Published in final edited form as:

Methods. 2015 January 1; 0: 146–157. doi:10.1016/j.ymeth.2014.11.015.

Benchmarking Methods and Data Sets for Ligand Enrichment Assessment in Virtual Screening

Jie Xia^{a,b}, Ermias Lemma Tilahun^b, Terry-Elinor Reid^b, Liangren Zhang^{a,*}, and Xiang Simon Wang^{b,*}

^aState Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing 100191, P.R. China

^bMolecular Modeling and Drug Discovery Core for District of Columbia Developmental Center for AIDS Research (DC D-CFAR); Laboratory of Cheminformatics and Drug Design, Department of Pharmaceutical Sciences, College of Pharmacy, Howard University, Washington, DC 20059, U.S.A.

Abstract

Retrospective small-scale virtual screening (VS) based on benchmarking data sets has been widely used to estimate ligand enrichments of VS approaches in the prospective (i.e. real-world) efforts. However, the intrinsic differences of benchmarking sets to the real screening chemical libraries can cause biased assessment. Herein, we summarize the history of benchmarking methods as well as data sets and highlight three main types of biases found in benchmarking sets, i.e. “analogue bias”, “artificial enrichment” and “false negative”. In addition, we introduced our recent algorithm to build maximum-unbiased benchmarking sets applicable to both ligand-based and structure-based VS approaches, and its implementations to three important human histone deacetylase (HDAC) isoforms, i.e. HDAC1, HDAC6 and HDAC8. The Leave-One-Out Cross-Validation (LOO CV) demonstrates that the benchmarking sets built by our algorithm are maximum-unbiased in terms of property matching, ROC curves and AUCs.

Keywords

Benchmarking methodology; Decoy sets; Structure-based virtual screening; Ligand-based virtual screening; Artificial enrichment; Analogue bias

© 2014 Elsevier Inc. All rights reserved.

*Correspondence should be addressed X.S.W. (x.simon.wang@gmail.com) or L-R.Z. (liangren@bjmu.edu.cn): Xiang Simon Wang, Ph.D., Howard University College of Pharmacy, 2300 4th St. NW, Washington, DC 20059, U.S.A., Liangren Zhang, Ph.D., State Key Laboratory of Natural and Biomimetic Drugs, Peking University School of Pharmaceutical Sciences, 38 Xueyuan Rd, Beijing 100191, China.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Since the first seminal publication by Kuntz et al [1], virtual screening (VS) has become an indispensable technique in the early-stage drug discovery to identify bioactive compounds against a specific target in a cost-effective and time-efficient manner [2]. A large collection of review-type literatures have discussed various VS approaches and provided perspectives of this technique [3–16, 115]. In general, VS aims to filter out thousands of nonbinders *in silico* and ultimately to reduce the cost related to bioassay and chemical synthesis [17, 18]. Depending on the availability of three-dimensional structures of biological targets, VS approaches are typically classified into Structure-based Virtual Screening (SBVS) and Ligand-based Virtual Screening (LBVS) [19]. The SBVS approaches, often referred to be molecular docking, employ the three-dimensional target structure to identify molecules that potentially bind to the target with appreciable affinity and specificity [10, 16, 20]. The latter is normally similarity-based, which identifies compounds of novel chemotypes but with similar activities by mining the information of known ligands [5, 11, 12, 17, 21–23].

To date, a wide variety of screening tools for both SBVS and LBVS have been developed [24–41]. Among them, DOCK [24], AutoDock [25], FlexX [26], Surflex [27], LigandFit [28], GOLD [29], Glide [30], ICM [31], and eHiTS [32] are popular tools for SBVS and updated regularly. For LBVS, QSAR modeling workflow [22] has been made publicly accessible to scientific communities by being incorporated into Chembench [33]. Catalyst [34], PHASE [35], and LigandScout [36] are classic algorithms for pharmacophore modeling. Needless to say, similarity search based on 2D structural fingerprints also plays a pivotal role in LBVS [23]. To date, new approaches are still emerging at a rapid pace. The recent successes of integrating Machine Learning (ML) as well as other cheminformatic techniques to improve accuracy of scoring functions [15] are encouraging, e.g. SFCScore(RF) [37], libSVM plus Medusa [38], and the development of novel descriptors [39] or fingerprints [40, 41].

With such a large number of VS approaches, it is of utmost importance for the users to learn which one is the optimal method for the specific target(s) under study. For this purpose, the objective assessments for all viable approaches become indispensable. Usually, the performance of each approach is measured by ligand enrichment from retrospective small-scale VS with a benchmarking set, as evidenced by numerous literatures [5, 14, 42–56]. Ligand enrichment is a metric to assess the capacity to place true ligands at the top-rank of the screen list among a pool of a large number of decoys, which are presumed inactives that are not likely to bind to the target [57, 58]. The combination of true ligands and their associated decoys is known as the benchmarking set [59]. This type of assessment is expected to uncover the merits and deficits of each approach for a specific target/task, thus being able to provide advices on method selection for prospective VS campaigns. Particularly, when new algorithms are developed, an objective assessment is normally needed to compare with the prior ones, thus to decide the necessity of the update. Also, in SBVS the assessment can assist in the optimization of receptor structures as well as the selection of the best comparative model(s) for screening purpose [60]. In fact, these types of studies have become the normal practice in both SBVS and LBVS in recent years. Nevertheless, ligand enrichment assessment based on a highly-biased or unsuitable

benchmarking set will not reflect the realistic enrichment power of various approaches for prospective VS campaigns. For example, as mentioned by Cleves and Jain, “2D-biased” data sets could cause questionable assessment when comparing SBVS and LBVS approaches [61]. In this way, the quality of the benchmarking sets becomes rather crucial for a fair and comprehensive evaluation.

In our opinion, benchmarking sets can be classified into two major types according to their initial designing purposes, i.e. the SBVS-specific and the LBVS-specific. Datasets such as Directory of Useful Decoys (DUD) [57] and its recent DUD-Enhanced (DUD-E) [58], virtual decoy sets (VDS) [62], G protein-coupled receptors (GPCRs) ligand library (GLL) and GPCRs Decoy Database (GDD) [63], Demanding Evaluation Kits for Objective *in Silico* Screening (DEKOIS) [64] and DEKOIS 2.0 [65], Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB) belong to SBVS-specific benchmarking sets. By contrast, only 3 datasets, i.e. DUD LIB VS 1.0 [66], database of *reproducible* virtual screens (REPROVIS-DB) [67] and Maximum Unbiased Validation (MUV) [68] are specifically designed for the purpose of LBVS. A detailed introduction of each data set is given in Table 1. To date, DUD and DUD-E have been intensively employed as gold standard data sets among the community [38, 69–74], while much fewer citations of DUD LIB VS 1.0 [56, 75] and MUV [76, 77] have been reported. In order to broaden the application domain of currently available LBVS-specific benchmarking sets, we recently proposed an unbiased method to build LBVS-specific benchmarking sets [78]. Herein, we review the development of both SBVS-specific and LBVS-specific benchmarking methods/sets and discuss their merits and deficits. In the end, we give a brief introduction to our in-house method and its application to build benchmarking sets for three human Histone Deacetylase (HDAC) isoforms which are under intensive studies.

2. Currently available benchmarking sets

2.1 SBVS-specific benchmarking sets and methods

2.1.1 Early-stage of benchmarking sets—The usage of benchmarking sets to evaluate docking approaches dates back to early 2000. The first pioneering benchmarking sets were created by Rognan et al. [79], and covered two popular targets: thymidine kinase (TK) and estrogen receptor α subtype (ER α). The data set for each target was composed of 10 antagonists (ligands) and 990 decoys. The method to build the benchmarking sets was relatively simple: First, 10 known ligands were collected for each target; then compounds in Advanced Chemical Directory (ACD) v.2000-1, (Molecular Design Limited, San Leandro) were filtered to eliminate chemical reagents, inorganic compounds, and molecules with unsuitable molecular weights (MWs); at last, 990 compounds were randomly selected as decoys from the remaining compounds. Through these benchmarking sets, the authors addressed issues such as the performances of different docking programs and the accuracy of consensus scoring rationally. Because of the success of this study, parameter optimization based on benchmarking sets had been regarded as a necessity prior to the screening of large chemical libraries. Later on, the benchmarking set for TK was applied to the comparative evaluation of 8 docking tools [43] while similar method was adopted to build benchmarking

sets for the assessment of GLIDE [80]. These decoy sets are available at <http://bioinfo-pharma.ustrasbg.fr/labwebsite/download.html> (accessed in Jun. 2014).

Besides ACD, MDL Drug Data Report (MDDR) had also been employed as the main source of decoys during the years of 2002–2005. In Shoichet's group, MDDR was first processed by removing those compounds containing unwanted functional groups such as phosphine. Next, 95,000 remaining “drug-like” compounds were combined with a certain number of ligands for each target [81, 82]. Diller et al [83] selected 32,000 compounds randomly from MDDR as decoys and put them together with over 1,000 known kinase inhibitors across 6 targets. In their study, they kept properties of decoys such as MW, number of rotatable bonds (RBs), H-bond acceptor (HBA) and H-bond donor (HBD) in line with those ligands. Though the criteria for property matching were not strict, this practice should be considered to be an improvement in benchmarking methods. These two types of benchmarking sets were not widely used, however, due to the commercial feature of MDDR.

In 2005, Jain's group at UCSF also released their own decoy sets of “ZINC negative set” to supplement then limited, not publicly-accessible benchmarking sets [84]. To build the sets, they retrieved 20 ligands for each target from PDBbind (<http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp>) [85]. A total of 1,000 decoys (also called “negative ligands”) were randomly collected from ZINC drug-like subset. This particular benchmarking set was used to optimize the performance of Surflex-Dock and is available at <http://www.jainlab.org/downloads.html>. It should be noted that in the above benchmarking sets, the ratio of decoys per ligand was set arbitrarily and the physicochemical properties of ligands and decoys were not strictly matched. Besides, the issue that “drug-like” decoys can be true binders had not carefully been addressed. Those datasets were thus considered to be bias-uncorrected benchmarking sets [57] for comparison purpose.

2.1.2 DUD, DUD clusters and charge-matched DUD—Analysis of early-stage benchmarking sets indicated that molecular size may cause overoptimistic ligand enrichment from SBVS [87–89]. The similar situation applies for other low-dimensional physicochemical properties as well [88]. It is known that poor property matching between ligands and decoys causes the “artificial enrichment”. To reduce such a benchmarking bias, the basic principle is that the decoys should resemble the ligands in the aspect of physicochemical properties. In 2006, Huang et al. compiled DUD [57] which ensured similar physicochemical properties but dissimilar 2-D topology between ligands and decoys. DUD contains 95,316 decoys and 2,950 ligands against 40 protein targets. The targets were selected based on the availability of annotated ligands, crystal structures and previous docking studies, and covered 6 classes: nuclear hormone receptors, kinases, serine proteases, metalloenzymes, folate enzymes, and other enzymes. The number of ligands for each target varies from 12 to 416, while the decoys are from the ZINC “drug-like” subset (<https://zinc.docking.org/>, accessed in Jun. 2014) [86]. The method for decoy generation was simple but effective. Firstly, Tanimoto coefficient (T_c) based on type 2 substructure keys of CACTVS was calculated to measure pairwise topological similarity between ligands and all compounds in ZINC “drug-like” subset before similar compounds in ZINC with $T_c > 0.9$ were filtered. Next, 32 physicochemical properties of ligands and the remaining compounds in ZINC were calculated and only 5 properties, i.e. MW, HBAs, HBDs, RBs and $LogP$ were

highlighted because of their direct association with drug-likeness defined by the Lipinski's rule of five. Subsequently, for each ligand 36 most similar compounds in physical properties were chosen as the final decoys in DUD. The comparison between DUD and prior benchmarking sets such as MDDR, Rognan's decoys and Jain's decoys showed that (1) physical properties of DUD decoys span the same range and contain maxima at about the same location as ligands while the situation was different in other datasets; (2) DUD reduced artificial enrichment bias in MDDR, Rognan's and Jain's decoys. In addition, DUD is publicly accessible at <http://dud.docking.org/>. It is the first bias-corrected, publicly available and SBVS-specific benchmarking set that covers the widest range of protein targets. Since its release in 2006, DUD has become the standard benchmarks and been employed in various types of studies such as structural model validation [90] and software evaluation [74, 91].

In 2007, Good and Oprea identified another type of benchmarking bias named “analogue bias” in DUD and found that it causes overoptimistic ligand enrichment as well [92]. To reduce the bias, they applied lead-like filter (i.e. MW < 450 and Alog P < 4) and reduced-graph filter to refine the DUD ligand sets [92]. These filters ensure lead-like features and chemical diversity of the ligands, which help evaluate the scaffold hopping ability of VS approaches. The refined DUD ligand dataset, known as “DUD Clustering” is available at <http://dud.docking.org/clusters/>. It is recommended to combine “DUD Clusters” with original DUD decoy sets when conducting the ligand enrichment study.

In 2010, when Shoichet's group was developing a context-dependent ligand desolvation method [93], they found that the original DUD decoys are more neutral than ligands, which render those decoys not suitable for evaluating purpose. To address this issue, they built an updated version of ZINC [86] and collect DUD decoys by matching of net formal charge (FC) together with prior physicochemical properties. The resulting data set is known as “charge-matched DUD” and available at <http://dud.docking.org/>.

2.1.3 Supplementary benchmarking sets to DUD—Multiple research groups have been building decoy sets to address the apparent weaknesses in DUD [62–64, 92, 94]. Good and Oprea generated World of Molecular BioAcTivity (WOMBAT) data sets at the same time in order to expand the limited number of diverse ligands in “DUD clusters” [92]. Their method was similar to the one of building “DUD clusters”, with the same goal of reducing “analogue bias”. Notably, the source of the ligands was WOMBAT [95] which contained more active compounds. The data sets were released at <http://dud.docking.org/>.

To expand the limited chemical space of DUD decoys, Wallach and Lilien created VDS (<http://compbio.cs.toronto.edu/VDS/>, accessed in Jun. 2014) by ignoring the synthetic feasibility [62]. For comparison purpose, they started with DUD ligands and calculated 5 similar physicochemical properties for matching. In addition, they calculated mutual topological similarity (i.e. T_c) values based on MACCS fingerprints between ligands and virtual molecules. Only when the molecules were similar in 5 physicochemical properties but dissimilar in structure ($T_c < 0.9$) to the reference ligand, they were chosen as decoys. The final ratio of decoys per ligand was kept the same as with DUD, i.e. 36. It was proved that VDS was comparable to the DUD in general but less biased in terms of

physicochemical similarity. It should be noted that the decoys in VDS were artificial molecules. The synthetic unfeasibility of these decoys makes VDS different from those of real collections, e.g. commercial chemical libraries, which may cause “inductive bias” [61]. Nevertheless, VDS had been previously used to normalize the docking ranks [96].

In the middle of 2012, due to the lack of decoy makers in DUD, Garcia-Vallve et al. developed DecoyFinder [94], a tool to generate decoys for a given group of ligands. The protocol utilized 5 physicochemical properties (the same as DUD) for property matching and MACCS fingerprint-based T_c calculation for topological pairwise similarity. The default cutoff value for T_c was 0.75 though it can be customized by the users. This tool broadens the application domain of DUD method because it can build decoys from any source of ligands, at any ratio of decoys per ligand and for any targets in which the user is interested. This tool is publicly accessible at <http://urvnutrigenomica-ctns.github.io/DecoyFinder/> (accessed in Jun. 2014).

In the same year, Cavasotto et al. compiled benchmarking sets for GPCRs, i.e. GLL/GDD in order to increase target diversity [63]. The 25,145 ligands in GLL were retrieved from GPCR-Ligand Database (GLIDA) [97] (<http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/>), covering 147 human Class A GPCR targets. The source of GDD decoys was also the ZINC database (around 20 million, accessed in Jan. 2011 at <http://zinc.docking.org/>). For property matching between ligands and decoys, FC was also introduced in addition to 5 common physicochemical properties because it was observed that the majority of GPCR ligands have non-zero charges [2]. Similarly, molecules in ZINC whose $T_c < 0.75$ based on MACCS fingerprint with reference to any known ligands were considered to be potential actives thus were filtered. The ratio of decoys per ligand was 39 in this case. Apart from target diversity, another merit of GLL/GDD was that the ligands were partitioned into agonists and antagonists, which was proved to be important for the quality of ligand enrichment in SBVS [98, 99].

Another useful benchmarking set for docking study was DEKOIS (2011) [64]. It highlights the issue of “false negatives” (decoys that were validated to be binders by experiments) in particular, and addressed the problem of “artificial enrichment”. They introduced a consensus score that fused normalized “*Physicochemical similarity score* (PSS)” and “*Avoidance of LADS score*” to quantify physicochemical similarity and possibility of being latent actives (“false negatives”) in the decoy set (LADS), respectively. Meanwhile, they developed two metrics, i.e. *deviation from optimal embedding score* (DOE score) to measure the quality of property matching and *doppelganger score* to assess structural similarity between actives and decoys. These features made them unique compared to all the prior methods such as DUD, DUD-E, GLL/GDD and VDS. The formulas to calculate PSS are listed in eqs 1–3 [64]. The *Avoidance of LADS score* is the remainder after division of the difference between bit strings of the potential decoy and all actives by number of atoms of that decoy. These scores were normalized and fused into a consensus score.

$$f_j = \begin{cases} |x_{j,ref} - x_{j,min}|, & \text{if } |x_{j,ref} - x_{j,min}| \geq |x_{j,ref} - x_{j,max}| \\ |x_{j,ref} - x_{j,max}|, & \text{if } |x_{j,ref} - x_{j,min}| < |x_{j,ref} - x_{j,max}| \end{cases} \quad (1)$$

$$\delta_{i,j} = \begin{cases} 1 - \frac{|x_{ref,j} - x_{i,j}|}{f_j}, & \text{if } f_j \neq 0 \\ 1, & \text{if } f_j = 0 \end{cases} \quad (2)$$

$$PSS_i = \frac{1}{n} \sum_{j=1}^n \delta_{i,j} \quad (3)$$

In these eqs, $x_{j,ref}$ is the value of physicochemical property j of the reference ligand ref , while $x_{j,min}$ and $x_{j,max}$ are the minimum and maximum values of property j , respectively. f_j is the bigger interval of $|x_{j,ref} - x_{j,min}|$ and $|x_{j,ref} - x_{j,max}|$, while $\delta_{i,j}$ is a similarity score by property j between decoy i and reference ligand ref . PSS_i is the average similarity score of n properties between decoy i and reference ligand ref .

The following equations show how they normalized the data (eq 4) and calculated *DOE score* based on ROC curves (eqs 5–6).

$$x_{i,j,norm} = \frac{x_{i,j}}{(x_{j,95\%} - x_{j,5\%})} \quad (4)$$

$$ABC_a^{DOE} = \sum_{i=1}^n |0.5 [(TPR_i + TPR_{i-1}) \bullet (FPR_i - FPR_{i-1}) - (FPR_i + FPR_{i-1}) \bullet (FPR_i - FPR_{i-1})]| \quad (5)$$

$$DOE_{score} = \frac{1}{m} \sum_{a=1}^n ABC_a^{DOE} \quad (6)$$

In eq. 4, $x_{i,j,norm}$ is the normalized value of property j for compound i . $x_{j,95\%}$ and $x_{j,5\%}$ are the 95th and 5th percentile of property j , respectively. In eq. 5, TPR and FPR represent true positive rate and false positive rate, respectively. ABC is the absolute value of areas between the ROC curve for active a and the random distribution $f(x) = x$, while *DOE score* is the average value of ABCs for all n actives.

The workflow was implemented on Pipeline Pilot platform (version 6.1.5.0., Accelrys Software, Inc.). The ligands were collected from the BindingDB (<http://www.bindingdb.org/bind/index.jsp>) [100] while the decoys were from ZINC subset #10 named “Everything” (version 8, accessed in Nov. 2009). After filtering non-druglike compounds in ZINC, they classified ZINC compounds into 5 equally populated physicochemical bins. The combination of property bins constitutes different physicochemical cells. Next, they

assigned each compound into respective cell according to common physicochemical properties. Subsequently, a pool of 1,500 potential decoys was constructed by preselecting compounds from “parent cell” or “direct neighbor cells” for each active. 100 optimal decoys were then chosen from the pool of potential decoys according to their consensus scores. Finally, 30 diverse decoys were selected from optimal decoys based on the function class fingerprints of maximum diameter 6 (FCFP_6). By comparing DEKOIS and DUD, they proved that their datasets met or exceeded the standards of property matching in DUD and ensured better avoidance of “latent actives” in the decoys sets. The data set was publicly accessible at http://www.dekois.com/dekois_orig.html (accessed in Jun. 2014).

2.1.4 Recent unbiased benchmarking sets—Notably, the above supplementary benchmarking sets to DUD addressed only one or two specific problem(s) rather than all of them. To this end, individual groups have tried to optimize their original data sets so as to minimize all potential benchmarking biases [58, 65, 99]. In June 2012, Irwin and Shoichet released an enhanced version of DUD, DUD: Enhanced (DUD-E) for short [58], which addressed all the observed weaknesses in original DUD. The optimization included: 1) More targets (102), including 5 GPCRs and 2 ion channels, were added to increase target diversity; 2) More annotated ligands (100–600) were collected from ChEMBL09 (<https://www.ebi.ac.uk/chembl/>) [101] and clustering based on Bemis-Murcko atomic frameworks [102] was performed on those ligands to ensure structural diversity; 3) The net FC was also added as an important descriptor to improve property matching between ligands and decoys; 4) The topology (extended-connectivity fingerprints of maximum diameter 4, ECFP_4) related criteria were used to select the most 25% dissimilar decoys so as to further lower the chance of “false negatives”; 5) The ratio of decoys per ligand was adjusted to 50; 6) An online tool of decoy-maker for user-supplied ligands was also provided at <http://dude.docking.org/generate> to broaden its application domain. In spite of its recent release, applications of DUD-E have already been reported [69, 70, 103].

NRLiSt BDB was built and released in 2014, for benchmarking of Nuclear Receptors ligands [99]. Like other SBVS-specific benchmarking sets, this database was composed of targets with crystal structures available, as well as their ligands and decoys. But different from DUD and DUD-E, more than one holo-crystal structures per target were included in this data sets. In addition, ligands from ChEMBL [101] (<https://www.ebi.ac.uk/chembl/>) and PubChem [104] (<http://www.ncbi.nlm.nih.gov/pccompound>, both accessed in Jun. 2014) were manually curated and divided into agonists and antagonists based on their activities’ nature. The decoys were generated by DUD-E decoy maker. It was specifically aimed to evaluate the performance of SBVS approaches on NRs, to understand the NRs’ function and modulation, and to support the drug discovery efforts targeting NRs. NRLiSt BDB is currently available on-line at <http://nrlist.drugdesign.fr>, and it can be considered as extended benchmarking sets for DUD-E.

In 2013, Boeckler’s group also realized the weaknesses in DEKOIS, in particular its hidden analogue bias, thus updated DEKOIS to version 2.0 [65] by: (1) including 3 additional physicochemical properties for larger matching; (2) updating *LADS score* for avoidance of latent actives; and (3) clustering initial ligands for higher chemical diversity. These updates further reduced three main biases, i.e. artificial enrichment, false negatives and analogue

bias. The additional properties they applied were negative charges, positive charges and number of aromatic rings. The *LADS score* was calculated by eq. 7. In this equation, $N_{FCFP_6 fragments}$ is the total number of unique FCFP_6 fragments in the active set; n is the total number of shared FCFP_6 fragments between the preselected decoy structure and all the actives; i is the index of each shared FCFP_6 fragment; $N_{i(HeavyAtoms)}$ is the numbers of heavy atoms in the shared FCFP_6 fragment i while $f_{i(FCFP_6 fragments)}$ represents the frequency of shared FCFP_6 fragment i in the active set. Those initial ligands from BindingDB [100] were clustered into 40 bins by a maximum dissimilarity partition according to T_c based on FCFP_6. Then the most potent ligand from each bin constituted the final ligand set. The other procedures were the same as the prior one used in DEKOIS.

$$LADS_{score} = \frac{\sum_{i=1}^n \left(N_{i(HeavyAtoms)} \cdot f_{i(FCFP_6 fragments)} \right)}{N_{FCFP_6 fragments}} \quad (7)$$

In summary, DEKOIS 2.0 includes 80 targets that are popular in the area of medicinal chemistry, with 40 ligands and 1200 (i.e. 30×40) decoys for each target. The data sets are available at <http://www.dekois.com/> (last accessed in Jun. 2014).

2.2 LBVS-specific benchmarking sets

2.2.1 MUV sets—MUV is the first reported LBVS-specific benchmarking sets. It addressed the benchmarking bias related to LBVS such as “artificial enrichment”, “analogue bias” and “false negatives”. Both of its ligands and decoys were taken from PCBioAssay (Primary and confirmatory bioassays) in PubChem [104] (<http://www.ncbi.nlm.nih.gov/pccompound>). In particular, the inactives from primary bioassays against a target formed the “Potential Decoys” data set, while the actives that had gone through both primary and confirmatory bioassays, i.e. being annotated with EC₅₀, were included into “Potential Actives” data set. In addition to that, “Potential Actives” dataset was further refined to reduce “false positives” by assay artifacts filter including “Hill slope filter”, “frequency of hits filter”, and the “autofluorescence and luciferase inhibition filter”. It should be pointed out that MUV only took experimental inactives as the source of decoys, which is a great advantage over other benchmarking sets.

The method to build MUV was the refined nearest neighbor analysis in spatial statistics [105]. First, 17 physicochemical properties were used for calculating pairwise Euclidean distances. Then actives inadequately embedded in decoys were removed from the “Potential Actives” dataset by a “chemical space embedding filter”. In order to ensure spatially random distribution of actives and decoys, the data sets of actives were adjusted to a common level of spread-over while the sets of decoys were chosen with a common level of separation from the actives. These spread or separation levels were controlled by three novel functions, i.e. ΣG , ΣF , ΣS (cf. eqs. 8–10). ΣG was applied to quantify self-similarity of actives while ΣF was applied to measure the degree of separation from the actives. ΣS was used to calculate the level of random distribution of actives and decoys. In the equations below, n is the number of actives while m is the number of decoys. $I_t(i,j)=1$ if the distance of active i to its nearest neighbor active is smaller than a defined distance t . Similarly, $I_t(j,i)=1$ when the

distance of decoy j to its nearest neighbor active is smaller than t . Therefore, $G(t)$ (or $F(t)$) represents the proportion of actives (or decoys) whose distances to the nearest neighbor active are less than t , respectively. As t varies from the minimum to the maximum, ΣG (or ΣF) is the sum of $G(t)$ s (or $F(t)$ s) over a range of distances $t_{(i)}$. ΣS is the difference between ΣG and ΣF .

$$G(t) = \frac{\sum_i I_t(i, j)}{n}; i=1, \dots, n; \sum G = \sum_i G(t_i) \quad (8)$$

$$F(t) = \frac{\sum_i I_t(j, i)}{m}; j=1, \dots, m; \sum F = \sum_i F(t_i) \quad (9)$$

$$\sum S = \sum_i (F(t_i) - G(t_i)) \quad (10)$$

Second, to address the bias of “false negative”, similarity search using simple descriptors and MAX-rule data fusion were performed to retrieve 5 most similar decoys in each MUV decoy set. These decoys were loaded into SciFinder Scholar and checked for their activities against the target. Finally, 30 actives and 15,000 decoys constituted the final benchmark for each target. Based on 18 MUV benchmarking sets, it was confirmed that the MUV strategy reduced effectively the artificial enrichment as well as analogue bias, measured by mean(ROC AUCs) and mean(Retrieval Rate(RTR)). The data sets and MUV decoy maker were also made public at the site of <http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html>. Since then, multiple studies on evaluation of LBVS approaches have been reported based on MUV data sets, including the comparison of similarity coefficients [106], Bayesian Networks [76], similarity search and pharmacophore modeling [77].

2.2.2 DUD LIB V1.0—In order to make DUD benchmarking set suitable for evaluating “scaffold hopping” ability for LBVS approaches, Jahn et al. converted DUD into DUD LIB V1.0 [66]. Its ligands were directly gathered from “DUD clusters” [92]. Due to the difference in property distributions of ligands (i.e. “DUD clusters”) and DUD decoys, MW < 450 and Alog P < 4 were also used as a lead-like filters to screen DUD decoys. The ligands and the remaining decoys constitute the current format of DUD LIB V1.0 which was released at <http://dud.docking.org/> (accessed in Jun. 2014). Since then, it has been used to evaluate optimal assignment LBVS approach developed by Jahn’s group, also to compare 2D fingerprint- and 3D shape-based similarity search methods [56]. This data set ensured property matching feature, though it contains weaknesses similar to DUD.

2.2.3 REPROVIS-DB—REPROVIS-DB was distinct from other typical benchmarking sets in that all of its data were derived from prior successful applications [67]. Its unique feature of reproducibility made it resemble the real-world chemical collection. REPROVIS-DB includes reference compounds, screening database, qualifying hits, LBVS methods (more 3D methods than 2D), compound selection criterion (database ranking or expert knowledge)

and the references for each dataset. The method to build REPROVIS-DB was relatively simple by collecting the above data and preparing the structures for reference compounds and hits. Nevertheless, this benchmarking set reduces the complexity and analogue bias effectively. Ligand enrichment assessments based on REPROVIS-DB help determine whether the new method can reproduce or perform better than the reported LBVS approaches thus show potentials for prospective applications. The data sets are available at <http://www.lifescienceinformatics.uni-bonn.de> (accessed in Jun. 2014). As REPROVIS-DB covers only a small collection of targets and was not extendable to those without any prior LBVS applications, it has not been widely used in the community.

3. Typical benchmarking bias

The benchmarking bias had been discussed intensively during the development of benchmarks [61, 92, 107–109]. The designing concept of benchmarking sets was to mimic the real-world chemical collections of actives and inactives. Ideally, the power of ligand enrichments by VS approaches measured by the benchmarking sets should be consistent with their performances in the real-world prospective applications. However, in many cases their performances are normally overestimated or underestimated from retrospective VS based on benchmarking sets [67]. Such a biased estimation from the benchmarking calculations is known as “benchmarking bias”. Apparently, the causes of the bias lie in the dissemblance of benchmarking sets to real chemical collections. Although the difference is unavoidable, the benchmarking sets can be made to be maximum-unbiased. To this end, it was proposed to reduce three main types of benchmarking biases, i.e. “analogue bias” [68], “artificial enrichment [88]”, and “false negative” (or “latent active” [64]).

The first type of bias is associated with the composition of the ligand set. As shown in Table 1, the ligands in those benchmarking sets were mostly collected from literatures or public databases of various sources. In most medicinal chemistry-related literature [61, 92], structure-activity relationship (SAR) studies based on certain scaffold(s) constitute the main topic. The corresponding ligand sets derived were thus dominant with limited chemotypes. Therefore, in principle, highly similar chemical structures, i.e. analogues, can be easily retrieved from those benchmarking sets. The inclusion of such a large amount of analogues makes ligand enrichment much easier than the real-world screening, thus causes the “analogue bias” during benchmarking evaluations. Cleves and Jain compared the performances of GSIM-2D similarity search with Surflex-Dock based on DUD (PPAR-gamma case which originally contained 83.53% analogues), and found GSIM-2D outperformed Surflex-Dock. However, when further studying the relationship of ROC performances of GSIM-2D to chemotypes, they uncovered that the performance of GSIM-2D decreases along with the increase of chemical diversity. Therefore, they concluded that with this kind of comparison DUD may not reflect realistic performances of approaches in prospective VS campaign due to the high percentage of analogues [61]. However, DUD has been still used to compare LBVS and SBVS approaches in the sequential studies [54, 110]. The results based on DUD show consistently that 2-D fingerprint based LBVS approaches ranked higher than 3-D LBVS approaches [110] as well as SBVS approaches [54]. The analysis of the analogues bias in DUD definitely casts doubt on these types of studies and associated conclusions. Another problem with analogues bias

is that the abilities of scaffold hopping for different approaches would not be able to evaluate with those data sets [92]. Therefore, the analogue bias needs to be dealt with in retrospective VS, in particular in LBVS [111]. To address it, solutions such as structure clustering had been proposed to remove analogues [58, 65, 92].

The other two types of biases are related to the composition of decoys. To avoid “artificial enrichment”, the common practice is now to make decoys resemble ligands in low-dimensional physicochemical properties [57, 88]. To the best of our knowledge, except for bias-uncorrected benchmarking sets such as Rognan’s decoys at the early stage [79], most of currently available benchmarking sets have followed this principle though different low-dimensional physicochemical properties were employed [57, 58, 63, 64, 68]. In comparison to “artificial enrichment”, “False negative” had been a challenging problem to deal with during the buildup of benchmarking sets, which normally lowers the ligand enrichment [64]. In practice, the decoys are selected so as to be topologically dissimilar to all the ligands by applying 2-D dissimilarity filter [57], for which different sets used different metrics to measure the structural dissimilarity [57, 58, 62, 63]. However, it was reported that some decoys in DUD appear to bind to their associated targets [58] even after the filtering steps.

From over a decade of efforts on building and updating SBVS-specific benchmarking sets, it appears that the recent benchmarking sets, i.e. DUD-E and DEKOIS 2.0, have corrected most of the biases associated with prior benchmarking sets. The quality of current SBVS-specific benchmarking sets can be regarded to be close to perfectness, though there seems a need to standardize the practice for decoy-building, e.g. ratio of decoys per ligand and clustering techniques. By contrast, LBVS-specific benchmarking sets have a shorter history, which dates back to 2009. Among them, MUV sets were carefully designed based on a robust theory, i.e. spatial statistics. Other sets were based on relatively simple practices, for instance, by directly collecting data from literature and/or clustering. In addition, there exists a common misconception to match up LBVS with SBVS approaches based on DUD sets, as mentioned by Cleves and Jain [61]. In fact, MUV was supposed to be applicable to SBVS approaches by design. However, the chemical space of its decoys was highly restricted to the experimental inactives from PubChem, which confines its application to more biological targets. Therefore, there are huge demands for unbiased benchmarking sets that are applicable to both LBVS and SBVS approaches, thus making a fair comparison possible.

4. Our benchmarking methodology and its applications to HDACs

4.1 Algorithm description

The unique feature of our in-house method is that it is in pursuit of spatial random distribution of compounds to be included in the decoy set on the premise of good property matching. In SBVS-specific benchmarking sets such as DUD-E [58] and DEKOIS [64], the decoys were prepared to be topologically dissimilar as much as possible to the ligands in order to avoid “false negative”. As such, direct usage of these benchmarking sets can cause biased assessment of LBVS approaches. Specifically, topological similarity (*Tsim*) to the ligands is considered in our method in addition to dissimilarity. To balance both factors, on one hand, the dissimilarity is set to a hypothetical degree, i.e. *Tc* equals to 0.75 based on MACCS fingerprints. This degree is also a common threshold used in GLL/GDD [63] and

Decoy Finder [94] in order to distinguish actives from inactives, thus to minimize the possibility of “false negative”. On the other hand, *Tsim* between a query ligand and its potential decoys is balanced by considering both *Tsim* of other ligands to this ligand as well as *Tsim* of other ligands to potential decoys (cf. Algorithm 1). In this way, the benchmarking sets become more challenging in ligand enrichment for LBVS approaches, in particular for 2-D similarity search. Meanwhile, the decoys in the benchmarking sets are unlikely to be actives; to achieve good property matching (i.e. to avoid “artificial enrichment”), the threshold of a metric of property similarity, *Psim*, is also applied in our in-house method.

The detailed algorithm is represented in the form of pseudocodes, which includes three consecutive parts, i.e. Select Diverse Ligands, Select Potential Decoys and Select Final Decoys (cf. Algorithm 1). In addition to the pseudocodes, we listed important notes here: (1) the input of the algorithm is a ligand set containing all compounds retrieved from literatures but sorted from strong to weak binding affinities according to their IC_{50} values, named as S_{AL} ; (2) *Tsim* is also measured by *Tc* based on MACCS fingerprints; (3) Similar to DUD-E, 6 (*n*) physicochemical properties (or descriptors), i.e. Log*P*, MW, HBAs, HBDs, RBs and net FC, are employed for property matching (*Psim*). They are normalized to [0,1] according to the range of physicochemical properties of final ligands (*FL*); (4) *Psim* is a formula derived from Euclidean distance (*Ed*), i.e. $Psim = 1 - Ed / \sqrt{n}$; (5) The ratio of decoy per ligand is defined as *r*, normally set as from 30 to 100. The algorithm has been implemented on Pipeline Pilot (version 7.5, Accelrys Software, Inc.) [78].

4.2 Case study: HDAC1, HDAC6 and HDAC8 isoforms

4.2.1 Data curation and validation—Human HDACs regulate the homeostasis of histone acetylation and are associated with various diseases such as cancers [112] and Alzheimer’s disease [113]. HDACs inhibitors have emerged to be promising therapies for these diseases. So far, HDACs are classified into 4 subclasses according to domain organization and sequence identity, i.e. Class I, IIA, IIB, III, and IV [114]. Recently, three HDAC isoforms i.e. HDAC1 (Class I), HDAC6 (Class I) and HDAC8 (Class IIB) had received much attention and are currently being intensively studied. Herein, we applied our method to build dual-purpose benchmarking sets for these HDAC isoforms. On one hand, it is to prove the efficacy of our novel methodology to deal with the overestimation issue of LBVS approaches (mainly 2D similarity search). On the other hand, these benchmarking sets for the above targets will be used to determine the optimal VS approaches, in both SBVS and LBVS, for our long-term effort in HDACs drug discovery.

At first, the inhibitors for each target were collected from ChEMBL-2013 database (<https://www.ebi.ac.uk/chembl/>, accessed in Jun. 2014) [101]. Among those compounds, only those with ChEMBL confidence score > 4 and $IC_{50} \leq 1 \mu M$ were retained [58]. Next, salts were removed from those compounds and the stereochemistry/charges of ligand were standardized. The non-drug-like ligands with RBs > 20 or MW > 600 were filtered according to the criterion of DUD-E [58]. Subsequently, different entities (i.e. charged ligands) of those remaining ligands at pH range of 7.3–7.5 were generated. The above ligand processing was performed using a series of components in Pipeline Pilot (version 7.5, Accelrys Software, Inc.), e.g. “Strip Salts”, “Standardize Molecule” and “Custom Filter”.

The data set after ligand processing was one source of input of the implementation. Another input was all purchasable molecules (~18 million) from ZINC at <http://zinc.docking.org/> (accessed in Jun. 2014). In these cases, $r=39$ was used as the ratio of decoys per ligand. Via three steps of our algorithm, diverse ligands and their topologically-optimized decoys were selected to form the final benchmarking sets for each target. Finally, the method was validated by Leave-One-Out (LOO) Cross-Validation (CV) [68]. In each iteration, a ligand was chosen as the query and its decoys were also removed. The remaining ligands and decoys constitute the validation set. Both the query and compounds in the validation set were represented by 6 physicochemical properties or MACCS fingerprints. The *Psim*-based similarity search or *Tsim*-based similarity search was conducted, in which *Psim* or *Tsim* was sorted and the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) were calculated. The validation was repeated multiple times depending on the number of ligands in the data set. After LOO CV, the overall quality of those benchmarking sets was measured by mean(ROC AUCs) from multiple validations. As a supplement to property matching, the distribution curves of each property between ligands and associated decoys were also generated.

4.2.2 Composition and quality of benchmarking sets—In HDAC1 benchmarking sets, there are 180 diverse ligands and 7,020 decoys. For HDAC6 and HDAC8, the sets contain 96 diverse ligands and 3,744 decoys, 39 diverse ligands as well as 1521 decoys, respectively. The property distribution curves (cf. Figure 1) show that 6 physicochemical properties match well between ligands with their corresponding decoys. For all three targets, we found that (1) the physicochemical properties of decoys and ligands span the same range; (2) peaks on the distribution curve of ligands and decoys are within the same interval; (3) some properties of ligands and decoys even show similar fractions within intervals, e.g. HBAs and net FC. ROC curves from *Psim*-based similarity search (cf. Figure 2A) indicate that the curves for all the targets are close to random distribution, i.e. $y=x$, whose AUC equals to 0.5. The mean(ROC AUCs)s are 0.472, 0.471 and 0.454, respectively. Those curves and values of mean(ROC AUCs) indicate that it is challenging to distinguish ligands from decoys based to property similarities. Because of the strict property matching between ligands and decoys, these three sets appear to be applicable to evaluation of SBVS approaches as well.

ROC curves from *Tsim*-based similarity search (cf. Figure 2B) show that most of the curves are close to the random distribution for HDAC1, HDAC6 and HDAC8, though some curves for HDAC8 are slightly distant from the random line. The similar trend is also observed in mean(ROC AUCs)s, which are 0.528 for HDAC1, 0.526 for HDAC6 and 0.652 for HDAC8, respectively. Therefore, it is difficult to enrich ligands from the background of those decoys by simple 2-D similarity searches. This indicates that the enrichment bias related to the composition of decoys has been significantly reduced, which makes it fair to compare SBVS and LBVS approaches using these benchmarking sets.

By comparing mean(ROC AUCs) from *Psim*-based and *Tsim*-based similarity search, it shows those benchmarking sets are more challenging for the former approach. It is expected considering the details of our algorithm design. In Algorithm 1, the priority is placed on property matching rather than topological similarity because the algorithm picks up decoys

of high *Psim* values first. As a result, the final decoys are largely limited to the chemical space of potential decoys defined by the *Psim*-based selection. We plan to improve the algorithm in this regard to obtain an optimal balance between property matching (mainly for SBVS) and topological similarity (mainly for LBVS). In addition, the limit in chemical space of the source of decoys seems to be a common problem for all benchmarking sets so far. Though VDS enlarged the chemical space by ignoring synthetic feasibility [62], it also impaired the similarity between benchmarking sets and the real-world “drug-like” chemical collections. ZINC was widely utilized as the source of decoys because its compounds cover the largest “drug-like” chemical space. In order to improve the quality of decoys, its future version with larger chemical space is needed. Nevertheless, other challenging issues need to be addressed in the future. Our methodology is based on the hypothesis that *Tc* based on MACCS fingerprints equals to or greater than 0.75 can distinguish actives and inactives. Although this is a common practice [63, 94], it is still questionable about how to balance the topological similarity of decoys to ligands while to avoid being actives when accommodating benchmarking sets for both SBVS and LBVS. Recently, benchmarking sets such as DUD-E and MUV started to include experimental decoys [58, 68]. However, the amount of experimental decoys is far less than enough and the associated chemical space is very limited. Only when high throughput screening (HTS) can confirm the activities of all potential decoys, the risk of “false negative” can be avoided. In our Algorithm 1, we employ 6 physicochemical properties for property matching. Similarly, most of the bias-corrected benchmarking sets such as DUD-E and DEKOIS 2.0 used the same set of properties. MUV is different in that it used 17 physicochemical properties. However, the number and types of properties to employ in the practice of benchmarking remain to be addressed.

5. Conclusions

In this article, we reviewed most available benchmarking sets and methods, and discussed the associated biases, i.e. “artificial enrichment”, “analogue bias” and “false negative”. The history of benchmarking sets shows that, from the first bias-uncorrected benchmarking sets to the latest, bias-corrected sets, more metrics and complex algorithms were involved to measure the biases and quality of the benchmarking sets. To date, almost every source of biases for SBVS-specific benchmarking sets have been carefully dealt with and recent SBVS-specific benchmarking sets such as DUD-E, DEKOIS 2.0 are regarded as maximal-unbiased. However, the amount and quality of LBVS-specific benchmarking sets are still limited in despite of the pioneering efforts by MUV sets. We present here a major problem found in normal LBVS-specific benchmarking sets, i.e. “analogue bias” related to the composition of decoys which causes overoptimistic enrichment. This type of bias becomes evident when comparing LBVS and SBVS approaches using SBVS-specific benchmarking sets such as DUD [61]. To reduce such bias, we recently developed a novel algorithm in the premise of good property matching feature. The algorithm employs traditional physiochemical properties and similarity metrics in an explicit and rational combination for filtering based on both properties (*Psim*) and topological metrics (*Tsim*). The validations by LOO CV on benchmarking sets we built for human HDAC1, HDAC6 and HDAC8 show that our algorithm (implemented on Pipeline Pilot) works in both aspects with success, i.e. reducing analogue bias and artificial enrichment significantly. Therefore, we anticipate that

(1) these benchmarking sets can be applied to compare LBVS and SBVS approaches in an unbiased manner; (2) Our method can also be employed to build LBVS-specific benchmarking sets for more popular targets as originally designed. All of our benchmarking data sets for HDACs are available for downloading at <http://xswlab.org/>.

Acknowledgements

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 (5P30A1087714-02), the National Institute on Minority Health and Health Disparities of 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), and National Natural Science Foundation of China (NSFC 81373272, 81172917).

Abbreviations

ACD	advanced chemical directory
AUC	area under curve
DEKOIS	demanding evaluation kits for objective <i>in silico</i> screening
DOE score	deviation from optimal embedding score
DUD	directory of useful decoys
DUD-E	DUD-enhanced
ECFP_4	extended-connectivity fingerprints of maximum diameter 4
Ed	euclidean distance
ERα	estrogen receptor α
FC	formal charge
FCFP_6	function class fingerprints of maximum diameter 6
FL	final ligands
GDD	GPCR decoy database
GLIDA	GPCR-ligand database
GLL	GPCR ligand library
GPCR	G protein-coupled receptor
HBAs	number of hydrogen bond acceptors
HBDs	number of hydrogen bond donors
HDAC	human histone deacetylase
LADS	latent actives in the decoy set
LBVS	ligand-based virtual screening
LOO CV	leave-one-out cross-validation

MDDR	MDL drug data report
ML	machine learning
MUV	maximum unbiased validation
MW	molecular weight
NRLiSt BDB	nuclear receptors ligands and structures benchmarking database
PCBioAssay	primary and confirmatory bioassays
<i>Psim</i>	property similarity
PSS	physicochemical similarity score
RBs	number of rotatable bonds
REPROVIS-DB	database of reproducible virtual screens
ROC	receiver operating characteristic
SBVS	structure-based virtual screening
Tc	Tanimoto coefficient
TK	thymidine kinase
<i>Tsim</i>	topological similarity
VDS	virtual decoy sets
VS	virtual screening
WOMBAT	world of molecular bioactivity

References

1. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. Journal of molecular biology. 1982; 161:269–288. [PubMed: 7154081]
2. Irwin JJ. Journal of computer-aided molecular design. 2008; 22:193–199. [PubMed: 18273555]
3. Ou-Yang SS, Lu JY, Kong XQ, Liang ZJ, Luo C, Jiang H. Acta pharmacologica Sinica. 2012; 33:1131–1140. [PubMed: 22922346]
4. Lavecchia A, Di Giovanni C. Current medicinal chemistry. 2013; 20:2839–2860. [PubMed: 23651302]
5. Braga RC, Andrade CH. Current topics in medicinal chemistry. 2013; 13:1127–1138. [PubMed: 23651486]
6. Ma DL, Chan DSH, Leung CH. Chem. Soc. Rev. 2013; 42:2130–2141. [PubMed: 23288298]
7. Ma XH, Zhu F, Liu X, Shi Z, Zhang JX, Yang SY, Wei YQ, Chen YZ. Current medicinal chemistry. 2012; 19:5562–5571. [PubMed: 23016548]
8. Barbosa AJM, Del Rio A. Current topics in medicinal chemistry. 2012; 12:866–877. [PubMed: 22352914]
9. Ripphausen P, Nisius B, Bajorath J. Drug discovery today. 2011; 16:372–376. [PubMed: 21349346]
10. Waszkowycz B, Clark DE, Gancia E. Wiley Interdiscip. Rev.-Comput. Mol. Sci. 2011; 1:229–259.
11. Schuster D, Wolber G. Current pharmaceutical design. 2010; 16:1666–1681. [PubMed: 20222852]
12. Kim KH, Kim ND, Seong BL. Expert opinion on drug discovery. 2010; 5:205–222. [PubMed: 22823018]

13. Villoutreix BO, Eudes R, Miteva MA. Combinatorial chemistry & high throughput screening. 2009; 12:1000–1016. [PubMed: 20025565]
14. Plewczynski D, Spieser SAH, Koch U. Combinatorial chemistry & high throughput screening. 2009; 12:358–368. [PubMed: 19442065]
15. Melville JL, Burke EK, Hirst JD. Combinatorial chemistry & high throughput screening. 2009; 12:332–343. [PubMed: 19442063]
16. Tuccinardi T. Combinatorial chemistry & high throughput screening. 2009; 12:303–314. [PubMed: 19275536]
17. Ripphausen P, Nisius B, Bajorath J. Drug discovery today. 2011; 16:372–376. [PubMed: 21349346]
18. Bajorath J. Nature reviews. Drug discovery. 2002; 1:882–894.
19. Stahura FL, Bajorath J. Current pharmaceutical design. 2005; 11:1189–1202. [PubMed: 15853666]
20. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. The AAPS journal. 2012; 14:133–141. [PubMed: 22281989]
21. Horvath D. Methods Mol Biol. 2011; 672:261–298. [PubMed: 20838973]
22. Tropsha A, Golbraikh A. Current pharmaceutical design. 2007; 13:3494–3504. [PubMed: 18220786]
23. Willett P. Methods Mol Biol. 2011; 672:133–158. [PubMed: 20838967]
24. Ewing TJ, Makino S, Skillman AG, Kuntz ID. Journal of computer-aided molecular design. 2001; 15:411–428. [PubMed: 11394736]
25. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. Journal of computational chemistry. 2009; 30:2785–2791. [PubMed: 19399780]
26. Rarey M, Kramer B, Lengauer T, Klebe G. Journal of molecular biology. 1996; 261:470–489. [PubMed: 8780787]
27. Jain AN. Journal of medicinal chemistry. 2003; 46:499–511. [PubMed: 12570372]
28. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. Journal of molecular graphics & modelling. 2003; 21:289–307. [PubMed: 12479928]
29. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Proteins. 2003; 52:609–623. [PubMed: 12910460]
30. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Journal of medicinal chemistry. 2004; 47:1739–1749. [PubMed: 15027865]
31. Abagyan R, Totrov M, Kuznetsov D. Journal of computational chemistry. 1994; 15:488–506.
32. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP. Journal of molecular graphics & modelling. 2007; 26:198–212. [PubMed: 16860582]
33. Walker T, Grulke CM, Pozefsky D, Tropsha A. Bioinformatics. 2010; 26:3000–3001. [PubMed: 20889496]
34. Kurogi Y, Guner OF. Current medicinal chemistry. 2001; 8:1035–1055. [PubMed: 11472240]
35. Dixon SL, Smondyrev AM, Rao SN. Chemical biology & drug design. 2006; 67:370–372. [PubMed: 16784462]
36. Wolber G, Langer T. Journal of chemical information and modeling. 2005; 45:160–169. [PubMed: 15667141]
37. Zilian D, Sottriffer CA. Journal of chemical information and modeling. 2013; 53:1923–1933. [PubMed: 23705795]
38. Hsieh JH, Yin S, Wang XS, Liu S, Dokholyan NV, Tropsha A. Journal of chemical information and modeling. 2012; 52:16–28. [PubMed: 22017385]
39. Nunez S, Venhorst J, Kruse CG. Journal of chemical information and modeling. 2010; 50:480–486. [PubMed: 20356089]
40. Hamza A, Wei NN, Zhan CG. Journal of chemical information and modeling. 2012; 52:963–974. [PubMed: 22486340]
41. Perez-Nueno VI, Rabal O, Borrell JI, Teixido J. Journal of chemical information and modeling. 2009; 49:1245–1260. [PubMed: 19364101]

42. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. *Journal of medicinal chemistry*. 2006; 49:5912–5931. [PubMed: 17004707]
43. Kellenberger E, Rodrigo J, Muller P, Rognan D. *Proteins*. 2004; 57:225–242. [PubMed: 15340911]
44. Zhou Z, Felts AK, Friesner RA, Levy RM. *Journal of chemical information and modeling*. 2007; 47:1599–1608. [PubMed: 17585856]
45. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL 3rd. *Journal of medicinal chemistry*. 2004; 47:3032–3047. [PubMed: 15163185]
46. Wang R, Lu Y, Fang X, Wang S. *Journal of chemical information and computer sciences*. 2004; 44:2114–2125. [PubMed: 15554682]
47. Wang R, Lu Y, Wang S. *Journal of medicinal chemistry*. 2003; 46:2287–2303. [PubMed: 12773034]
48. Cheng T, Li X, Li Y, Liu Z, Wang R. *Journal of chemical information and modeling*. 2009; 49:1079–1093. [PubMed: 19358517]
49. Ashtawy HM, Mahapatra NR. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*. 2012; 9:1301–1313.
50. Huang SY, Grinter SZ, Zou X. *Physical chemistry chemical physics : PCCP*. 2010; 12:12899–12908. [PubMed: 20730182]
51. Leach AR, Gillet VJ, Lewis RA, Taylor R. *Journal of medicinal chemistry*. 2010; 53:539–558. [PubMed: 19831387]
52. Chen Z, Li HL, Zhang QJ, Bao XG, Yu KQ, Luo XM, Zhu WL, Jiang HL. *Acta pharmacologica Sinica*. 2009; 30:1694–1708. [PubMed: 19935678]
53. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kretsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD. *Journal of chemical information and modeling*. 2007; 47:1504–1519. [PubMed: 17591764]
54. von Korff M, Freyss J, Sander T. *Journal of chemical information and modeling*. 2009; 49:209–231. [PubMed: 19434824]
55. Sanders MP, Barbosa AJ, Zarzycka B, Nicolaes GA, Klomp JP, de Vlieg J, Del Rio A. *Journal of chemical information and modeling*. 2012; 52:1607–1620. [PubMed: 22646988]
56. Hu G, Kuang G, Xiao W, Li W, Liu G, Tang Y. *Journal of chemical information and modeling*. 2012; 52:1103–1113. [PubMed: 22551340]
57. Huang N, Shoichet BK, Irwin JJ. *Journal of medicinal chemistry*. 2006; 49:6789–6801. [PubMed: 17154509]
58. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. *Journal of medicinal chemistry*. 2012; 55:6582–6594. [PubMed: 22716043]
59. Nicholls A. *Journal of computer-aided molecular design*. 2008; 22:239–255. [PubMed: 18253702]
60. Katritch V, Rueda M, Abagyan R. *Methods Mol Biol*. 2012; 857:189–205. [PubMed: 22323222]
61. Cleves AE, Jain AN. *Journal of computer-aided molecular design*. 2008; 22:147–159. [PubMed: 18074107]
62. Wallach I, Lilien R. *Journal of chemical information and modeling*. 2011; 51:196–202. [PubMed: 21207928]
63. Gatica EA, Cavasotto CN. *Journal of chemical information and modeling*. 2012; 52:1–6. [PubMed: 22168315]
64. Vogel SM, Bauer MR, Boeckler FM. *Journal of chemical information and modeling*. 2011; 51:2650–2665. [PubMed: 21774552]
65. Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM. *Journal of chemical information and modeling*. 2013; 53:1447–1462. [PubMed: 23705874]
66. Jahn A, Hinselmann G, Fechner N, Zell A. *Journal of cheminformatics*. 2009; 1:14. [PubMed: 20150995]
67. Ripphausen P, Wassermann AM, Bajorath J. *Journal of chemical information and modeling*. 2011; 51:2467–2473. [PubMed: 21902278]
68. Rohrer SG, Baumann K. *Journal of chemical information and modeling*. 2009; 49:169–184. [PubMed: 19434821]

69. Zhang X, Wong SE, Lightstone FC. *Journal of chemical information and modeling*. 2014
70. Zhang Y, Yang S, Jiao Y, Liu H, Yuan H, Lu S, Ran T, Yao S, Ke Z, Xu J, Xiong X, Chen Y, Lu T. *Journal of chemical information and modeling*. 2013; 53:3163–3177. [PubMed: 24266594]
71. Sastry GM, Inakollu VS, Sherman W. *Journal of chemical information and modeling*. 2013; 53:1531–1542. [PubMed: 23782297]
72. Schumann M, Armen RS. *Journal of computational chemistry*. 2013; 34:1258–1269. [PubMed: 23420703]
73. Schneider N, Lange G, Hindle S, Klein R, Rarey M. *Journal of computer-aided molecular design*. 2013; 27:15–29. [PubMed: 23269578]
74. Neves MA, Totrov M, Abagyan R. *Journal of computer-aided molecular design*. 2012; 26:675–686. [PubMed: 22569591]
75. Liu X, Jiang H, Li H. *Journal of chemical information and modeling*. 2011; 51:2372–2385. [PubMed: 21819157]
76. Abdo A, Chen B, Mueller C, Salim N, Willett P. *Journal of chemical information and modeling*. 2010; 50:1012–1020. [PubMed: 20504032]
77. Tiikkainen P, Markt P, Wolber G, Kirchmair J, Distinto S, Poso A, Kallioniemi O. *Journal of chemical information and modeling*. 2009; 49:2168–2178. [PubMed: 19799417]
78. Xia J, Jin H, Liu Z, Zhang L, Wang XS. *Journal of chemical information and modeling*. 2014; 54:1433–1450. [PubMed: 24749745]
79. Bissantz C, Folkers G, Rognan D. *Journal of medicinal chemistry*. 2000; 43:4759–4767. [PubMed: 11123984]
80. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. *Journal of medicinal chemistry*. 2004; 47:1750–1759. [PubMed: 15027866]
81. Irwin JJ, Raushel FM, Shoichet BK. *Biochemistry*. 2005; 44:12316–12328. [PubMed: 16156645]
82. McGovern SL, Shoichet BK. *Journal of medicinal chemistry*. 2003; 46:2895–2907. [PubMed: 12825931]
83. Diller DJ, Li R. *Journal of medicinal chemistry*. 2003; 46:4638–4647. [PubMed: 14561083]
84. Pham TA, Jain AN. *Journal of medicinal chemistry*. 2006; 49:5856–5868. [PubMed: 17004701]
85. Wang R, Fang X, Lu Y, Wang S. *Journal of medicinal chemistry*. 2004; 47:2977–2980. [PubMed: 15163179]
86. Irwin JJ, Shoichet BK. *Journal of chemical information and modeling*. 2005; 45:177–182. [PubMed: 15667143]
87. Kuntz ID, Chen K, Sharp KA, Kollman PA. *Proceedings of the National Academy of Sciences of the United States of America*. 1999; 96:9997–10002. [PubMed: 10468550]
88. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, Watson P. *Journal of chemical information and computer sciences*. 2004; 44:793–806. [PubMed: 15154744]
89. Pan Y, Huang N, Cho S, MacKerell AD Jr. *Journal of chemical information and computer sciences*. 2003; 43:267–272. [PubMed: 12546562]
90. Fan H, Irwin JJ, Webb BM, Klebe G, Shoichet BK, Sali A. *Journal of chemical information and modeling*. 2009; 49:2512–2527. [PubMed: 19845314]
91. Brozell SR, Mukherjee S, Balus TE, Roe DR, Case DA, Rizzo RC. *Journal of computer-aided molecular design*. 2012; 26:749–773. [PubMed: 22569593]
92. Good AC, Oprea TI. *Journal of computer-aided molecular design*. 2008; 22:169–178. [PubMed: 18188508]
93. Mysinger MM, Shoichet BK. *Journal of chemical information and modeling*. 2010; 50:1561–1573. [PubMed: 20735049]
94. Cereto-Massague A, Guasch L, Valls C, Mulero M, Pujadas G, Garcia-Vallve S. *Bioinformatics*. 2012; 28:1661–1662. [PubMed: 22539671]
95. Olah, MMM.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, TI. WOMBAT: World of Molecular Bioactivity. In: Oprea, TI., editor. *Cheminformatics in Drug Discovery*. New York: Wiley-VCH; 2004. p. 223-239.

96. Wallach I, Jaitly N, Nguyen K, Schapira M, Lilien R. *Journal of chemical information and modeling*. 2011; 51:1817–1830. [PubMed: 21699246]
97. Okuno Y, Yang J, Taneishi K, Yabuuchi H, Tsujimoto G. *Nucleic acids research*. 2006; 34:D673–D677. [PubMed: 16381956]
98. Ben Nasr N, Guillemain H, Lagarde N, Zagury JF, Montes M. *Journal of chemical information and modeling*. 2013; 53:293–311. [PubMed: 23312043]
99. Lagarde N, Ben Nasr N, Jeremie A, Guillemain H, Laville V, Labib T, Zagury JF, Montes M. *Journal of medicinal chemistry*. 2014
100. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. *Nucleic acids research*. 2007; 35:D198–D201. [PubMed: 17145705]
101. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. *Nucleic acids research*. 2012; 40:D1100–D1107. [PubMed: 21948594]
102. Bemis GW, Murcko MA. *Journal of medicinal chemistry*. 1996; 39:2887–2893. [PubMed: 8709122]
103. Coleman RG, Carchia M, Sterling T, Irwin JJ, Shoichet BK. *PloS one*. 2013; 8:e75992. [PubMed: 24098414]
104. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH. *Nucleic acids research*. 2012; 40:D400–D412. [PubMed: 22140110]
105. Rohrer SG, Baumann K. *Journal of chemical information and modeling*. 2008; 48:704–718. [PubMed: 18380447]
106. Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P. *Journal of chemical information and modeling*. 2012; 52:2884–2901. [PubMed: 23078167]
107. Hawkins PC, Warren GL, Skillman AG, Nicholls A. *Journal of computer-aided molecular design*. 2008; 22:179–190. [PubMed: 18217218]
108. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T. *Journal of computer-aided molecular design*. 2008; 22:213–228. [PubMed: 18196462]
109. Nicholls A. *Journal of computer-aided molecular design*. 2008; 22:239–255. [PubMed: 18253702]
110. Venkatraman V, Perez-Nueno VI, Mavridis L, Ritchie DW. *Journal of chemical information and modeling*. 2010; 50:2079–2093. [PubMed: 21090728]
111. Jain AN, Nicholls A. *Journal of computer-aided molecular design*. 2008; 22:133–139. [PubMed: 18338228]
112. West AC, Johnstone RW. *The Journal of clinical investigation*. 2014; 124:30–39. [PubMed: 24382387]
113. Guan JS, Haggarty SJ, Giacometti E, Dannenberg JH, Joseph N, Gao J, Nieland TJ, Zhou Y, Wang X, Mazitschek R, Bradner JE, DePinho RA, Jaenisch R, Tsai LH. *Nature*. 2009; 459:55–60. [PubMed: 19424149]
114. Roper S, Esteller M. *Molecular oncology*. 2007; 1:19–25. [PubMed: 19383284]
115. Karthikeyan, M.; Vyas, R. *Practical Chemoinformatics*. First ed.. New York: Springer; 2014.

Fig. 1.
Property distributions of ligands (blue) and decoys (red) in our LBVS-specific benchmarking sets for human HDAC1, HDAC6 and HDAC8, respectively.

Fig. 2. ROC curves and mean(ROC AUCs) from (A) *Psim*-based similarity search with six physicochemical properties and (B) *Tsim*-based similarity search with MACCS fingerprints against our benchmarking data sets for human HDAC1, HDAC6 and HDAC8, respectively.

Table 1

List of current benchmarking data sets, made available from years of 2000 to 2014. The source of ligands, decoys, ratio of decoys per ligand and public access address are included.

Publication Year	Name	Author(s)	Source of ligands	Source of decoys	Ratio of decoys per ligand	Public access (accessed in Jun. 2014)
2000–11	Rognan's decoy set	Rognan	literature	Advanced Chemical Directory (ACD)	undefined	http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html
2002–09	Shoichet's decoy set	Shoichet	literature	MDL Drug Data Report (MDDR) ver2000.2	undefined	
2003–06	Diller's decoy set	Diller	literature	MDL Drug Data Report (MDDR)	undefined	
2005–06	Jain's decoy set	Jain	PDBbind database	ZINC "drug-like" subset	undefined	http://www.jainlab.org/downloads.html
2006–10	Directory of Useful Decoys (DUD)	Irwin & Shoichet	literature, PDBbind, etc.	ZINC "drug-like" subset	36	http://dud.docking.org/
2007	Directory of Useful Decoys (DUD) Clusters	Good & Oprea	DUD ligands	DUD decoys	undefined	http://dud.docking.org/
2007	WOMBAT datasets	Good & Oprea	WOMBAT	DUD decoys	undefined	http://dud.docking.org/
2010	Charge Matched DUD	Shoichet	DUD	ZINC database	36	http://dud.docking.org/
2011–01	Virtual Decoy Sets (VDS)	Wallach and Lilien	DUD	virtual molecules	36	http://compbio.cs.toronto.edu/VDS
SBVS	Demanding Evaluation Kits for Objective <i>in Silico</i> Screening (DEKOIS)	Boeckler	BindingDB	ZINC database (version 8, November 2009)	30	http://www.dekois.com/dekois_orig.html
	G protein-coupled receptor (GPCR) Ligand Library (GLL) / Decoy Database (GDD)	Cavasotto	GLIDA	ZINC database (accessed in Jan. 2011)	39	http://cavasotto-lab.net/Databases/GDD/Download/
2012-01	DecoyFinder	Cereto-Massague & Garcia-Vallve	user-supplied	ZINC database or others	any (user-defined)	http://urvnutrigenomica-ctms.github.io/DecoyFinder/
2012-06	Directory of Useful Decoys Enhanced (DUD-E)	Irwin & Shoichet	ChEMBL	ZINC	50	http://dude.docking.org/
2013-05	DEKOIS 2.0	Boeckler	BindingDB	ZINC	30	http://www.dekois.com/
2014-01	Nuclear Receptors ligands and structures benchmarking database (NRLiSt BDB)	Matthieu Montes	ChEMBL database or the Pubchem database	DUD-E decoy maker & ZINC	50	http://nrlist.drugdesign.fr
2009-01	Maximum Unbiased Validation (MUV)	Baumann	PubChem (actives with EC ₅₀)	PubChem (inactives)	500	http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html
LBVS	DUD LIB V1.0	Good and Oprea	DUD clusters	DUD decoys	undefined	http://dud.docking.org/
	DataBase of REProducible Virtual Screens (REPROVIS-DB)	Bajorath	literature (quantifying hits)	literature (screening database)	undefined	http://www.lifescienceinformatics.uni-bonn.de

Algorithm 1

Our methodology to build LBVS-specific benchmarking set. All Ligand set is defined as $S_{AL}[]$, ZINC set is defined as $S_z[]$, Final Ligand set is defined as $S_{FL}[]$, Potential Decoy set is defined as $S_{PD}[]$, output of All Final Decoy set is defined as $S_{AFD}[]$, *Index* is the unique identity of each compound for selection purpose.

Step 1. Select Diverse Ligands ($S_{AL}[]$)

```

1:  $Tsim_{AL}[] \leftarrow$  Tanimoto coefficient  $\{S_{AL}, S_{AL}\}$  // get topological similarity between all ligand pairs, temporary ID for each ligand is  $i$  (target ligand set) and  $j$  (reference ligand set),  $i, j = 1, 2, \dots, N_{AL}$ 
2:  $Tsim_{AL}[] \leftarrow$  set element  $Tsim_{AL}(i,j)=a$  ( $a < 0$ ) where  $i=j$  // set self-similarity to a negative value
3:  $rval \leftarrow 1$  // set initial ID of current target ligand to 1
4: while size  $rval \neq 0$  do
5:    $Pair[] \leftarrow$  get  $(i,j)$  of  $Tsim_{AL}[]$  0.75 // get IDs of two ligands whose similarity is over the threshold (0.75)
6:    $rval \leftarrow$  get  $i$  of the first element in  $Pair[]$  // get index of current target ligand based on  $Pair[]$ 
7:   if  $Pair[]$  null do
8:      $Tsim_{AL}[] \leftarrow$  set  $Tsim_{AL}(i,j)=0$  where  $i=rval$  or  $j=rval$  // set similarity of current target ligand to any reference ligand to 0
9:   end if
10: end while
11:  $Index_{SL}[] \leftarrow$  get index of  $Tsim_{AL,AL}(i,j) = 0$  where  $j=1$  // get indexes of all ligands whose mutual similarity is within threshold
12:  $S_{FL}[] \leftarrow S_{AL}[Index_{SL}[]]$  // the final ligand set with selected diverse ligands
13: return  $S_{FL}[]$ 

```

Step 2. Select Potential Decoys ($S_{FL}[], S_z[]$)

```

1:  $P_{FL}[] \leftarrow$  physicochemical properties  $\{S_{FL}\}$  // calculate physicochemical properties of Final Ligands
2:  $P_z[] \leftarrow$  physicochemical properties  $\{S_z\}$  // calculate physicochemical properties of ZINC compounds
3:  $P_{FL\_R}[] \leftarrow$  get ranges of each property based on  $P_{FL}[]$  // get the maximum and minimum values for each property
4:  $Index_z[] \leftarrow$  get index of  $P_z[]$  within  $P_{FL\_R}[]$  // get the indexes of compounds within all physicochemical properties threshold
5:  $S_{zs}[] \leftarrow S_z[Index_z[]]$  // ZINC Subset with selected compounds
6:  $Tsim_{FL}[] \leftarrow$  Tanimoto coefficient  $\{S_{FL}, S_{FL}\}$  // calculate mutual topological similarity within Final Ligands
7:  $Tsim_{FL\_min} \leftarrow$  get min of  $Tsim_{FL}[]$  // get the minimum of all values based on  $Tsim_{FL}[]$ 
8:  $Tsim_{FL\_zs}[] \leftarrow$  Tanimoto coefficient  $\{S_{FL}, S_{zs}\}$  // calculate mutual topological similarity between Final Ligands and ZINC Subset compounds
9:  $Tsim_{FL\_zs\_maxmin}[] \leftarrow$  get all mins and maxs of  $Tsim_{FL\_zs}[]$  // get min and max similarity values to all Final Ligands for each ZINC Subset compound
10:  $Index_{zs}[] \leftarrow$  get index of  $Tsim_{FL\_min}$   $Tsim_{FL\_zs\_maxmin}[] < 0.75$  // get indexes of compounds within topological similarity threshold (max < 0.75 & min  $Tsim_{FL\_min}$ )
11:  $S_{PD}[] \leftarrow S_{zs}[Index_{zs}[]]$  // Potential Decoys with selected compounds from ZINC Subset
12: return  $S_{PD}[]$ 

```

Step 3. Select Final Decoys ($S_{FL}[], S_{PD}[]$)

```

1:  $P_{PD}[] \leftarrow$  physicochemical properties  $\{S_{PD}\}$  // calculate physicochemical properties of Potential Decoys
2:  $P_{FL\_norm}[] \leftarrow$  normalize  $P_{FL}[]$  // normalize the values into [0,1] based on  $P_{FL\_R}[]$ 
3:  $P_{PD\_norm}[] \leftarrow$  normalize  $P_{PD}[]$  // normalize the values into [0,1] based on  $P_{FL\_R}[]$ 
4:  $N_{FL} \leftarrow$  size  $S_{FL}[]$  // count the number of Final Ligands
5:  $Index_{FL}[] \leftarrow$  get index of  $S_{FL}[]$  // get indexes of Final Ligands

```

```

6:  $Index_{PD}[] \leftarrow$  get index of  $S_{PD}[]$  // get indexes of Potential Decoys
7:  $S_{AFD}[] \leftarrow \text{null}$  // define a dataset of All Final Decoys
8: for  $i \leftarrow 1$  to  $N_{FL}$  do //  $i$  is temporary ID for each Final Ligand
9:    $Index_{AFD}[] \leftarrow$  get index of  $S_{AFD}[]$  // get indexes of already-made decoys in All Final Decoy set
10:   $S_{PDnew}[] \leftarrow S_{PD}[Index_{PD} - Index_{AFD}]$  // select potential decoys that are not in All Final Decoy set to form a new Potential Decoy set
11:   $Index_{CURR}[] \leftarrow$  get index of ligand  $i$  // get index of the current ligand
12:   $S_{CURR}[] \leftarrow S_{FL}[Index_{FL}] = Index_{CURR}[]$  // ligand set containing current selected ligand
13:   $P_{CURR\_norm}[] \leftarrow$  get normalized properties of  $S_{CURR}$  // get the normalized properties of the current ligand from  $P_{FL\_norm}[]$ 
14:   $Psim_{PDnew,CURR}[] \leftarrow$  property similarity  $\{S_{PDnew}[], S_{CURR}[]\}$  // calculate property similarity between potential decoys and current ligand based on  $P_{CURR\_norm}[]$  and  $P_{PDnew\_norm}[]$ 
15:  for  $k \leftarrow 1$  to 10 do //  $k$  controls different thresholds
16:     $Thr(k) = 1 - 0.05 \times (k - 1)$  // set the  $Psim$  threshold based on  $k$ 
17:     $Num\_thr(k) \leftarrow$  count number of  $Psim_{PDnew,CURR}[]$   $Thr(k)$  // count numbers of potential decoys whose  $Psim$  over each threshold
18:  end for
19:   $cutoff\_k \leftarrow$  get  $k$  of  $Num\_thr(k)$   $r$  &  $Num\_thr(k+1) > r$  // get  $k$  when number of potential decoys within threshold
20:   $cutoff\_max \leftarrow Thr(cutoff\_k)$  // get the upper threshold of  $Psim$ 
21:   $cutoff\_min \leftarrow Thr(cutoff\_k + 1)$  // get the lower threshold of  $Psim$ 
22:   $Index_{PD1}[] \leftarrow$  get index of  $Psim_{PDnew,CURR}[]$   $cutoff\_max$  // get index of the 1st pool of selected potential decoys within threshold
23:   $S_{FD1}[] \leftarrow S_{PD}[Index_{PD1}[]]$  // Final Decoy set 1
24:   $Index_{PD2}[] \leftarrow$  get index of  $cutoff\_max$   $Psim_{PDnew,CURR}[]$   $cutoff\_min$  // get index of the 2nd pool of potential decoys within threshold
25:   $S_{FD2\_mid} \leftarrow S_{PD}[Index_{PD2}[]]$  // Final Decoy set 2 intermediate
26:   $S_{FL\_other} \leftarrow S_{FL}[Index_{FL}] - Index_{CURR}[]$  // the set of other ligands except for the current one
27:   $Tsim_{FL\_other, CURR}[] \leftarrow$  Tanimoto\_coefficient  $\{S_{FL\_other}, S_{CURR}\}$  // calculate mutual topological similarity between other final ligands and current final ligand
28:   $Tsim_{FL\_other, FD2\_mid}[] \leftarrow$  Tanimoto coefficient  $\{S_{FL\_other}, S_{FD2\_mid}\}$  // calculate mutual topological similarity between other final ligands and each compound in Final decoys 2 intermediate
29:   $Tsim_{FD2\_mid, CURR}[] \leftarrow$  get values of  $|Tsim_{FL\_other, FD2\_mid}[] - Tsim_{FL\_other, CURR}[]|$  //  $Tsim$  difference
30:   $S_{Tsim_{FD2\_mid, CURR}}[] \leftarrow$  sort  $Tsim_{FD2\_mid, CURR}[]$  // sort the topological difference from low to high
31:   $Index_{FD2}[] \leftarrow$  get indexes of the top ( $r$ -size  $S_{FD1}$ ) values of  $S_{Tsim_{FD2\_mid, CURR}}[]$  // get indexes of the top ( $r$ -size  $S_{FD1}$ ) decoys from Final Decoy set 2 intermediate
32:   $S_{FD2}[] \leftarrow S_{PD}[Index_{FD2}[]]$  // Final Decoy set 2
33:   $S_{FD}[] = S_{FD1} \cup S_{FD2}$  // Final Decoy set including Final Decoy set 1 and Final Decoy set 2
34:   $S_{AFD}[] = S_{AFD}[] \cup S_{FD}[]$  // All Final Decoy set containing separate final decoys for each ligand
35: end for
36: return  $S_{AFD}[]$ 

```