shown below:

$$simp_{\mathrm{T,R}} = 1 - \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_{i,\mathrm{T}} - p_{i,\mathrm{R}})^2} \tag{2}$$

p represents the scaled value of physicochemical property, n is the total number of physicochemical properties used for the calculation, and i is the index for individual property. T is for the target compound, and R is for reference compound; the "*simp*" represents the physicochemical similarity between target compound and reference compound. The other precise filtering criterion, i.e., "*simsdiff*", is defined in eq 3:

$$\overline{simsdiff_{i,k}} = \frac{1}{m-1}\sum_{j=1}^{m-1}|sims_{k,j} - sims_{i,j}| \tag{3}$$

m is the number of the ligands, i is the index for the query selected from the ligand set, ranging from 1 to m. $m - 1$ is the number of the ligands except for the query, thus the remaining ligands' index is set to j (from 1 to $m - 1$). The decoy index is set to k. $\overline{simsdiff_{i,k}}$ is used to record the average difference between two topological similarities, i.e. MACCS "*sims*", of which one is between the decoy k and remaining ligands j, and another is between the query i and the remaining ligands j. In this step, "*simp*" cutoff can be automatically updated for each run to make sure there are enough decoys, i.e., more than 39 for our FDs. Normally, we set it to be 0.95, but when fewer than 39 decoys are obtained, the value is decreased by 0.05 gradually until enough decoys are found. The decoys filtered by "*simp*"-based criteria are ranked according to "*simsdiff*" value, and the final 39 decoys at the top of the list are picked up. Ideally, "*simsdiff*" values for all the FDs are '0'. But since the chemical space of the ZINC database is limited, what we can do is to select the decoys with the lowest "*simsdiff*" values. When moving to the next ligand, we also make sure there are no duplicates in the new decoy set. After the corresponding decoys for each ligand are finalized, i.e. our UDS for LBVS, the whole benchmarking set for the specific target is constructed and ready for validation.

Validation. The LOO CV is applied to the retrospective similarity-based LBVS on our ULS/UDS. The LOO CV procedure is designed as follows. At each cycle, one ligand is moved out from ULS as a query, and its corresponding decoys are removed from the UDS as well. The remaining compounds of both ligands and decoys then constitute a screening set for internal validation purpose. All compounds are coded by six physicochemical properties and MACCS structural keys that were used in the early stage of our workflow, followed by the

Table 2. Metrics of Mean(ROC AUCs) from Leave-One-Out Cross-Validation Based on Similarity Search by Physicochemical Properties ("*simp*"), MACCS Keys ("*sims*") and External Validation by FCFP_6 Fingerprint

data set	GLL/GDD (*simp*)	eGLL/eGDD (*simp*)	ULS/UDS (*simp*)	GLL/GDD (*sims*)	eGLL/eGDD (*sims*)	ULS/UDS (*sims*)	GLL/GDD (FCFP_6)	ULS/UDS (FCFP_6)
5HT1F-AGO	0.557	0.474	0.491	0.797	0.717	0.554	0.801	0.651
5HT1F-ANTA	0.552	0.572	0.518	0.663	0.672	0.458	0.618	0.549
DRD5-AGO	0.508	0.498	0.436	0.713	0.680	0.552	0.720	0.662
DRD5-ANTA	0.526	0.527	0.452	0.658	0.622	0.531	0.665	0.590
HRH4-AGO	0.542	0.446	0.451	0.909	0.900	0.726	0.871	0.796
HRH4-ANTA	0.503	0.490	0.474	0.694	0.669	0.530	0.580	0.557
ACM4-AGO	0.504	0.490	0.467	0.665	0.638	0.506	0.590	0.515
ACM4-ANTA	0.516	0.486	0.486	0.629	0.592	0.498	0.669	0.662
OPRM-AGO	0.511	0.527	0.477	0.736	0.572	0.510	0.773	0.654
OPRM-ANTA	0.500	0.358	0.476	0.882	0.722	0.589	0.884	0.589
BRS3-ANTA	0.565	0.382	0.348	0.874	0.837	0.639	0.940	0.950
SSR2-ANTA	0.506	0.489	0.350	0.795	0.727	0.583	0.808	0.793
AG22-ANTA	0.533	0.580	0.415	0.830	0.876	0.694	0.803	0.705
PE2R3-AGO	0.630	0.653	0.367	0.941	0.931	0.728	0.938	0.745
PE2R3-ANTA	0.492	0.427	0.478	0.712	0.562	0.482	0.816	0.643
MTR1B-AGO	0.541	0.478	0.501	0.908	0.817	0.581	0.899	0.700
MTR1B-ANTA	0.555	0.543	0.484	0.873	0.858	0.598	0.833	0.720
Min	0.492	0.358	0.348	0.629	0.562	0.458	0.580	0.515
Max	0.630	0.653	0.518	0.941	0.931	0.728	0.940	0.950
Average	0.532	0.495	0.451	0.781	0.729	0.574	0.777	0.675

traditional similarity-based LBVS. On the basis of the ranked similarity values for the screening compounds and their observed activity values, i.e., 1 for ligands and 0 for decoys, we compute the ROC curves and their corresponding AUCs. This process is repeated m times if there are m ligands. In addition, we calculate the mean(ROC AUCs) as a metric to evaluate the quality of the benchmarking set. Similar metrics have been proposed and implemented previously by a couple of research groups, such as mean(ROC) in MUV[38] and the deviation from optimal embedding score (DOE score) in DEKOIS.[41] As a special ROC curve, the diagonal line $y = x$ indicates randomly assigning both classes, i.e., ligand or decoy. The AUC of this curve is 0.5 in this situation.[54] Accordingly, the enrichment curves moving toward the diagonal line (AUC = 0.5) indicate that those similarity-based LBVS approaches fail to distinguish ligands from decoys. In this way, ligand and decoy are in a random distribution around chemical spaces, and it meets our goal of reducing overoptimistic enrichment caused by artificial bias. Therefore, we deem ROC AUC = 0.50 to be the optimal embedding in the current workflow.

RESULTS AND DISCUSSION

Retrospective Similarity-Based LBVS Detects Analogue Bias in GLL/GDD. We encoded 17 representative data sets in GLL/GDD, eGLL/eGDD, and ULS/UDS with six physicochemical properties and MACCS structural keys and conducted retrospective LBVS based on calculated "*simp*" and MACCS "*sims*." Particularly, we designed the eGLL/eGDD set to be the intermediate after our "analogues excluding" strategy. As we know, there are two aspects that affect the screening performance: one is the composition of ligands, and another is the decoy building strategy. To make a fair comparison, we applied our "analogues excluding" strategy to the GLL as in the construction workflow of ULS/UDS and extract the decoys from the GDD accordingly. In this way, the eGLL/eGDD set contains the same composition of ligands as ULS/UDS. The results of the retrospective LBVS are shown in Table 2 and Figures 2−4. For "*simp*"-based VS, the average value of the metrics of mean(ROC AUCs)s for 17 GPCRs targets is at the same level (close to 0.50) for all three data sets, i.e., GLL/GDD, eGLL/eGDD, and ULS/UDS. In GLL/GDD, the minimum value is 0.492 and even the maximum value is only 0.630. In fact, for most of the GPCRs targets, the mean(ROC AUCs)s in eGLL/eGDD are similar to those of the other two sets. The average value is 0.495, and the range is from 0.358 to 0.653 (cf. Table 2, Figures 2 and 3). All three lines in Figure 2 (upper panel, "*simp*") are close to the random distribution curve for the majority of GPCRs targets, while fluctuating slightly over the same set of receptors such as BRS3-ANTA, AG22-ANTA, and PE2R3-AGO. Figure 3 shows more details about ROC curves from "*simp*"-based VS for 17 data sets. For most ROC curves in these plots, the red and blue curves of each iteration match well with the random distribution curve. However, for MACCS "*sims*"-based VS, the average value of mean(ROC AUCs)s in GLL/GDD is as high as 0.781 and fluctuates at the range from 0.629 to 0.941 (Table 2). Consistently, both the blue line (GLL/GDD) and red line (eGLL/eGDD) in Figure 2 are fairly distant from the line of random value. It is even more obvious from Figure 4 to observe that for most of the GPCRs targets, the ROC curves in red and blue are distant from random distribution curve. These results indicate that although GLL/GDD (eGLL/eGDD) reduced artificial enrichment significantly as represented by the ideal performance of "*simp*"-based VS (thus good for SBVS), there exists a large bias when topology-based similarity search is conducted with MACCS keys ("*sims*"). To be more specific,

dx.doi.org/10.1021/ci500062f | J. Chem. Inf. Model. 2014, 54, 1433−1450

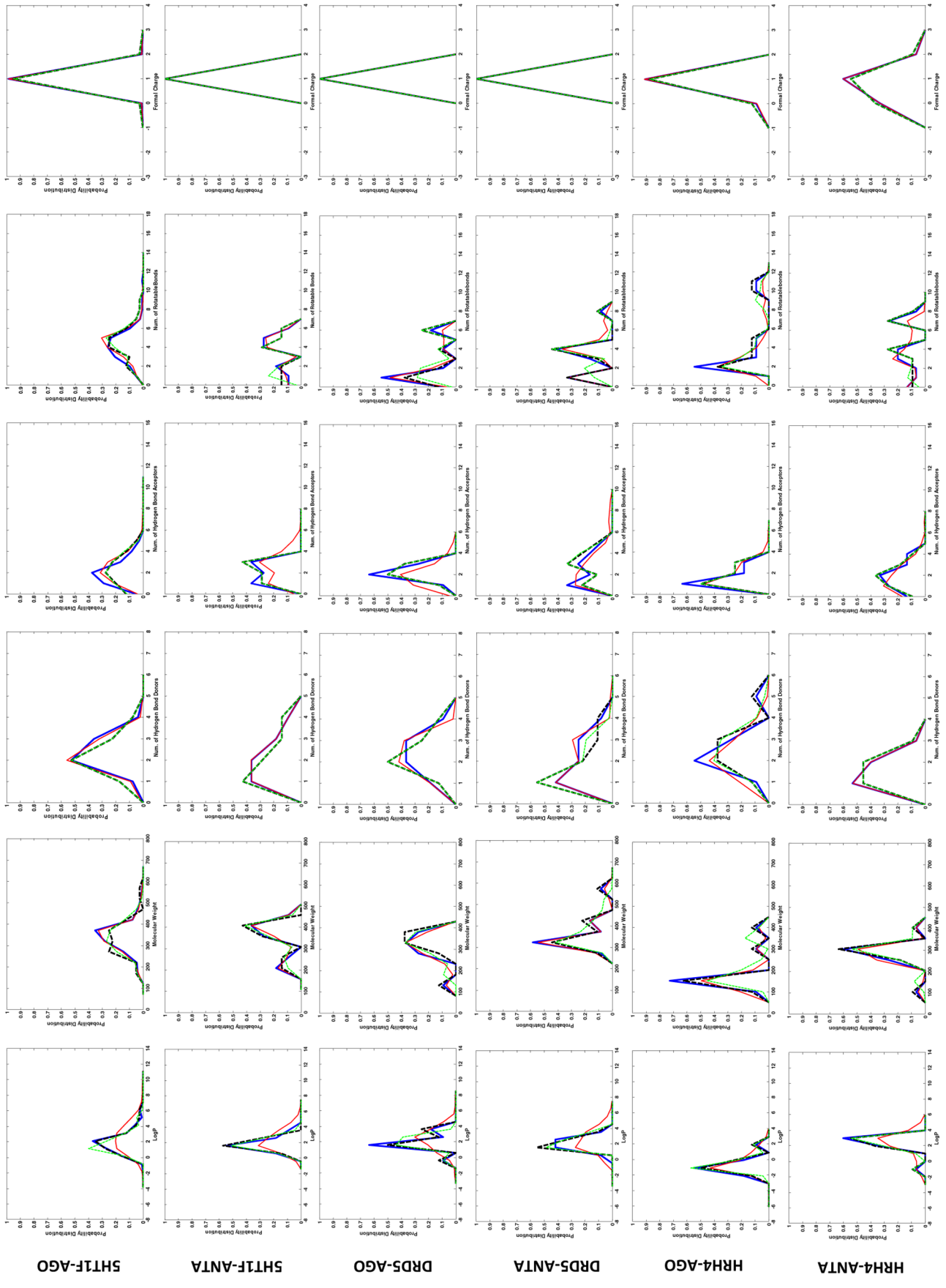


Figure 1. continued

dx.doi.org/10.1021/ci500062f | J. Chem. Inf. Model. 2014, 54, 1433−1450

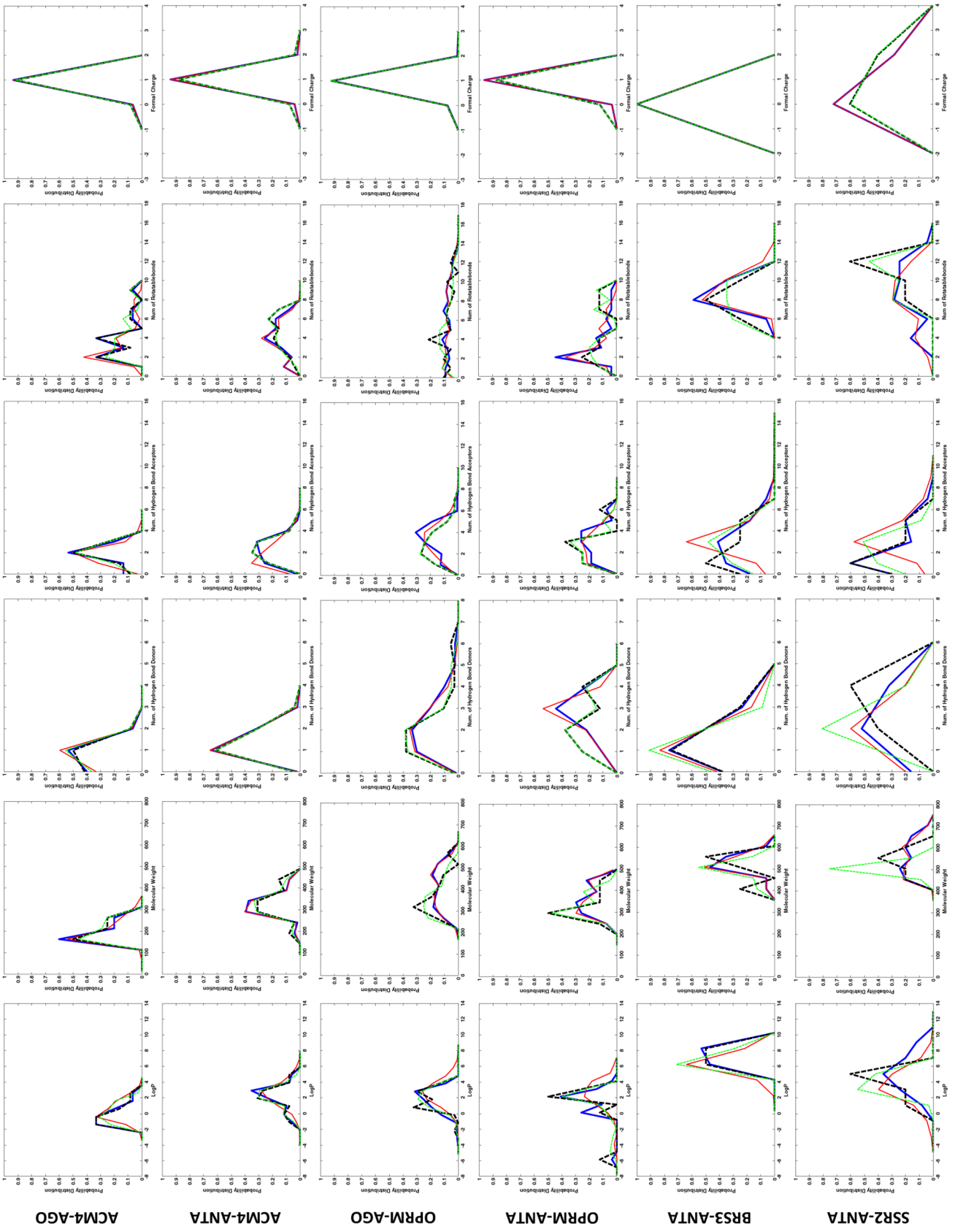


Figure 1. continued

dx.doi.org/10.1021/ci500062f | J. Chem. Inf. Model. 2014, 54, 1433−1450

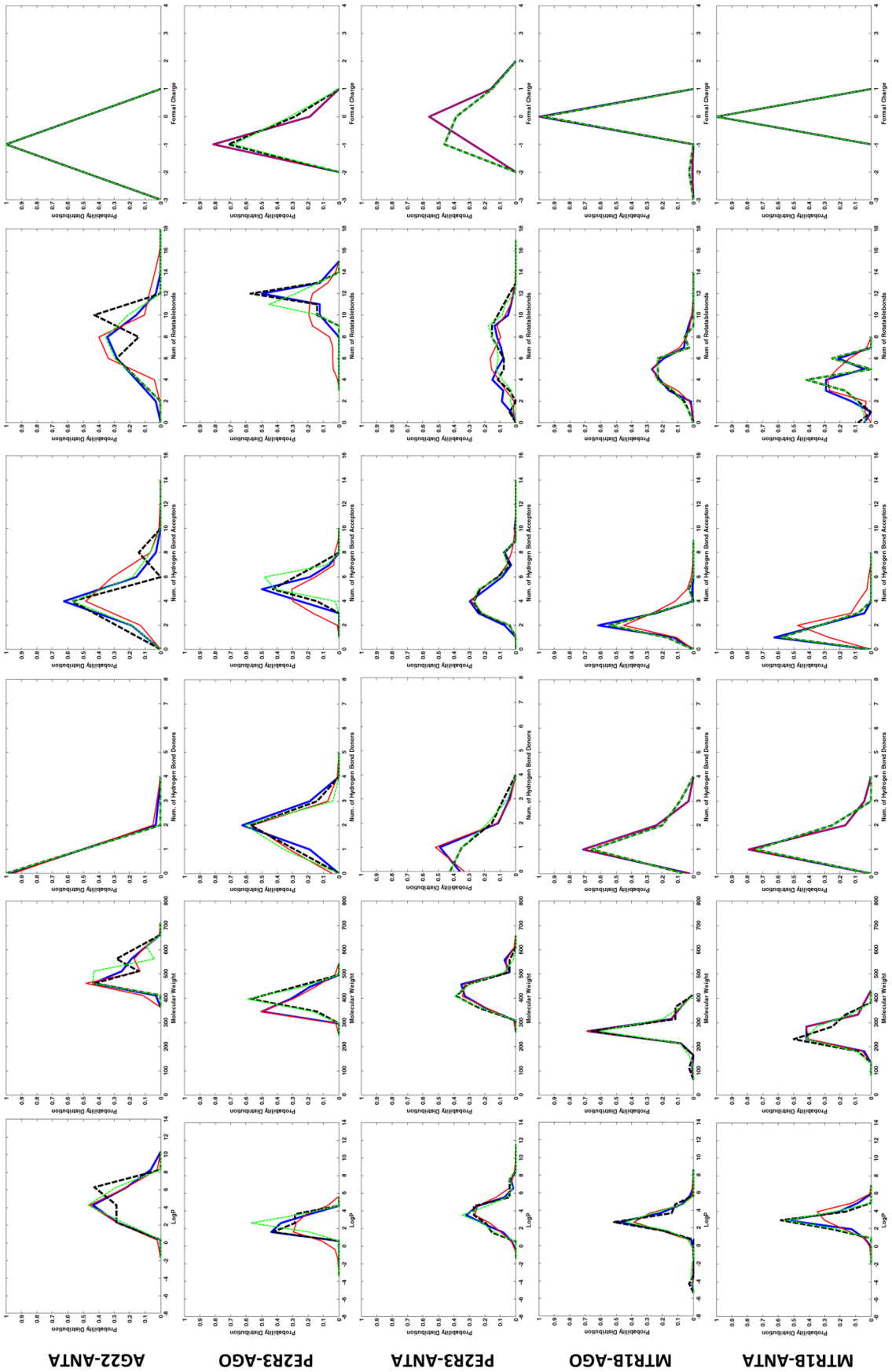


Figure 1. The physicochemical properties distributions of ligands and decoys in ULS/UDS and GLL/GDD for all 17 data sets. Color and sign: GLL, blue, full line; GDD, red, full line; ULS, black, dotted line; UDS, green, dotted line.

dx.doi.org/10.1021/ci500062f | J. Chem. Inf. Model. 2014, 54, 1433−1450

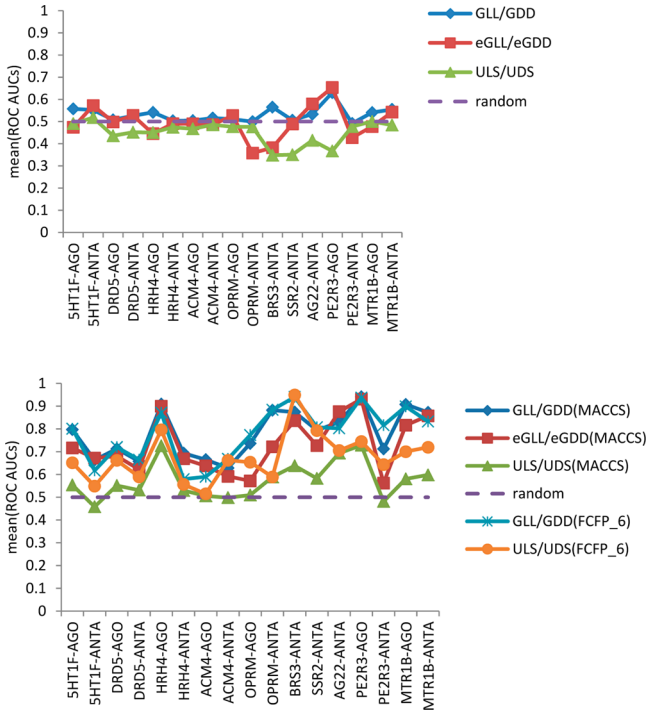


Figure 2. The performance of leave-one-out cross-validation in the metrics of mean(ROC AUCs) across 17 data sets from similarity search by physicochemical properties ("*simp*", upper panel), MACCS structural keys ("*sims*"), and FCFP_6 fingerprint (lower panel).

although each query's Euclidean distances of physicochemical properties to all decoys are close to its distances to other ligands, its chemical scaffold is quite different from decoys which makes the decoys rank low on the list. By contrast, other ligands to the same target are ranked high, thus becoming easy to identify. Therefore, a serious caveat exists for current standard, "*simp*"-based approaches to build the benchmarking decoy sets when applying to the problem of LBVS. Herein, we suggest adding a topology-based filtering strategy which takes into consideration the topological similarities not only between the query and its decoys, but also between its decoys with other ligands as well.

As we mentioned before, all the mean(ROC AUCs)s across 17 data sets in ULS/UDS are smaller for "*simp*"-based VS when compared with GLL/GDD. Among them, results for 5HT1F antagonists, 5HT1F agonists, MTR1B antagonists, and MTR1B agonists are the closest to 0.50, a value for random distribution. However, for PE2R3 agonists, BRS3 antagonists, SSR2 antagonists, and AG22 antagonists, their values appear to be distant to 0.50 (cf. Table 2). Apparently, there are more decoys ranking at the top of the list with high physicochemical similarity in these cases. The reason for this is likely that not enough compounds in the ZINC database meet filtering criteria for both "*simp*" and "*simsdiff*" for a certain number of ligands due to the limited chemical space in the database itself. Notably, for the same PE2R3 agonists, SSR2 antagonists, and AG22 antagonists, the value of mean(ROC AUCs)s changes slightly from GLL/GDD to eGLL/eGDD but substantially from eGLL/eGDD to ULS/UDS. Therefore, we think the decrease in values from GLL/GDD to ULS/UDS is mainly caused by our physicochemical properties-based and topology-based filtering strategies we applied because of the same composition of eGLL to ULS. For example, for PE2R3 agonists the value goes from 0.630 in GLL/GDD to 0.367 in ULS/UDS. This type of "antiscreening" phenomenon, i.e. the ROC AUCs fall below 0.50, is actually a situation that always happens in the real practice of virtual screening. In this scenario, the ratio of actives in a chemical library is normally lower than usual and there exist "false positive" molecules which rank high but are inactive in themselves. To recognize this type of molecule and lower the value of FPR ("1 − specificity") at the *x*-axis remains to be one of the major tasks of virtual screening methods.[55] In this sense, these kinds of data sets with mean(ROC AUCs) below 0.50 in our ULS/UDS are acceptable as it poses the challenge to current methods and will facilitate their advancement.

Our Workflow Makes ULS/UDS Unbiased Measured by "*simp*"- and MACCS "*sims*"-based VS. As demonstrated in its original article and our current study, the GLL/GDD has already achieved a good level for "*simp*"-based VS thus is useful to SBVS as well since physicochemical properties of ligands do play an important role in many scoring functions. This implies that the GLL/GDD methodology is an effective way to reduce artificial enrichment for SBVS, but not necessarily for LBVS. In our current workflow, we keep the physicochemical properties-based strategy but add topology-based mechanisms in order to (1) exclude analogues in the ligand set and (2) exclude decoys that do not meet our filtering criteria defined by the preliminary target-specific topology filter and the *simsdiff* filter. Through these strategies, we achieved our goal in our ULS/UDS benchmarking set as shown below. In the third and sixth columns in Table 2, the average value of mean(ROC AUCs) across 17 GPCRs targets for "*simp*"-based VS is 0.451, while the value for MACCS "*sims*"-based VS is 0.573. In Figure 2, the green lines (ULS/UDS) show the small difference from the random line of 0.5. These results indicate it is challenging to differentiate the ligands from decoys in our ULS/UDS using either "*simp*"-based or MACCS "*sims*"-based VS, thus ideal to evaluate the approaches of LBVS. In comparison to GLL/GDD, the average value of mean(ROC AUCs) in ULS/UDS was reduced significantly from 0.781 to 0.574 for MACCS "*sims*"-based VS. Depending on various GPCRs targets, the decreasing rate of mean(ROC AUCs) ranges from 36.00% (MTR1B-AGO) to 16.36% (AG22-ANTA). We also observe the significant differences between red curves (GLL/GDD) and green curves (ULS/UDS) in Figure 4.

External Validation by FCFP_6 Fingerprint Shows the Improvement by ULS/UDS. Since our workflow employs physicochemical properties and MACCS structural keys during the construction of ULS/UDS, it becomes necessary to verify their performance using other fingerprints as an independent validation. We employed the FCFP_6 fingerprint for this purpose because of its proven accuracy in recent years.[56] The results are collected in Table 3 and Figures 2 and 5 as well to make the comparison with GLL/GDD. The mean(ROC AUCs) of ULS/UDS is smaller than its corresponding value of GLL/GDD across all 17 data sets, and its average drops by 12.65% (0.675 vs 0.777). Consistently, most ROC curves in green (ULS/UDS) are below ROC curves in red (GLL/GDD). These data indicate that similar to the prior two LBVS approaches ("*simp*"- and "*sims*"-based), the bias in enrichment has been reduced largely in our data sets. It is especially true for the data set of OPRM-ANTA, in which the mean(ROC AUCs) falls to 0.589 (ULS/UDS) from 0.884 (GLL/GDD). Interestingly, for the same data sets such as ULS/UDS the values of mean(ROC AUCs) are higher for FCFP_6 fingerprint

dx.doi.org/10.1021/ci500062f | J. Chem. Inf. Model. 2014, 54, 1433−1450

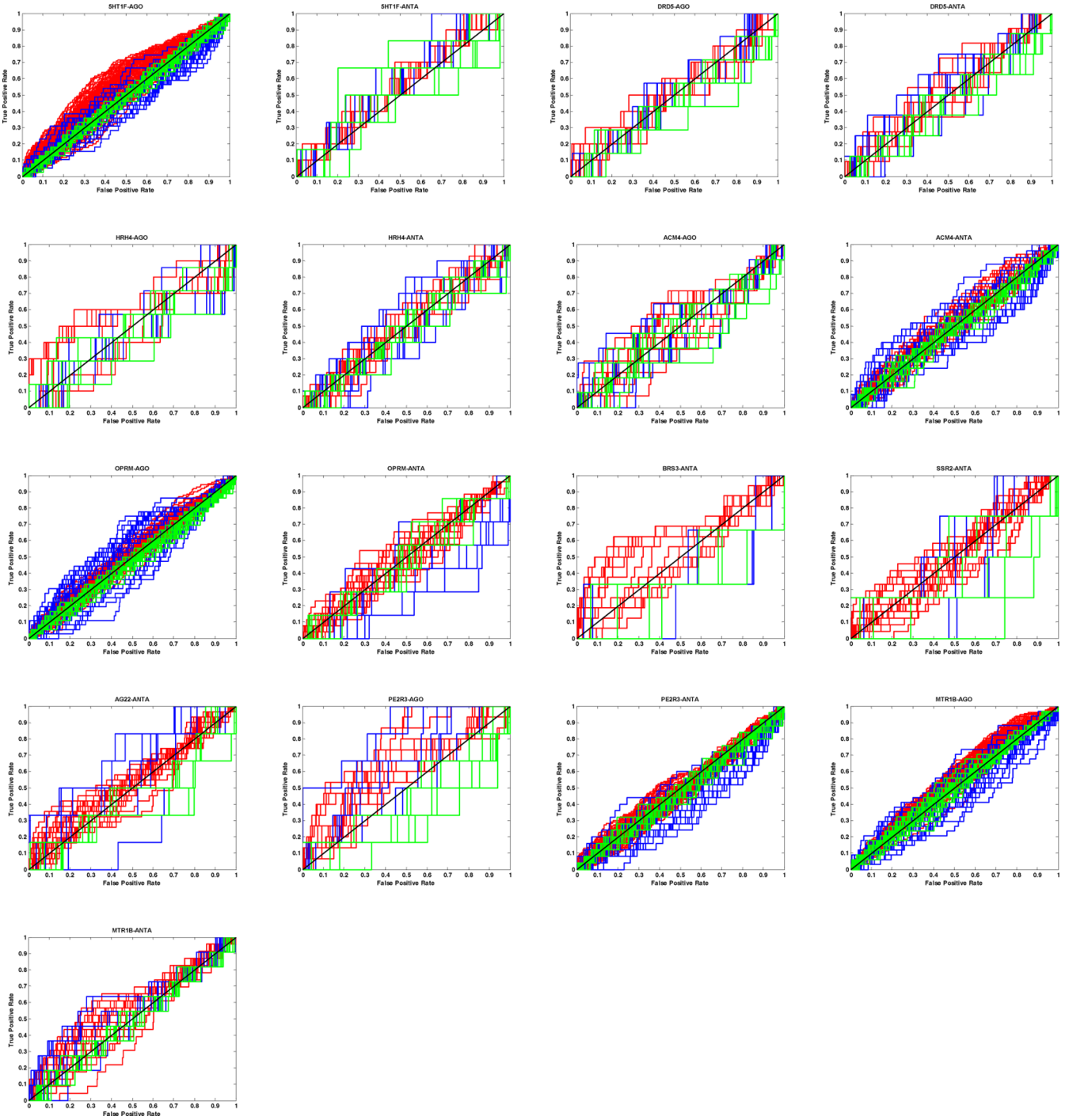


Figure 3. The ROC curves from similarity search by physicochemical properties ("*simp*") for all 17 data sets. For each data set, the curves are colored in red for GLL/GDD, blue for eGLL/eGDD, and green for ULS/UDS, respectively. The multiple curves in the same color represent different iterations in LOO CV for the specific benchmarking set, while the diagonal line in black shows the random distribution.

than MACCS keys consistently across all targets, suggesting that there exist certain systemic reason(s) derived from fingerprints themselves. On the other hand, the values of mean(ROC AUCs) for GLL/GDD with MACCS are similar to those with FCFP_6 (the average value is 0.781 vs 0.777). We can observe the similar trends from Figure 2, lower panel, in which the dark blue line (MACCS on GLL/GDD) comes close to the light blue line (FCFP_6 on GLL/GDD) while the line of FCFP_6 is above the line of MACCS keys based on our benchmarking set. In the future, we plan to employ additional LBVS approaches to check if it is a common observation and explore its implication to real screening.

The Underlying Mechanisms of Reducing Enrichment Bias. As mentioned before, we employed three major strategies in our workflow. Among them, the physicochemical properties-based (mainly "*simp*") strategy has been proved to be effective in randomly sampling and matching in properties for ligands and decoys. And it had been widely utilized in the generation of DUD, DUD-E, and GLL/GDD benchmarking data sets. To locate the exact mechanism(s) of reducing enrichment bias by

dx.doi.org/10.1021/ci500062f | J. Chem. Inf. Model. 2014, 54, 1433−1450

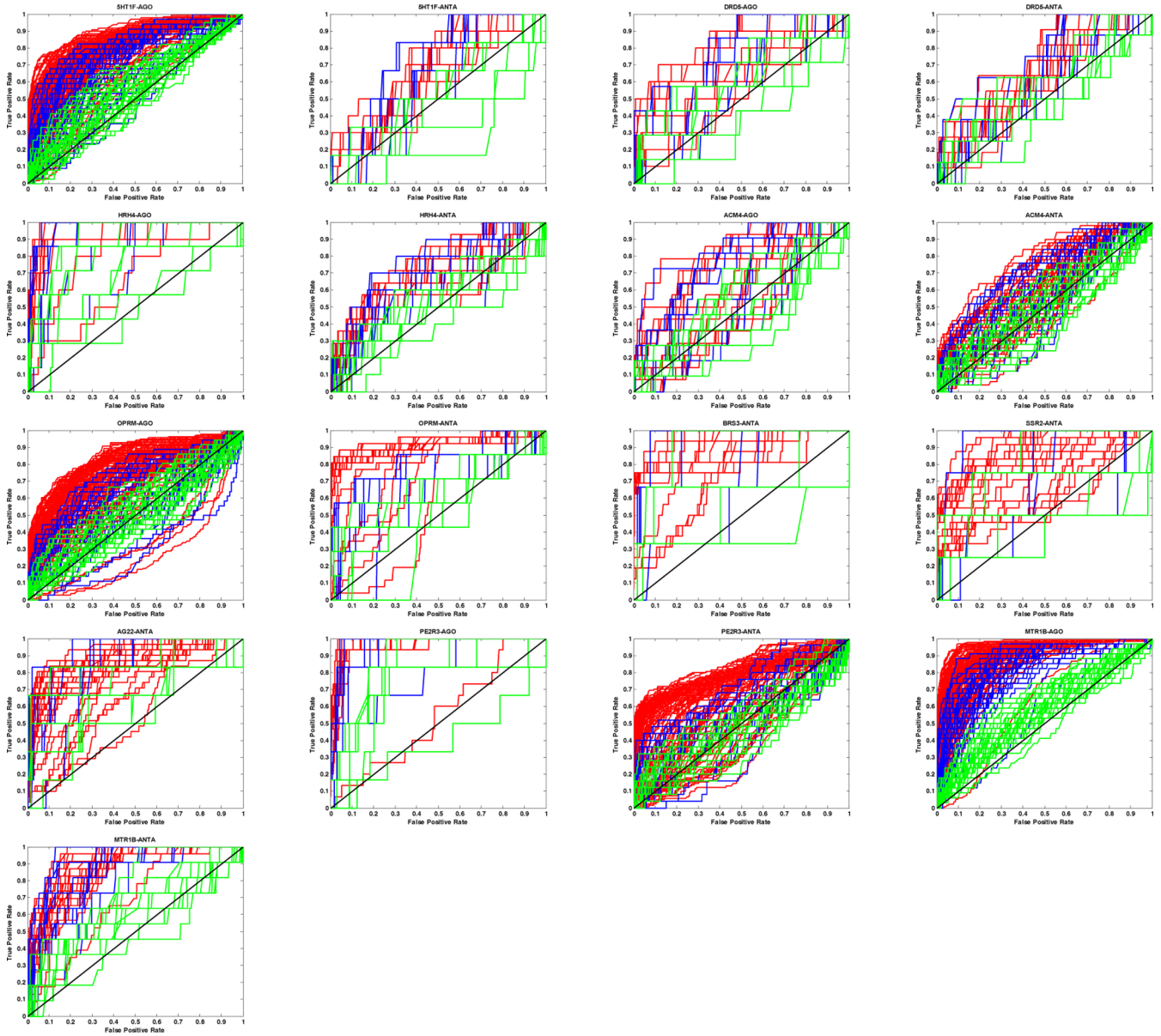


Figure 4. The ROC curves from similarity search by MACCS structural keys ("*sims*") for all 17 data sets. For each data set, the curves are colored in red for GLL/GDD, blue for eGLL/eGDD, and green for ULS/UDS, respectively. The multiple curves in the same color represent different iterations in LOO CV for the specific benchmarking set, while the diagonal line in black shows the random distribution.

our method, we analyzed in detail our two other strategies, i.e. the analogues excluding and topology-based strategy (a preliminary target-specific topology filter and "*simsdiff*"-based filtering). For the first one, the differences in composition between GLL and ULS may be enough to lower the enrichment bias. To test this possibility, we employed eGLL/eGDD for the comparison. Table 3 collects mean(ROC AUCs) values for 17 data sets and their statistical data for both GLL/GDD and eGLL/eGDD. In fact, the values are not so different between these two groups. The average mean(ROC AUCs) does not decrease largely, only with a small change from 0.781 to 0.729. The maximal value goes down from 0.941 to 0.931, while the minimal value goes from 0.629 to 0.562. In general, the average of decreasing rate is around 6.76%. In some cases such as 5HT1F antagonists and AG22 antagonists, the mean(ROC AUCs)'s increase by 1.36% and 5.58%, respectively. Nevertheless, there are cases with significant changes after excluding, for example, PE2R3 antagonists and OPRM agonists whose decreasing rates are greater than 20%. We postulate that it might be related to the exclusion of highly similar ligands, which leads to a higher excluding ratio. To confirm it, we plot the relationship between excluding ratio and decreasing rate from GLL/GDD to eGLL/eGDD (cf. Figure 6). Basically, the effect of analogues excluding to lower the enrichment bias is more obvious when the excluding ratio is above 0.70. In this region, most data sets show the decreasing rate above 10% (cf. Figure 6 and detailed data are in Table S1). Therefore, the effect of analogues excluding does exist but is limited to lower enrichment bias for LBVS. Because the difference between eGLL/eGDD and ULS/UDS mainly lies in our topology-based filtering strategies, i.e. a preliminary target-specific topology filter and "*simsdiff*"-based filtering, their powers are reflected apparently by comparing the columns in Table 2. The purpose of the preliminary topology filter is not

dx.doi.org/10.1021/ci500062f | *J. Chem. Inf. Model.* 2014, 54, 1433−1450

Table 3. Comparison of GLL and ULS in Term of Chemical Diversity

	no. of compds		no. of scaffolds		compound/scaffold ratio	
data set	GLL	ULS	GLL	ULS	GLL	ULS
5HT1F-AGO	131	40	74	34	1.77	1.18
5HT1F-ANTA	11	7	10	7	1.10	1.00
DRD5-AGO	11	8	10	8	1.10	1.00
DRD5-ANTA	12	9	11	9	1.09	1.00
HRH4-AGO	11	8	4	4	2.75	2.00
HRH4-ANTA	15	11	14	11	1.07	1.00
ACM4-AGO	15	12	9	9	1.67	1.33
ACM4-ANTA	51	26	42	26	1.21	1.00
OPRM-AGO	140	37	84	35	1.67	1.06
OPRM-ANTA	27	8	22	7	1.23	1.14
BRS3-ANTA	17	4	11	4	1.55	1.00
SSR2-ANTA	25	5	18	5	1.39	1.00
AG22-ANTA	32	7	20	7	1.60	1.00
PE2R3-AGO	16	7	6	5	2.67	1.40
PE2R3-ANTA	125	26	82	21	1.52	1.24
MTR1B-AGO	135	35	54	27	2.50	1.30
MTR1B-ANTA	24	12	15	9	1.60	1.33
min					1.07	1.00
max					2.75	2.00
average					1.62	1.18

only to eliminate the possibility of a "false negative" in the decoy sets but also help lower analogue bias, while our novel formula of "*simsdiff*" addresses directly the problem of enrichment bias. Generally speaking, across all 17 data sets the average mean(ROC AUCs)'s ("*sims*"-based VS) decreases to 0.574 from 0.729 aided by our method. In comparison, with only analogues excluding and simple physicochemical properties-based filtering in eGLL/eGDD, the enrichment bias still exists as their mean(ROC AUCs)s values stay distant from 0.50 for the majority of data sets. In some cases such as PE2R3 antagonists and OPRM agonists whose mean(ROC AUCs)'s are lowered to nearly 0.50, we consider this as a coincidence because in the process of building GDD they followed the principle of "first come, first served" and did not consider the effect of topological similarity in the ligand sets. In addition, the maximum value of mean(ROC AUCs) is 0.728 and the minimum value is 0.458 for ULS/UDS sets ("*sims*"-based VS), while the maximum/minimum values are as high as 0.931 and 0.562 in eGLL/eGDD. In particular, the data sets with the best performance are 5HT1F agonists, 5HT1F antagonists, MTR1B agonists, MTR1B antagonists, OPRM agonists, and OPRM antagonists. Their mean(ROC AUCs)'s by "*sims*"-based VS are also very close to 0.50. The green curves in Figure 3 also show the good quality of these data sets. Nevertheless, from Table 2 and Figures 2 and 4, we can see there are exceptions in ULS/UDS whose values are above 0.70 but still below the high values of eGLL/eGDD. For example, the decreasing rate for HRH4 agonists is 20.17% and 22.60% for PE2R3 agonists. After analysis, we find most decoys for these two targets have high values of "*simsdiff*" (above 0.10), which make the ROC curves distant from the random distribution level (cf. Tables S2 and S3). Therefore, our method is also restricted by limited chemical space of ZINC like other benchmarking data sets. In summary, these data prove that our topology-based filtering strategies (preliminary topology filter plus the "*simsdiff*" filter) contribute to the effect of lowering enrichment bias for LBVS more than analogues excluding.

Although our methods can achieve good results for the current data sets, there are several issues that need to be addressed in the sequel studies. First, for the problem of "false negative" in decoys, the authors of MUV and DUD-E proposed to include only true inactives that had been experimentally validated.[29,38] Since it is not possible to obtain enough real inactives for all the targets, we follow many groups to adopt a fingerprint-based Tc value (i.e., 0.75 for MACCS keys) as the cutoff to differentiate actives and inactives.[41] Second, although ZINC is an ideal source of decoys, its limited chemical space restricts our method in obtaining proper decoys for some specific targets, i.e. the decoy sets for HRH4 agonists and PE2R3 agonists. In VDS, the authors tried to create virtual decoys to enlarge the chemical space, which may be a good alternative but needs to be further justified for LBVS.[39] Third, to make it comparable with GLL/GDD, we only include six drug-like physicochemical properties though there were recommendations to use more.[45]

Physicochemical Properties Distributions of GLL/GDD and ULS/UDS. Similar to the prior publications[29,40] on benchmarking sets (DUD, DUD-E, GLL/GDD, etc.), we employed property distribution to check the match between ligand set and their decoys for all 17 GPCRs targets. From the plots in Figure 1, we can conclude that our UDS approximates to ULS closely for most targets in all six physicochemical properties, i.e., logP, MW, HBAs, HBDs, RBs, and FC. For example, in the data sets for the 5HT1F antagonists, 5HT1F agonists, MTR1B antagonists, and MTR1B agonists, the property distribution curves for ligands and decoys of ULS/GDD match closely, similar to or better than those in GLL/GDD, which is consistent with the results from "*simp*"-based VS (cf. Figure 2, upper panel). These examples are to demonstrate that our workflow affords a comparable property-matching ability to the GDD methodology, which explains the similar results for both benchmarking sets. However, to some targets like we mentioned before, i.e. PE2R3 agonists, BRS3 antagonists, SSR2 antagonists, and AG22 antagonists, the curves do not match tightly for our ULS/UDS. We have attributed this to two reasons in the prior paragraph. For BRS3 antagonists, the property distribution curve indeed proves our discussed point. At the graphs of logP, MW, and HBA in which the curve profile of ULS does not fit well to that of UDS, GLL also does not match GDD in that aspect. Interestingly, the curves of ULS overlap with the ones of GLL while the curves of UDS fit to the ones of GDD. This indicates that both methods cannot locate enough decoys that meet the "*simp*"-based criteria, caused by the limitation in chemical space of the original database. In fact, it is a common problem that occurred to DUD, DUD-E as well. The similar observation also exists in some cases of SSR2-ANTA, PE2R3-AGO, and other targets. In summary, the benchmarking set of ULS/UDS we built can be an alternative to GLL/GDD to evaluate docking methods to GPCRs.[44]

Scaffold Analysis of GLL vs ULS Shows that Our Analogues Excluding Improves Chemical Diversity. Scaffold analysis was conducted for 17 GPCRs target sets in GLL to estimate the chemical diversity in this published database.[40,53] For this analysis, we generated Murcko frameworks[57] using the Generate Fragments component in Pipeline Pilot to count the unique molecular scaffolds. In this component, the "Fragments To Generate" parameter was set

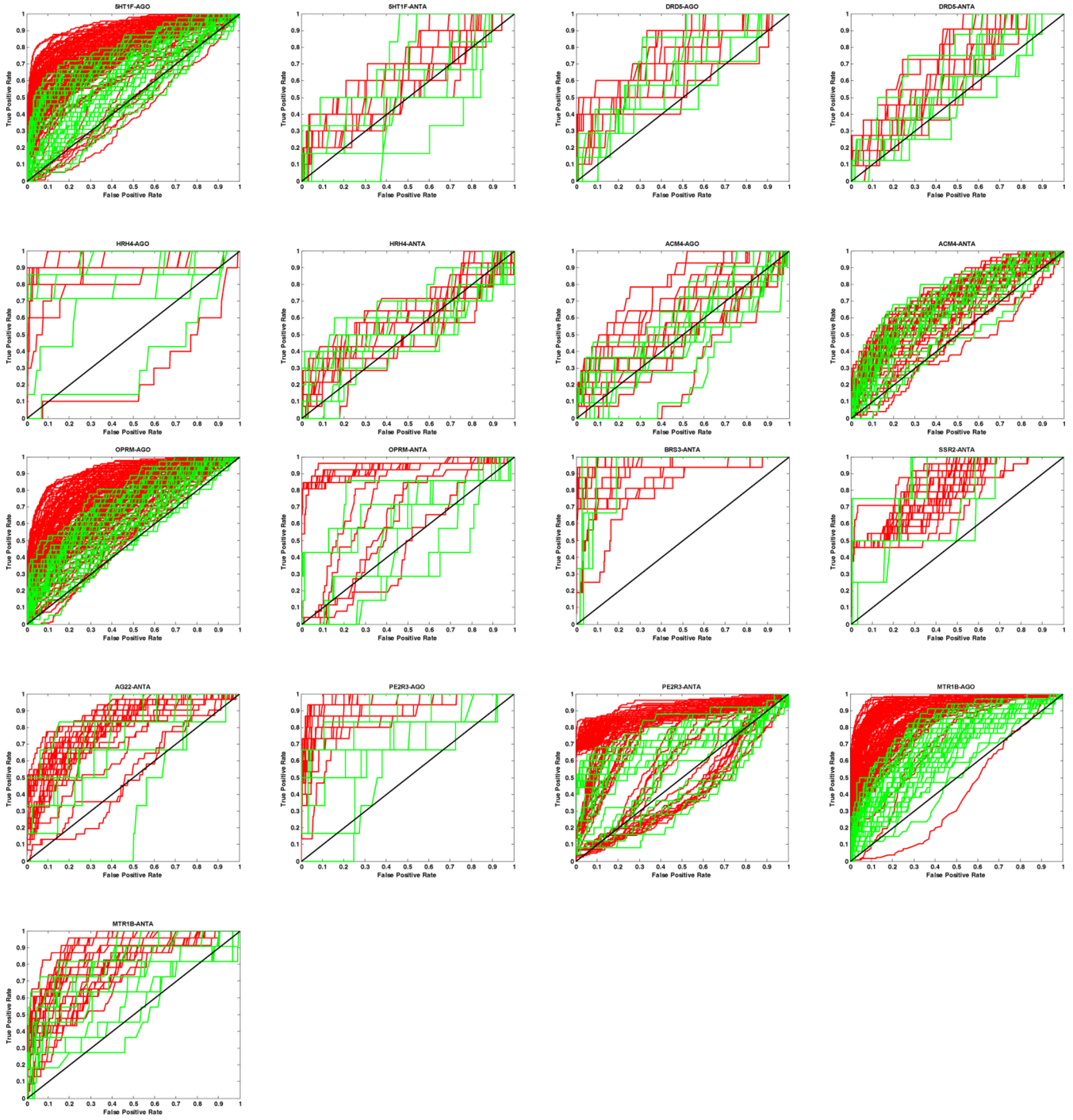


Figure 5. The ROC curves from similarity search by FCFP_6 fingerprint for all 17 data sets. For each data set, the curves are colored in red for GLL/GDD and green for ULS/UDS, respectively. The multiple curves in the same color represent different iterations in LOO CV for the specific benchmarking set, while the diagonal line in black shows the random distribution.

to "Murcko Assemblies," and other parameters were set as default values. After excluding analogues to constitute ULS, we carried out the same analysis to check the effect of our analogues excluding. Table 3 shows the comparison between GLL and ULS in terms of number of compounds, number of unique scaffolds, and the ratio of compounds to scaffolds. From this table, we can observe that after excluding analogues by using our strategy, the ratios of compounds/scaffolds decrease for all the targets and the average ratio decreases by 27.3%, from 1.62 to 1.18. The ratio of 1.18 means that ULS contains only 1.18 compounds per scaffold class, thus representing higher chemical diversity than GLL. At the same time, the number of ligands per receptor drops noticeably, e.g., from 135 to 35 for MTR1B-AGO, which can help reduce the computing cost of screening effort. From these two aspects, we concluded that our analogues excluding is effective in improving chemical diversity of the ligand set.

Effect of Decoys/Ligands Ratio on Quality of ULS/UDS. To the best of our knowledge, this issue has not been addressed before. In fact, the ratio of decoys to actives had been

dx.doi.org/10.1021/ci500062f | J. Chem. Inf. Model. 2014, 54, 1433−1450

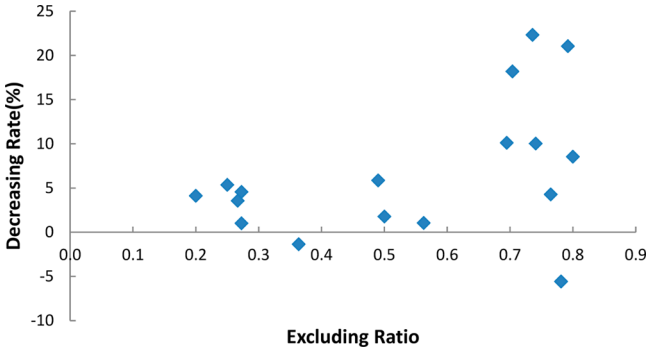


Figure 6. The relationship between the excluding ratio (removing ratios of analogues) and the decreasing rate of mean(ROC AUCs) from GLL/GDD to eGLL/eGDD.

set arbitrarily as different research groups defined different ratios in their studies, i.e., 36 in DUD[36] and VDS,[39] 39 in GDD,[40] 50 in DUD-E,[29] and 30 in DEKOIS,[41] respectively. In this study, we keep the same ratio as in GDD in order to compare our methodology with GDD in a fair way. In this section, we intend to examine the effect of various ratios on mean(ROC AUCs) of our data sets. To address this question, we select five representative GPCRs targets as samples whose current mean(ROC AUCs) values are distributed at different levels; i.e. the values for HRH4 agonists and PE2R3 agonists are above 0.70; the value for AG22 antagonists is in the range of [0.60, 0.70]; the value for ACM4 agonists is close to 0.50, and the one for 5HT1F antagonists is below 0.50. Apart from studying the ratios mentioned in other papers, we also increase the ratio to 100 so as to see the effect between 30 and 100. In this way, we have five points of ratio for each data set, i.e. 30, 36, 39, 50, and 100. The results of various ratios are shown in Figure 7 (cf. data in Tables S4 and S5). In general, there is no significant change from 30 to 100 for both types of screening

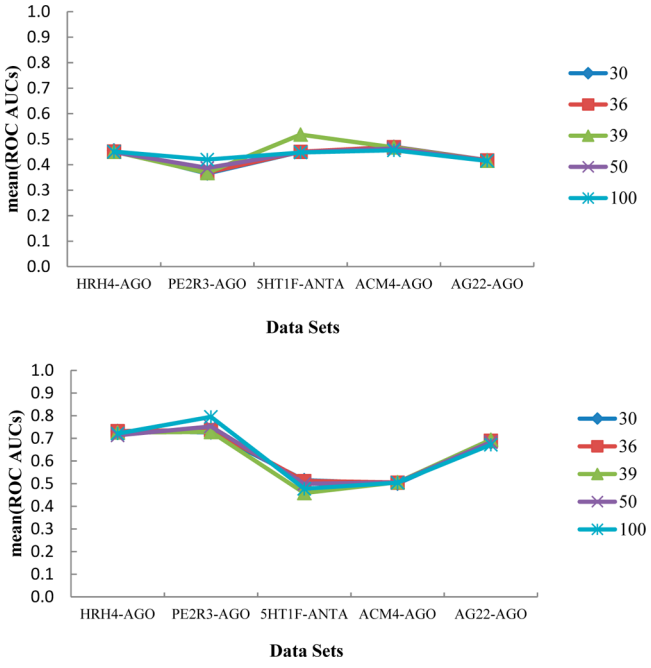


Figure 7. The effect of different decoy/ligand ratios on mean(ROC AUCs) from similarity search based on physicochemical properties ("*simp*," upper panel) and MACCS keys ("*sims*," lower panel).

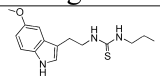
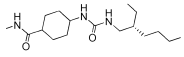
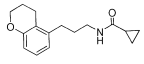
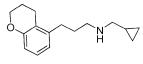
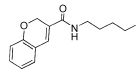
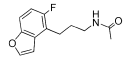
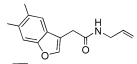
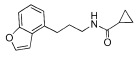
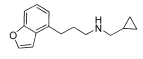
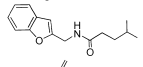
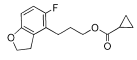
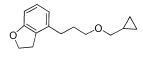
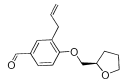
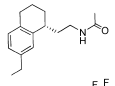
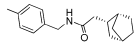
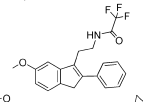
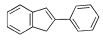
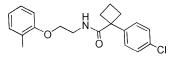
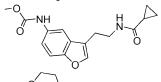
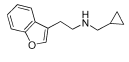
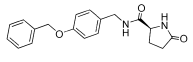
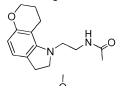
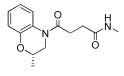
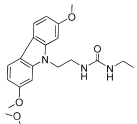
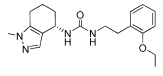
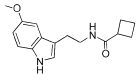
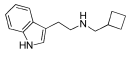
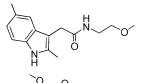
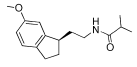
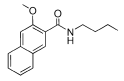
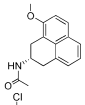
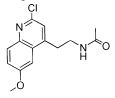
methods across the five data sets. All mean(ROC AUCs)s basically stay at the same level as 39, for both "*simp*"-based and "*sims*"-based VS. There is only a small spike for 39 at 5HT1F antagonists in comparison to other numbers with "*simp*" but are still close to 0.50. According to these results, we conclude that the effects of various decoys/ligands ratios (from 30 to 100) are nearly the same. In our current workflow, the number of 39 then appears to be a reasonable ratio for constructing the decoy sets and also is good for the purpose of comparison to GDD.

The Structural Features of Ligands and Decoys in ULS/UDS. As discussed before, values of "*simp*" and "*simsdiff*" are associated with the quality of the benchmarking set measured by mean(ROC AUCs). To obtain a detailed view of the individual benchmarking set, we chose the data set of MTR1B-AGO as an example to explore the structural features of ligands and decoys. The chemical structure of each ligand in ULS, its major scaffold, as well as its closest decoy in UDS are listed in Table 4, together with their "*simp*" and "*simsdiff*" values. Overall, we have the following observations: (1) The chemical structures of ligands are mostly different from each other, represented by unique scaffolds (Murcko frameworks). (2) The physicochemical properties of those decoys listed match well with those of the ligands, as shown by fairly high "*simp*" values (0.951−0.993). (3) In terms of chemical topology, the decoys resemble the ligands to a certain degree, with MACCS "*sims*" (Tc) at the range of [0.500, 0.745]. It should be noted that all MACCS "*sims*" are smaller than 0.75 (empirical threshold for active/inactive), which ensures the likelihood of decoys to be true inactives. (4) The "*simsdiff*" value can be applied here as a quantitative measure for how difficult it is to differentiate ligands from decoys. In this case, the "*simsdiff*" values of those decoys are extremely close to 0 with the highest value of 0.066, indicating that it is rather difficult to enrich the ligands by simple approaches such as similarity search. In summary, the structural features of the ligands and decoys in ULS/UDS meet the criteria for building benchmarking sets of high quality for LBVS.

CONCLUSIONS

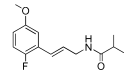
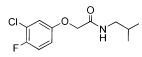
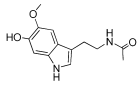
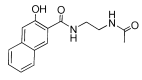
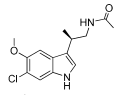
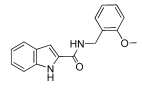
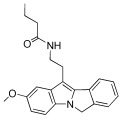
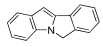
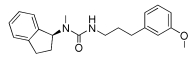
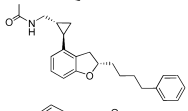
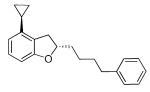
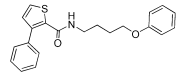
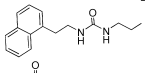
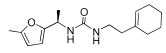
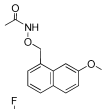
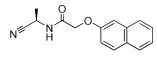
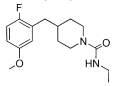
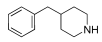
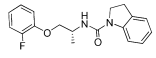
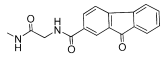
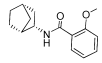
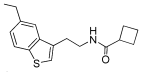
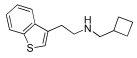
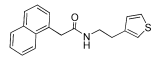
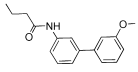
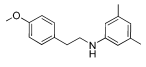
In the current study, we attempt to design an effective method to create benchmarking data sets for LBVS. As a means of validation, we applied our methods to a multitude of GPCRs targets. This kind of benchmarking study has become common in recent years for the purpose of virtual screening, though the main focus had been placed on the SBVS. Due to the lack of crystal structures, there is great need for unbiased benchmarking sets to evaluate different LBVS methods for GPCRs drug discovery. To be more specific, our methods can (1) ensure chemical diversity of ligands, (2) maintain the physicochemical similarity between ligands and decoys, and (3) make the decoys dissimilar in chemical topology to ligands. In addition, with the LOO CV based on MACCS or FCFP_6 fingerprint on 17 GPCRs' data sets, our ULS/UDS sets generated by this method reduced the "artificial enrichment" and "analogue bias" in GLL/GDD sets with great success. As our workflow includes analogues excluding, physicochemical properties-based filtering, and topology-based filtering, we move further to prove that our topology-based filtering strategies (mainly "*simsdiff*") account more for the effect of lowering the enrichment bias for LBVS than two other strategies, i.e., analogues excluding and "*simp*"-based filtering. Measured by the mean(ROC AUCs) from "*simp*"-based VS, we recovered the relationship in property distribution between our ULS and UDS sets. Its quality of

Table 4. Chemical Structures of Each Ligand and Its Scaffold (Murcko Framework) As Well As Its Corresponding Closest Decoy in ULS/UDS Benchmarking Set for MTR1B-AGO[a]

index	ligand	scaffold	closest decoy	*simp*	*simsdiff*	*sims*
1				0.952	0.057	0.500
2				0.975	0.042	0.745
3				0.964	0.040	0.696
4				0.982	0.040	0.738
5				0.962	0.037	0.700
6				0.979	0.040	0.743
7				0.964	0.027	0.661
8				0.961	0.029	0.649
9				0.953	0.035	0.730
10				0.986	0.031	0.739
11				0.953	0.056	0.727
12				0.993	0.040	0.694
13				0.992	0.040	0.744
14				0.960	0.033	0.720
15				0.955	0.038	0.712
16				0.960	0.046	0.745
17				0.963	0.033	0.705
18				0.968	0.056	0.738
19		**		0.967	0.014	0.722
20				0.966	0.042	0.708
21				0.954	0.058	0.712
22				0.951	0.039	0.698

dx.doi.org/10.1021/ci500062f | *J. Chem. Inf. Model.* 2014, 54, 1433−1450

Table 4. continued

index	ligand	scaffold	closest decoy	*simp*	*simsdiff*	*sims*
23				0.983	0.031	0.739
24				0.953	0.066	0.714
25				0.970	0.036	0.736
26				0.975	0.041	0.705
27				0.979	0.046	0.673
28				0.965	0.040	0.633
29				0.979	0.024	0.660
30				0.953	0.042	0.723
31				0.984	0.030	0.667
32				0.958	0.047	0.673
33				0.981	0.042	0.711
34				0.963	0.040	0.725
35				0.951	0.030	0.650
Min				0.951	0.014	0.500
Max				0.993	0.066	0.745

[**]No Murcko framework is generated due to nonring system in the structure. [a]Three similarity values, i.e. "*simp*," "*simsdiff*," and "*sims*," between each ligand and its closest decoy are also listed.

match is a popular metric to measure the performance of benchmarking sets, while a mismatch leads to the artificial enrichment in SBVS. Finally, we found out that the ratio for decoys and ligands in a range of 30 to 100 does not affect the quality of the benchmarking set, in which we employed the number of 39 for building our decoy sets (UDS).

Our methods mainly focus on generating decoy sets for application in LBVS. In fact, according to the outcome of our "*simp*"-based VS, it is challenging to differentiate ligands and decoys in ULS/UDS sets using similarity search based on six physicochemical properties, which is a basic criterion of benchmarking for molecular docking.[29,36,39,41,42] In the future, the benchmarking sets generated by our method can be extended to evaluate the methods of SBVS or even make a comparison between SBVS and LBVS in an unbiased manner. Our most immediate goal would be to apply this method to create benchmarking sets for each subtype in the chemokine receptor family for which the LBVS methods are still the most suitable tool for the discovery of subtype-selective chemokine receptor antagonists.

■ ASSOCIATED CONTENT

Supporting Information

The excluding ratios of analogues and decreasing rates of mean(ROC AUCs) for all the data sets, the "*simsdiff*" values for HRH4 agonists/decoys and the PE2R3 agonists/decoys, the data for Figure 6, as well as other supplementary data indicated in the text are available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: liangren@bjmu.edu.cn.

*E-mail: x.simon.wang@gmail.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by District of Columbia Developmental Center for AIDS Research (P30AI087714), National Institutes of Health Administrative Supplements for U.S.-China Biomedical Collaborative Research (5P30A10877714-02), and the National Institute on Minority Health and Health Disparities of the National Institutes of Health under Award Number G12MD007597. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We are also grateful to the China Scholarship Council (CSC) (201206010076), National Natural Science Foundation of China (NSFC 81373272, 81172917).

■ ABBREVIATIONS USED

GPCRs, G protein-coupled receptors; SBDD, structure-based drug design; LBDD, ligand-based drug design; VS, virtual screening; LBVS, ligand-based virtual screening; SBVS, structure-based virtual screening; *simp*, similarity in physico-chemical properties; *sims*, similarity in structure; *simsdiff*, *sims* difference; LOO, leave-one-out; CV, cross-validation; DUD, directory of useful decoys; MUV, maximum unbiased validation; VDS, virtual decoy sets; GLL, GPCR ligand library; GDD, GPCR decoy database; ULS, unbiased ligand set; UDS, unbiased decoy set; eGLL, excluded GLL; eGDD, excluded GDD; MW, molecular weight; HBAs, number of hydrogen bond acceptors; HBDs, number of hydrogen bond donors; RBs, number of rotatable bonds; FC, formal charge; PDs, potential decoys; FDs, final decoys; Tc, Tanimoto coefficient; DOE score, deviation from optimal embedding score; FCFP_6, function class fingerprints of maximum diameter 6

■ REFERENCES

(1) Lappano, R.; Maggiolini, M. G protein-coupled receptors: novel targets for drug discovery in cancer. *Nat. Rev. Drug Discovery* **2011**, *10* (1), 47−60.

(2) Granier, S.; Kobilka, B. A new era of GPCR structural and chemical biology. *Nat. Chem. Biol.* **2012**, *8* (8), 670−3.

(3) Stevens, R. C.; Cherezov, V.; Katritch, V.; Abagyan, R.; Kuhn, P.; Rosen, H.; Wuthrich, K. The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nat. Rev. Drug Discovery* **2012**, *12*, 25−34.

(4) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5* (12), 993−6.

(5) Shorr, R. G.; Lefkowitz, R. J.; Caron, M. G. Purification of the beta-adrenergic receptor. Identification of the hormone binding subunit. *J. Biol. Chem.* **1981**, *256* (11), 5820−6.

(6) Latek, D.; Modzelewska, A.; Trzaskowski, B.; Palczewski, K.; Filipek, S. G protein-coupled receptors - recent advances. *Acta Biochim. Pol.* **2012**, *59*, 515−529.

(7) Mason, J. S.; Bortolato, A.; Congreve, M.; Marshall, F. H. New insights from structural biology into the druggability of G protein-coupled receptors. *Trends Pharmacol. Sci.* **2012**, *33* (5), 249−60.

(8) Zhao, Q.; Wu, B. L. Ice breaking in GPCR structural biology. *Acta Pharmacol. Sin.* **2012**, *33* (3), 324−34.

(9) Congreve, M.; Langmead, C. J.; Mason, J. S.; Marshall, F. H. Progress in structure based drug design for G protein-coupled receptors. *J. Med. Chem.* **2011**, *54* (13), 4283−311.

(10) Jang, J. W.; Kim, M. S.; Cho, Y. S.; Cho, A. E.; Pae, A. N. Identification of structural determinants of ligand selectivity in 5-HT(2) receptor subtypes on the basis of protein-ligand interactions. *J. Mol. Graphics Modell.* **2012**, *38*, 342−53.

(11) Alkhalfioui, F.; Magnin, T.; Wagner, R. From purified GPCRs to drug discovery: the promise of protein-based methodologies. *Curr. Opin. Pharmacol.* **2009**, *9* (5), 629−35.

(12) Willett, P. Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.* **2011**, *672*, 133−58.

(13) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23−24), 1046−53.

(14) Berglund, A. E.; Head, R. D. PZIM: a method for similarity searching using atom environments and 2D alignment. *J. Chem. Inf. Model.* **2010**, *50* (10), 1790−5.

(15) Horvath, D. Pharmacophore-based virtual screening. *Methods Mol. Biol.* **2011**, *672*, 261−98.

(16) Gao, Q.; Yang, L.; Zhu, Y. Pharmacophore based drug design approach as a practical process in drug discovery. *Curr. Comput.-Aided Drug Des.* **2010**, *6* (1), 37−49.

(17) Caporuscio, F.; Tafi, A. Pharmacophore modelling: a forty year old approach and its modern synergies. *Curr. Med. Chem.* **2011**, *18* (17), 2543−53.

(18) Yang, S. Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today* **2010**, *15* (11−12), 444−50.

(19) Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des* **2007**, *13* (34), 3494−504.

(20) Tropsha, A.; Wang, S. X. QSAR modeling of GPCR ligands: methodologies and examples of applications. *Ernst Schering Found. Symp. Proc.* **2006**, No. 2, 49−73.

(21) Wang, X. S.; Tang, H.; Golbraikh, A.; Tropsha, A. Combinatorial QSAR modeling of specificity and subtype selectivity of ligands binding to serotonin receptors 5HT1E and 5HT1F. *J. Chem. Inf. Model.* **2008**, *48* (5), 997−1013.

(22) Costanzi, S.; Tikhonova, I. G.; Harden, T. K.; Jacobson, K. A. Ligand and structure-based methodologies for the prediction of the activity of G protein-coupled receptor ligands. *J. Comput.-Aided Mol. Des.* **2009**, *23* (11), 747−54.

(23) Sage, C.; Wang, R.; Jones, G. G-protein coupled receptors virtual screening using genetic algorithm focused chemical space. *J. Chem. Inf. Model.* **2011**, *51* (8), 1754−61.

(24) Vogt, I.; Ahmed, H. E.; Auer, J.; Bajorath, J. Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping. *Mol. Diversity* **2008**, *12* (1), 25−40.

(25) Taylor, C. M.; Rockweiler, N. B.; Liu, C.; Rikimaru, L.; Tunemalm, A. K.; Kisselev, O. G.; Marshall, G. R. Using ligand-based virtual screening to allosterically stabilize the activated state of a GPCR. *Chem. Biol. Drug Des.* **2010**, *75* (3), 325−32.

(26) Sukumar, N.; Das, S. Current trends in virtual high throughput screening using ligand-based and structure-based methods. *Comb. Chem. High Throughput Screening* **2011**, *14* (10), 872−88.

(27) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11* (9), 1189−202.

(28) Sanders, M. P.; Barbosa, A. J.; Zarzycka, B.; Nicolaes, G. A.; Klomp, J. P.; de Vlieg, J.; Del Rio, A. Comparative analysis of pharmacophore screening tools. *J. Chem. Inf. Model.* **2012**, *52* (6), 1607−20.

(29) Mysinger, M. 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* (14), 6582−94.

(30) Butkiewicz, M.; Lowe, E. W.; Mueller, R.; Mendenhall, J. L.; Teixeira, P. L.; Weaver, C. D.; Meiler, J. Benchmarking Ligand-Based Virtual High-Throughput Screening with the PubChem Database. *Molecules* **2013**, *18* (1), 735−56.

(31) von Korff, M.; Freyss, J.; Sander, T. Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* **2009**, *49* (2), 209−31.

(32) 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* (4), 1504−19.
(33) Hu, G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *J. Chem. Inf. Model.* **2012**, *52* (5), 1103−13.
(34) Zhou, H.; Skolnick, J. FINDSITE(comb): A Threading/Structure-Based, Proteomic-Scale Virtual Ligand Screening Approach. *J. Chem. Inf. Model.* **2013**, *53* (1), 230−40.
(35) Novikov, F. N.; Stroylov, V. S.; Zeifman, A. A.; Stroganov, O. V.; Kulkov, V.; Chilov, G. G. Lead Finder docking and virtual screening evaluation with Astex and DUD test sets. *J. Comput.-Aided Mol. Des.* **2012**, *26* (6), 725−35.
(36) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789−801.
(37) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37* (Web Serverissue), W623−33.
(38) 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* (2), 169−84.
(39) Wallach, I.; Lilien, R. Virtual decoy sets for molecular docking benchmarks. *J. Chem. Inf. Model.* **2011**, *51* (2), 196−202.
(40) Gatica, E. A.; Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* **2012**, *52* (1), 1−6.
(41) 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* (10), 2650−65.
(42) Cereto-Massague, A.; Guasch, L.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics* **2012**, *28* (12), 1661−2.
(43) Ripphausen, P.; Wassermann, A. M.; Bajorath, J. REPROVIS-DB: a benchmark system for ligand-based virtual screening derived from reproducible prospective applications. *J. Chem. Inf. Model.* **2011**, *51* (10), 2467−73.
(44) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 793−806.
(45) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 193−9.
(46) Rohrer, S. G.; Baumann, K. Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J. Chem. Inf. Model.* **2008**, *48* (4), 704−18.
(47) 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* (3−4), 169−78.
(48) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 141−6.
(49) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *J. Cheminform.* **2009**, *1*, 14.
(50) *MACCS Structural Keys*; MDL Information Systems Inc.: San Ramon, CA, 2005.
(51) Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. GLIDA: GPCR–ligand database for chemical genomics drug discovery–database and tools update. *Nucleic Acids Res.* **2008**, *36* (Database issue), D907−12.
(52) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(53) Tanimoto, T. *IBM Internal Report*; IBM Corp: Armonk, NY, 1957.
(54) Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27* (8), 861−874.
(55) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.
(56) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462* (7270), 175−81.
(57) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−93.