JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

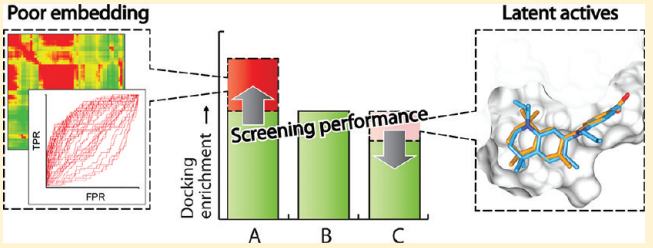
# DEKOIS: Demanding Evaluation Kits for Objective *in Silico* Screening — A Versatile Tool for Benchmarking Docking Programs and Scoring Functions

Simon M. Vogel,[†,‡] Matthias R. Bauer,[†,‡] and Frank M. Boeckler[*,†]

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

Supporting Information

**ABSTRACT:** For widely applied *in silico* screening techniques success depends on the rational selection of an appropriate method. We herein present a fast, versatile, and robust method to construct *demanding evaluation kits for objective in silico screening* (DEKOIS). This automated process enables creating tailor-made decoy sets for any given sets of bioactives. It facilitates a target-dependent validation of docking algorithms and scoring functions helping to save time and resources. We have developed metrics for assessing and improving decoy set quality and employ them to investigate how decoy embedding affects docking. We demonstrate that screening performance is target-dependent and can be impaired by latent actives in the decoy set (LADS) or enhanced by poor decoy embedding. The presented method allows extending and complementing the collection of publicly available high quality decoy sets toward new target space. All present and future DEKOIS data sets will be made accessible at www.dekois.com.




## INTRODUCTION

The concerted use of computer-based methods to screen large compound libraries against biological targets is referred to as *in silico* screening or virtual screening since the late 1990s.[1] Success stories have been reported demonstrating that *in silico* screening is able to provide valuable contributions to the identification of new lead compounds in drug discovery.[2−8] However, the success rate of this approach is difficult to determine based on literature data, because of the "publication bias" favoring positive results over failure. Furthermore, when applied to novel targets, it is hardly feasible to predict the performance of a virtual screening method prior to experimental verification. Possible failure can be due to the particular design and use of *in silico* screening methods or due to poor druggability.[9,10] While methods can be optimized, poor druggability is an intrinsic problem that cannot be overcome.

*In silico* screening is a knowledge-driven approach. It strongly depends on the amount and quality of structural information available. There are ligand-based screening techniques, which use either one-dimensional (1D) filters (e.g., physicochemical properties of known actives), two-dimensional (2D) filters (e.g., substructure matching), or three-dimensional (3D) filters (e.g., pharmacophore filters) and structure-based screening methods, which rely on information about the three-dimensional structure of the target. Docking programs predict the binding mode and the interaction quality of a ligand bound to a receptor and belong to the most popular tools among structure-based methods. A substantial number of docking programs with an even larger number of docking algorithms and scoring functions have emerged in the last decades ever since the first docking program was written.[11−13] Fortunately, this leaves much choice to the user but also requires the user to make a good decision based on rational considerations. The power of docking programs to discriminate between actives and inactives is extensively studied in many evaluation studies,[12] as this is the crucial criterion describing screening performance. However, meaningful benchmarking of docking programs is methodologically not trivial.[14−20] Also, there is evidence that performance of docking programs is target-dependent.[21−23] Therefore, it appears desirable to facilitate a target-specific tailor-made evaluation to assess the quality of the docking tools and scoring functions or combinations thereof.

One of the most crucial issues in conducting an evaluation study for docking programs is the design of the test set, which consists of known actives and putative inactives, often referred to as decoys. Significant discrepancies in molecular weight distribution between actives and decoys can make docking results appear to be artificially good.[24,25] A useful docking program should however be able to discriminate between actives and inactives solely on the basis of structural information, even if there is no difference in low-dimensional properties between actives and inactives. The physicochemical properties of the actives should therefore be mimicked by the decoys.[26,27]

In recent years, several efforts have been made to generate benchmarking sets for docking programs. The underlying methods to generate benchmarking sets differ substantially.[28−31]

**Received:** April 1, 2011
**Published:** July 21, 2011

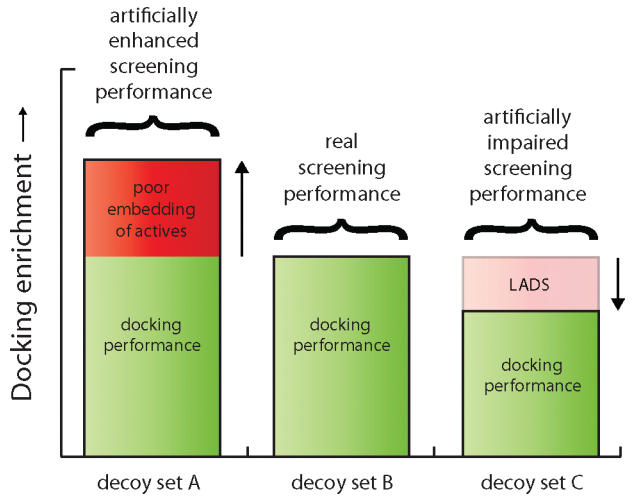



**Figure 1.** Proposed effect of decoy set composition on the apparent docking performance. Poor physicochemical matching of actives with decoys can make the docking performance look artificially good (decoy set A), whereas the presence of latent actives in the decoy set (LADS) can lead to an artificially impaired performance (decoy set C). It should be noted that under certain circumstances a poor embedding of actives can also lead to an impaired screening performance. This happens sometimes, when the actives are embedded within much larger decoys.

A recent method that accounts for the avoidance of analogue bias and artificial enrichment employing spatial statistics is the refined nearest neighbor analysis (MUV Data Sets).[32,33] Baumann et al. applied their method on PubChem[34] bioactivity data yielding a collection of benchmark sets with active and inactive status for each compound.[27] Quite recently, Wallach et al. introduced their virtual decoy sets, which are based on *in silico* decoy generation.[35] One of the biggest efforts to provide challenging but fair benchmarking sets is the DUD data set.[36] It provides 36 decoys for each active against 40 targets. It has been shown that enrichment factors for decoys from the DUD test sets are lower compared to earlier methods to construct decoy sets, indicating that biased information regarding low-dimensional properties was reduced. This makes the DUD data set a useful and frequently applied tool to evaluate the quality of docking programs. However, these ready-made test sets do not allow evaluation of screening tools for targets not included in the DUD data set. As the matching between physicochemical properties of actives and decoys in the DUD is done using an elaborate, multistep protocol, the expendability toward novel targets is necessarily restricted. We aim to overcome this limitation of the DUD by creating an efficient and versatile method which automatically generates decoy sets of good quality.

Beneath effects leading to poor embedding of actives within the decoy set (Figure 1, decoy set A), which artificially enhance the observable docking performance over the real docking performance, other effects can cause an artificial impairment of the observed docking performance (Figure 1, decoy set C). The presence of *latent actives in the decoy set* (LADS) which essentially are false false positives constitutes such an effect that systematically decreases docking performance. Therefore LADS need to be avoided whenever possible by suitable methods. It should be noted that the dimension of the LADS issue can obviously depend on the degree of matching between biological target space and chemical compound space. Thus, in target classes where various bioactive "privileged structures" occur frequently in usual screening libraries, the LADS phenomenon is more often perceived.

Based on the putative pitfalls in the construction of decoy sets as discussed above, we postulate that a test set should meet the following criteria, which we implement in our protocol:

I Actives and decoys should not be easily separable by low dimensional filters
II Latent actives in the decoy set (LADS) should be avoided whenever possible
III To keep the separation of actives and decoys challenging, the structural diversity of decoys should be maximized
IV The decoy selection should be tailor-made for each single active to ensure a good and equal representation of the physical property space of each active by the same number of decoys ("embedding of actives")

In addition, we will introduce several metrics that help us to assess and improve the decoy set quality and optimize the active set used. We postulate that optimization of the active set can be another crucial step to ensure a reasonable evaluation of the docking performance (i.e., to strongly differentiate between real and artificial docking enrichments). When matching decoys to actives, the composition of the active set can greatly influence the quality of the decoy set. The more structural or physicochemical redundancy there is in the active set, the more difficult it becomes to achieve a challenging set of decoys. Such redundancy is usually caused by publications of elaborate structure−activity relationship studies in prevalent lead optimization strategies − a phenomenon commonly referred to as "analog bias".[37−39]

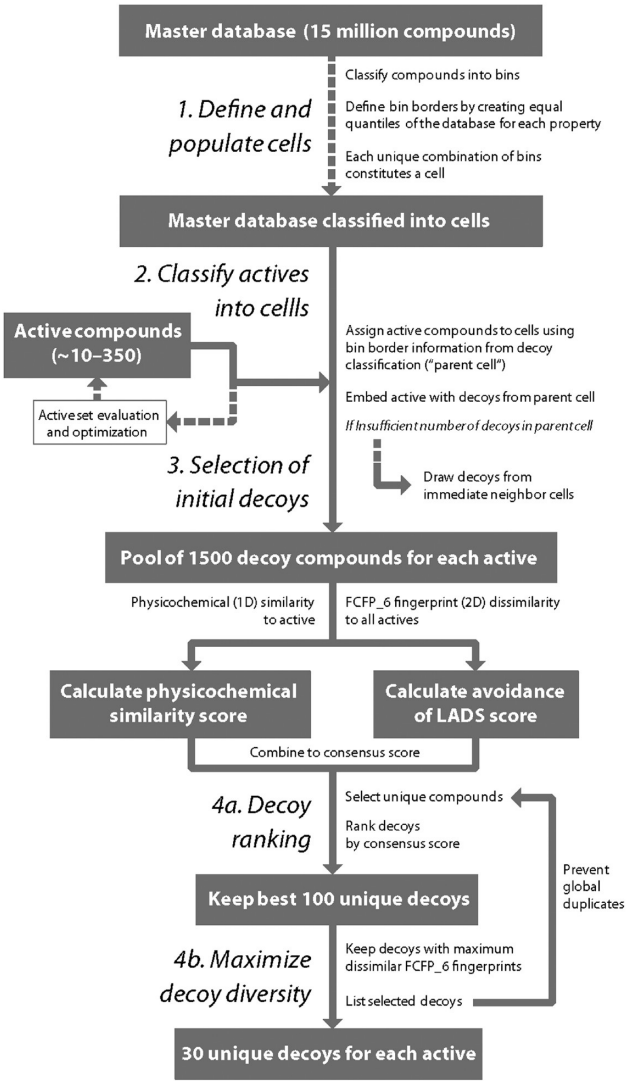
## METHODS

**Construction of the Master Database.** A database of potential decoy compounds (master database) was generated from the ZINC database subset #10 "everything" (version 8, November 2009) comprising more than 21 million compounds.[40] From these compounds, we removed counter ions, molecules with non-organic atoms (atoms other than H, C, N, O, P, S, F, Cl, Br, I) and excluded molecules with extremely high or low values for the physicochemical properties molecular weight (MW > 1000), octanol-water partition coefficient (logP < −8 or logP > 10), number of hydrogen bond acceptors (HBA > 20), number of hydrogen bond donors (HBD > 20), and number of rotatable bonds (RB > 20). Subsequently, the master database size has been reduced to a technically feasible size of 15 million molecules by randomly removing compounds. Molecular weight and number of rotatable bonds are calculated using Pipeline Pilot (version 6.1.5.0., Accelrys Software, Inc.) built-in functions. Octanol-water partition coefficient, number of hydrogen bond acceptors, and number of hydrogen bond donors are calculated using JChem Pipeline Pilot components (version 5.3, ChemAxon Ltd.) with pH set to 7.4. Based on the evaluation of various microspecies and their proportion at a given pH in this component, floating point values are retrieved for HBD and HBA as well.

**Construction of Decoy Sets.** The generation of DEKOIS is an automated workflow, processed by a protocol designed for the software platform Pipeline Pilot (see Scheme 1). The protocol basically consists of four consecutive steps, of which the first step is optional if processed once.

In the first step, each compound in a database of 15 million molecules (master database) is classified into bins by evaluating five physical properties: molecular weight (12 bins), octanol−water partition coefficient (8 bins), number of hydrogen bond acceptors (4 bins), number of hydrogen bond donors (4 bins),

**Scheme 1. Workflow for the Automated Construction of Decoy Sets**




and number of rotatable bonds (7 bins). The classification into bins is done in a way ensuring that each designated physicochemical property bin has the same population of bin members (see Supporting Information Figure S1). The member with the highest value in each bin defines the border of a bin. In a following step, five-dimensional bins (cells) are created for each combination of the five possible physicochemical property bins, forming a total of 10,752 cells. Each compound's physicochemical properties are now characterized by its cell number. Once the master database is classified into cells, this information can be used for further construction of decoy sets.

A subsequent statistical analysis of the physicochemical properties distribution of all bioactive compounds stored in the publicly available database BindingDB[41] has revealed a substantially higher number of hydrogen bond donors (HBD) for bioactive molecules compared to ZINC molecules. To account for this systematical bias, we have decreased the number of compounds with a low count of HBD by randomly deleting such compounds from heavily populated cells (~1 million compounds). Subsequently, we have augmented the database with ZINC compounds possessing a high count of HBD until the final size of 15 million compounds was reached. This is done for technical reasons to avoid instabilities and to ensure robust performance of the protocol, while facilitating the quality of the physicochemical matching on a consistently high level. We believe that exchange of less than 7% of the entire database to compensate for unequal cell population will improve the diversity of the master database and is unlikely to introduce substantial bias.

For the second step, the protocol user provides a set of bioactive molecules. These actives are then assigned to cells according to the previously established bin border information which explicitly defines each cell. We have implemented an optional "active set optimization" function, which can help to resolve issues with heavily clustered active sets that can occur due to the "analog bias". In principle, we reduce (a) structural redundancy of the actives as calculated with functional class fingerprints and (b) physicochemical redundancy by removing actives from cells, which are strongly populated by the active set. Avoiding clustered actives makes it easier to find suitable decoys that homogenously embed the remaining actives and, thus, increases the decoy set quality.

In the third step, for each active 1500 decoys are preselected as an initial pool of decoys, which is subject to further refinement. It should be noted that this and all following steps are iteratively processed for every single active. In this preselection process, decoys that match the cell of the active are randomly selected until 1500 decoys are assigned to each active. If the cell of the active ("parent cell") does not contain enough compounds to provide the necessary amount of decoys, cells that differ in only one physicochemical property bin from the parent cell ("direct neighbor cells") are employed to supply the remaining decoys. Because of possible differences in population density of these adjacent cells, we first select a random adjacent cell and then a random decoy from this cell. This is important to ensure that decoys are picked uniformly around the physicochemical property space of the active. If the parent cell and the direct neighbor cells cannot provide the necessary amount of decoys together, the initial pool of decoys will be reduced for this particular active.

In the fourth step, the initial pool of 1500 decoys per active is refined to deliver a final set of 30 decoys per active. This decoy set is optimized regarding two features: high physicochemical similarity between actives/decoys and the absence of latent actives in the decoy set (LADS). To estimate the similarity, the difference between each active and each decoy from the initial pool in terms of MW, logP, number of HBA, number of HBD, and number of RB is calculated and summarized in a *physicochemical similarity score (PSS)*. This score consists of the arithmetic mean of the normalized similarity scores for each property. Normalization in the step is required because large numerical differences in physicochemical properties would lead to an unbalanced *PSS*.

First, for every active and every property, the decoys with the minimal and maximal property value are found in the corresponding pool of decoys. Then the distance between the active and the minimum as well as between the active and the maximum is calculated. The larger difference is retrieved as shown in eq 1

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

with $x_{j,ref}$ being the value of physicochemical property j for the reference active *ref*, while $x_{j,min}$ and $x_{j,max}$ are the minimum and maximum values for the property j, respectively. $f_j$, being the bigger interval of $|x_{j,ref} - x_{j,min}|$ and $|x_{j,ref} - x_{j,max}|$, is used for normalization of the distance between every decoy and the active to a value between 0 and 1, although usually the distance between maximum and minimum is employed as the denominator. This is done to account for cases where the active might be located outside the span of the maximum and minimum, which essentially would lead to values exceeding 1.

The resulting relative distance (0 = smallest, 1 = largest) is converted into a similarity score $\delta_{i,j}$ by inverting the scale (0 = lowest similarity, 1 = highest similarity) as shown in eq 2

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

With $\delta_{i,j}$ being the difference in physicochemical property j of compound i to the active *ref* and $x_{i,j}$ being the value of physicochemical property j for compound i. The *physicochemical similarity score* (*PSS*) is the arithmetic mean in normalized differences of compound i to the active *ref* in terms of the physicochemical properties j with *n* being the total number of physicochemical properties

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

To avoid the occurrence of LADS, functional class fingerprints (FCFP_6) for all decoys are calculated and disassembled into their fingerprint bit strings. We generated fingerprint bit strings of all actives in the same way and build up a pool of bit strings that potentially represent bioactive substructures ("negative list"). For each decoy all bit strings from the negative list are removed from its own bit strings. The number of remaining bit strings is divided by the number of atoms in the decoy yielding the *avoidance of LADS score*. The more fingerprint bit strings one decoy has in common with the active set the lower its *avoidance of LADS score*.

In the initial pool of decoys maximal and minimal *avoidance of LADS scores* as well as minimal and maximal *PSS* are used to normalize respective score values for each decoy. Both scores are then combined to a consensus score. Based on this consensus score, the 100 best ranked decoys per active are selected. The final set of 30 decoys per active is then selected by the maximum dissimilarity method based on the decoys' functional class fingerprints (FCFP_6) to enhance structural diversity in the decoy sets. To omit duplicates in the final decoy set, the final 30 decoys are not considered for the selection of the 100 best decoys for all following actives.

**Molecule and Binding Pocket Preparation for Docking.** To prepare molecules for docking we used the MOE suite's[42] "molecule wash" function to deprotonate strong acids and protonate strong bases. Energy minimization of all molecules was then performed by using the MMFF94x force field at a gradient of 0.001 rmsd. Existing chirality was preserved and partial charges were calculated according to the parameters of the force field. As described within the DUD errata section,[43] FXA, thrombin, and trypsin active sets contain erroneous molecular data. We employed a Pipeline Pilot protocol to repair these ligand sets (see the Supporting Information).

All crystal structures were downloaded from the Protein Data Bank (PDB).[44] Identical and redundant protein chains with nonessential cofactors, ions, water molecules and ligands were discarded. Subsequently, protonation of the protein−ligand complex was performed with the MOE 'Protonate 3D' function at standard settings ($T$ = 300 K, pH = 7.0, ionic strength $I$ = 0.1 mol/L). The binding mode of the ligand was further investigated with respect to water, metal ion, or cofactor interactions. Cofactors and metal ions were treated as part of the protein. Essential water molecules which interact with the ligand were specifically included in the docking run setup.

**Docking Experiments.** All docking experiments were performed using the docking program GOLD v3.2.[45] in combination with the scoring functions Goldscore[46] and Chemscore[47] at default parameters (van der Waals = 6.0, Hydrogen Bonding = 3.0). The search efficiency for the genetic algorithm was increased from standard 100% to 200% at automatic mode. Binding site residues were defined by specifying crystal structure ligand coordinates, the active site radius setting remained at a default value of 10 Å with the 'detect cavity' option enabled. GOLD set atom types for the ligands and terminated a docking run early if the top 3 solutions were within 1.5 Å rmsd. Water molecules were specified in GOLD by switching state settings to 'toggle' and orientation mode to 'spin'. For each target, we employed a test docking of the ligand from the crystal structure into its binding pocket. If pose retrieval was unsatisfying we optimized water handling, metal coordination, and binding site definitions and continued with the best settings in respect of pose retrieval.

**Metrics for Quality Assessment of Decoy Sets.** The *deviation from optimal embedding score* (DOE score) is a numerical metric for the decoy set quality that is derived from spatial analysis of distances of molecules in a multidimensional physicochemical space. Distances in physicochemical space are calculated from each active to all remaining actives and decoys after data normalization following eq 4

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

With $x_{i,j,norm}$ being the normalized value of physicochemical property j for compound i and $x_{j,95\%}$ and $x_{j,5\%}$ being the 95th and fifth percentile for physicochemical property j, respectively. For each active, all remaining molecules − sorted in an ascending order by distance to the active − are either classified as an active (raising the *true positive rate* = *TPR*) or as a decoy (raising the *false positive rate* = *FPR*). The absolute value of the areas between this curve for active a to the random distribution f($x$) = $x$ is calculated according to eq 5

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

This area becomes zero for an optimal embedding of actives and 0.5 at maximum for a complete spatial separation of actives and decoys. The DOE score is calculated as arithmetic mean of all $ABC^{DOE}$ values according to eq 6 with $m$ being the total number of actives.

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

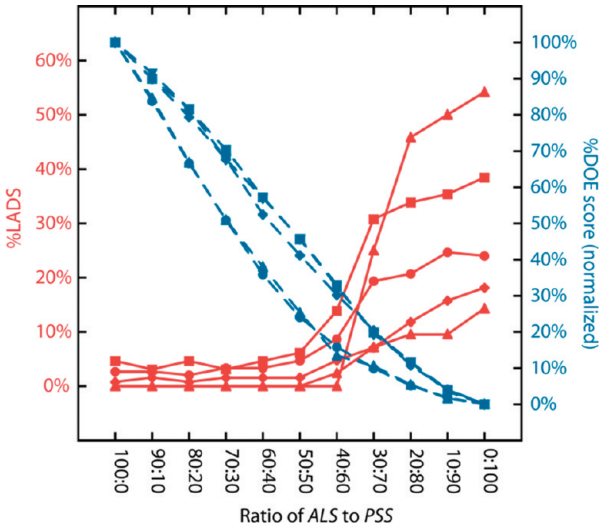



**Figure 2.** Rate of LADS and relative DOE score at different ratios of *avoidance of LADS score* (*ALS*) to *physicochemical similarity score* (*PSS*). %LADS denotes the percentage of LADS that have survived the LADS filter at the given ratio. The normalized DOE score is calculated as $(DOE^{x}-DOE^{min})/(DOE^{max}-DOE^{min})$. Optimal performance for both criteria is always found at 0%. Results are depicted for the following decoy sets: EGFR (●), DHFR (■), COX2 (♦), p38 (▲), FXA (▼).

## RESULTS

**Balancing of Decoy Embedding Quality and Avoidance of LADS.** When improving physicochemical similarity of decoys to actives, the dilemma arises that the danger for decoys to match significant structural features of actives is increased. Putative decoys containing structural elements of bioactive analogs are prone to exhibit bioactivity themselves. Thus, although docking programs may correctly identify this bioactivity, the compound nevertheless is classified as a decoy and will corrupt the statistical assessment of the screening performance. Even for compounds with only residual bioactivity problems occur: screening performance benchmarks use a binary classification system which only discriminates between the active and inactive state. Virtual states in between these do not fit into this classification system and hence confound benchmarking experiments. To estimate the quantity of latent actives in decoy sets constructed with the presented method, we have used a stochastic approach. These results are presented in the Supporting Information and underline the importance of utilizing a LADS filter, because the chance for latent actives is significantly increased when constructing decoy sets using our protocol compared to randomly assembling decoys.

Therefore we have implemented a structure based filter that avoids decoys with structural fingerprints of the active set. However, stringently avoiding structural elements of actives has adverse effects on the effort to mimic the physicochemical properties of the actives. Thus, these two features have to be optimized at the same time. In order to find the optimal ratio, we construct decoy sets for five large active sets with varying weighting schemes for *avoidance of LADS score* (*ALS*) to *physicochemical similarity score* (*PSS*) (Figure 2).

To simulate a hide-and-seek scenario for LADS, only actives originating from a cell which hosts at least one other active are retained. Actives are then divided randomly into an even sized test (LADS) and training set (actives) in a way that each hidden active in the decoy set has at least one corresponding compound in the active set. This is necessary to ensure that LADS are not easily removed due to their dissimilar physicochemical properties.

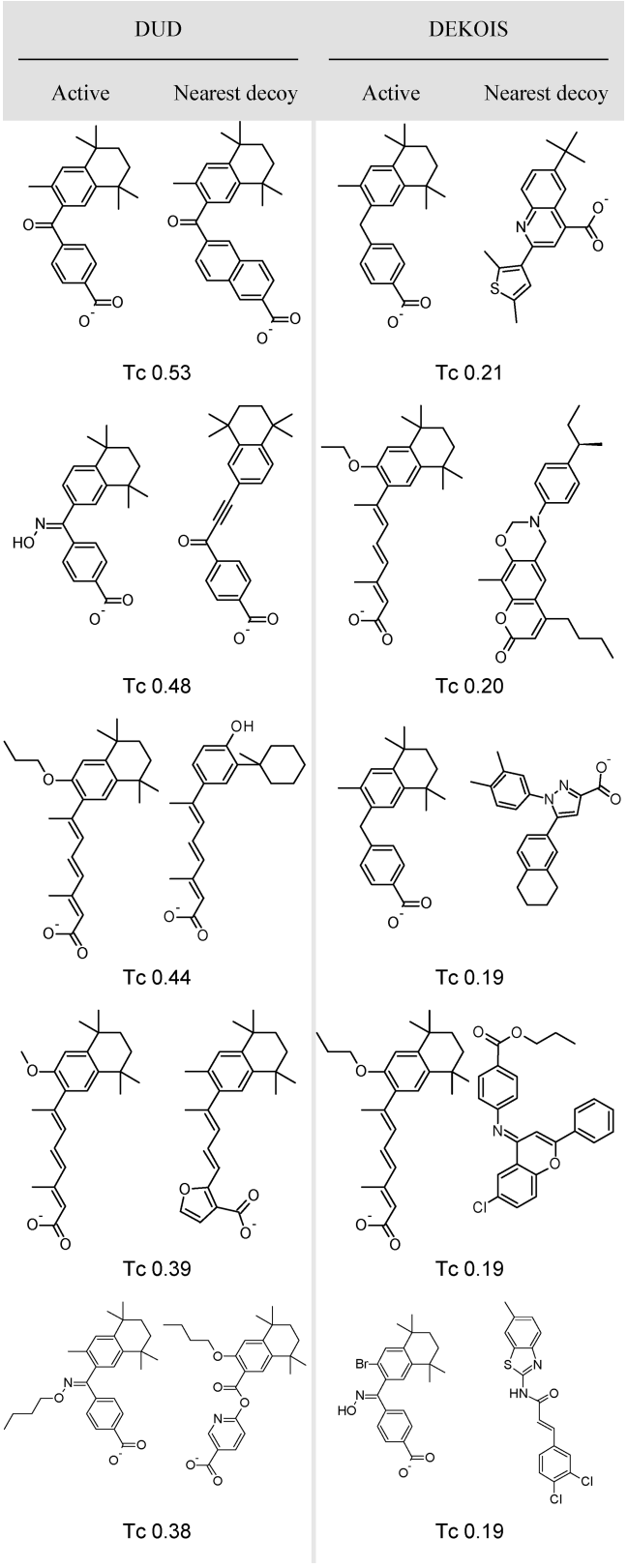
In Figure 2, the weight of the *avoidance of LADS score* is decreasing from left to right, while the weight of the *physicochemical similarity score* is increasing from left to right. At both extremes only one score is represented in the consensus score, while in between weighted mixtures are used to give the resulting consensus score. Interestingly, the rate of LADS never reaches 100%, even if the weight of the *avoidance of LADS score* is set zero. This is due to the fact that LADS can be rejected for the final selection of decoys either because of their nonoptimal matching physicochemical properties or within the scope of diversity optimization. The rate of LADS deviates only marginally from 0 when the *avoidance of LADS score* is contributing equally strong or stronger (50:50 to 100:0) to the consensus score than the PSS. Under this threshold (beginning with 40:60) the rate of LADS increases steeply. It should be noted that for some of the test sets 0% LADS is not achieved for any of the weight settings, indicating that the LADS filter does not necessarily guarantee complete avoidance. This is likely to happen when LADS possess structural elements that are not or only incompletely represented in the active set.

The physicochemical matching of decoys to actives is quantified as normalized DOE score: $(DOE^{x}-DOE^{min})/(DOE^{max}-DOE^{min})$. It should be noted that a low DOE score corresponds to good embedding of actives in the decoy set. A normalized DOE score of 100% refers to the highest DOE score of each decoy set and is generally seen when the physicochemical matching of decoys to actives has not been optimized at all. The normalized DOE score is decreasing continuously with increasing weight of *PSS*. While initially this trend is almost linear, starting from a ratio of 40:60 for several test sets the rate of progress is reduced. At the ratio of 50:50, the LADS filter result is still optimal, while the DOE score has been significantly improved. For a ratio of 40:60, the further gain in DOE reduction comes at the cost of a visible increase of LADS that are not successfully retained by the filter. A reasonable compromise of *avoidance of LADS score* to *PSS* is therefore estimated to be at a ratio 50:50. Using this parameter, we have generated DEKOIS as described in the Methods section for all active sets present in the DUD.

**Application of the Avoidance of LADS Filter.** In order to exemplify the efficacy of the avoidance of LADS filter, we compare the structural similarity between actives from the retinoid X receptor set as it is provided by the DUD with decoys from DUD and DEKOIS (Table 1). Structural similarity is expressed as Tanimoto coefficient calculated based on Scitegic's functional class fingerprints (FCFP_6). First a pairwise comparison of every decoy versus every active is conducted by calculating the Tanimoto coefficients (Tc). The pairs are ranked by their Tc values. Starting from the highest value only the first occurrence of every decoy and every active is kept. The five decoys with the highest structural similarity to any unique active are shown in Table 1. DUD decoys for RXR have generally higher Tc values than DEKOIS, indicating that they are structurally closer to actives than DEKOIS. Strong structural resemblance is even visually detectable for the higher Tc values. It seems that some decoys in the DUD set contain privileged structural elements for molecular recognition in RXR.

We further investigate how close structural similarity may affect docking programs in discriminating between actives and

**Table 1. Comparison of Five Structurally Similar Decoy/Active Pairings in the DUD- and DEKOIS-RXR Set Ranked by Their Tanimoto Coefficient (Tc)**




decoys. We exemplify this issue by docking the DUD and DEKOIS set for RXR using GOLD/Chemscore. Figure 3 shows the three best scoring decoys from DUD and DEKOIS, respectively, in superposition with a closely related active. Similar to the 2D comparison (Table 1), the docked poses reveal a strikingly good superposition of DUD decoys and actives. Moreover, the three shown DUD decoys are even scored slightly higher by GOLD/Chemscore than any of the 20 actives. In contrast, the LADS-filtered DEKOIS show only partial matching of particular substructures, such as a nitrophenyl moiety (in Figure 3E) or a carboxypyrazole moiety (in Figure 3F), which match a carboxyphenyl moiety in the corresponding active. The docking scores for the DEKOIS are substantially lower compared to those of DUD decoys, with the best scored decoy ranking lower than 14 (out of 20) actives.

If bioactivity data were available and DUD decoys were validated nonbinders, this decoy set would have been extremely useful, as it is very challenging for a docking program to distinguish between actives and decoys. In utilizing such a decoy set for benchmarking, one would be able to screen for docking programs that can discriminate between actives and decoys based on subtle differences. However, as we aim to construct decoy sets without the necessity to have bioactivity data available, such a close structural relationship between actives and decoys is undesirable. The risk of introducing LADS and so biasing the benchmarking experiment is unreasonably high.

**DOE Score and Decoy Embedding Quality Heat Map.** Thorough analysis of physicochemical property distributions is required to analyze decoy set quality in terms of physicochemical matching of decoys to actives. Docking experiments can help to estimate the quality of decoy sets. When comparing decoy sets, lower docking enrichments are usually associated with a higher decoy set quality.[24,26] However, the presence of LADS can also lead to this effect and therefore obscure the interpretability of the docking results. Moreover, docking is computationally demanding and thus an expensive way for just characterizing decoy sets.

Given these issues with docking-based quality assessments, there is a great need for quick and reliable methods that help to determine decoy set quality in terms of physicochemical matching in a standardized way. Ideally, this metric summarizes physicochemical decoy set quality into one value to facilitate an easy comparability between different sets.

For an in-depth analysis of decoy sets a metric that identifies problematic clusters, where physicochemical matching of decoys to actives is suboptimal, is desirable. It would be convenient to detect, locate, and thereby help to resolve areas with insufficient decoy embedding.

We have developed the *deviation from optimal embedding score* (*DOE score*), which calculates distances of actives and decoys in a five-dimensional property space — based on MW, logP, number of HBA, number of HBD, and number of RB — and computes how well actives are embedded by decoys. In principle, the DOE plot is a superposition of numerous ROC plots. For each active, the distances to all remaining actives and decoys are calculated. Starting with the nearest neighbor, the *true positive rate* will be raised if the neighbor is an active. The *false positive rate* will be increased if the neighbor is a decoy. Ideally, there are equal numbers of decoys between any two actives. As a result true and false positive rates are equal, indicating a perfect embedding of actives with decoys (equal to the "line of no discrimination"). Clusters of actives without sufficient embedding within decoys will lead to positive or negative enrichment and thus to a "deviation from optimal embedding". For each ROC curve

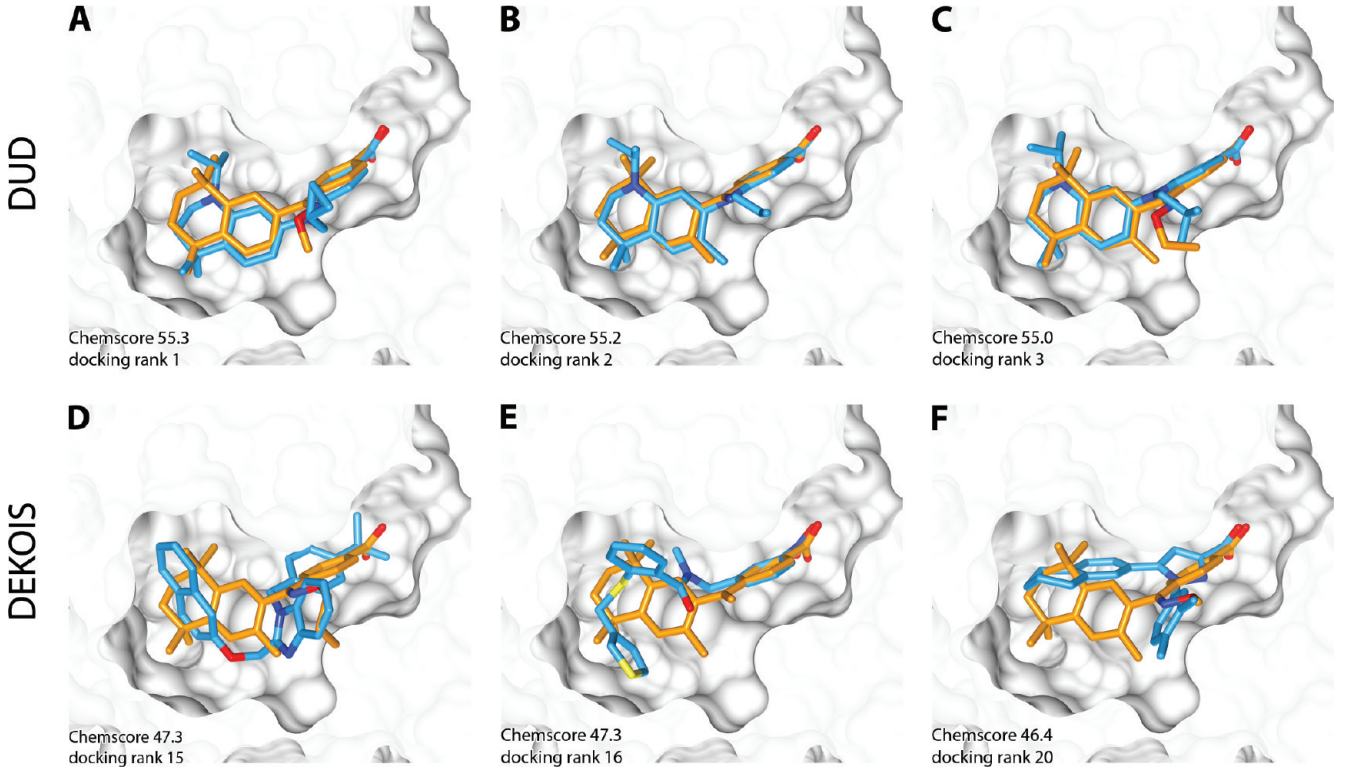



**Figure 3.** Superimposed docking poses of actives from the RXR set (orange) and top-ranked decoys (blue) from DUD (A−C) and DEKOIS (D−F), respectively, in the RXR binding pocket (PDB code 1mvc). Only the highest scoring pose is depicted for each compound. The docking rank is referring to the decoy's position in a sorted list of actives and decoys of the DUD and DEKOIS set, respectively.

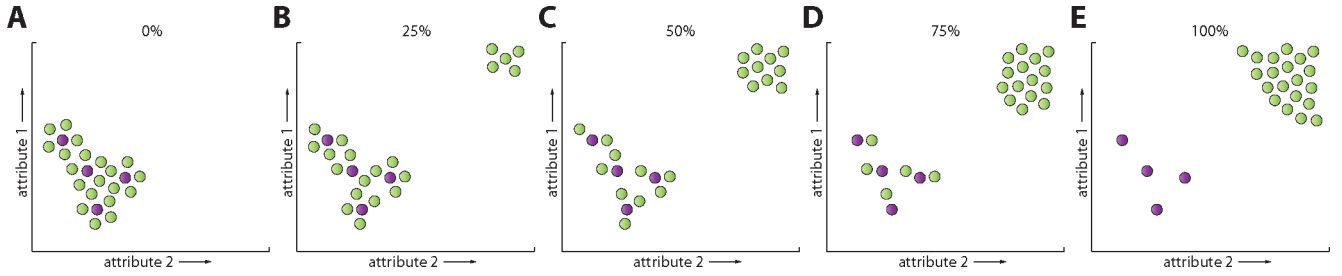



**Figure 4.** Schematic plot showing effects of decoy scrambling on distribution of decoys in a two-dimensional physicochemical property space. For every active five closely matching decoys are assigned initially (A). These property-matched decoys are randomly substituted with 25% (B), 50% (C), 75% (D), and 100% (E) dissimilar decoys. Actives are shown in purple, decoys are shown in green.

(based on one active as a starting point) the sum of all absolute values of the areas between this curve and the "line of no discrimination" is calculated to yield a deviation score for this curve ($ABC^{DOE}$). The DOE score is the arithmetic mean of the deviation scores for all ROC curves. Consequently, a DOE score of zero refers to an ideal embedding, and a DOE score of 0.5 refers to a maximally poor embedding.

The *decoy embedding heat map* visualizes how many decoys there are between any two actives. In using a color code, where red refers to a poor embedding and green to a good embedding, this plot intuitively highlights problematic areas in the decoy set.

**Decoy Scrambling Experiments.** In this series of experiments, we have gradually decreased the quality of decoy sets to study the impact on DOE plots (Figure 5A−E), *decoy embedding heat maps* (Figure 5F−J), and $pROC^{AUC}$ values[39] in docking experiments (Figure 6). The original decoy sets are impaired by substituting high quality decoys with 25%, 50%, 75%, and 100% decoys with completely different physicochemical properties ("dissimilar decoys"). The effect of this decoy set disintegration is shown schematically in a reduced two-dimensional space in Figure 4. In the original decoy set with 0% dissimilar decoys, the active set is well embedded with property matched decoys. Although each active is formally assigned five decoys, the actual number of decoys residing in closer proximity to a given active than the neighbor active is generally higher. By substituting these property-matched decoys with decoys that possess completely different properties, the embedding of actives within decoys is reduced. In the set with 100% dissimilar decoys, the actives are exposed completely.

**Effects on Quality Assessment Metrics.** We have applied these decoy scrambling experiments, as shown schematically in Figure 4, to a decoy set that we have constructed for estrogen receptor antagonists (ER-antagonist). This set contains 39 bioactive compounds from the DUD and 30 property-matched

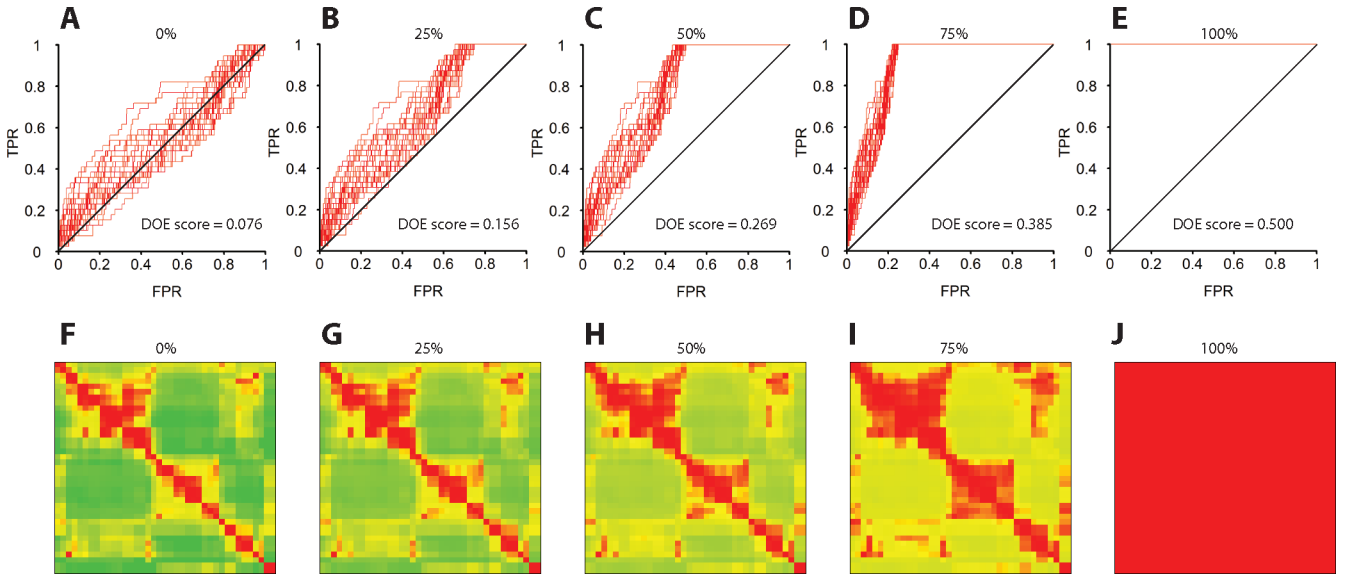



**Figure 5.** DOE plots (A−E) and *decoy embedding heat maps* (F−J) of ER-antagonist decoy sets with a varying amount of artificial impairment. The original decoy sets (A, F) are scrambled with 25% (B, G), 50% (C, H), 75% (D, I) and 100% (E, J) low-quality decoys. The decoys originate from a cell that is diametrically opposed to the actives in terms of physicochemical properties, thus introducing a simulated bias. The diagonal line in DOE plots A−E represents the line of no discrimination. The *decoy embedding heat maps* F−J show a red to yellow color for 0−100 decoys and a yellow to green color for 100−900 decoys between any two actives. Actives are shown horizontally and vertically in the same order. TPR, true positive rate; FPR, false positive rate.

decoys per active selected by our protocol, making a total of 1170 decoys. This original decoy set serves as a starting point for the decoy scrambling experiments and has a rather low DOE score of 0.076 indicating good embedding.

In the resulting DOE plots (Figure 5A−E), the diagonal black line from the left bottom to the top right is the "line of no discrimination" which represents an ideal embedding of actives within decoys. The staircase-shaped red lines represent ROC-like curves for each active. In Figure 5A (the original DEKOIS set) the ROC curves enclose the line of no discrimination with only small positive and negative deviation. With increasing amounts of dissimilar decoys (25%, 50%, 75%, and 100% in Figure 5B−E, respectively), however, the red lines are increasingly shifted away from the black line in a systematic manner: the overall average slope of the ROC curves is increased, reflecting the fact that 100% true positive rate is reached at a decreasingly small false positive rate. This is due to a complete spatial separation of dissimilar decoys from similar decoys and actives (as depicted schematically in Figure 4). The actives are embedded within a decreasingly large set of original (similar) decoys. The increasing fraction of dissimilar decoys is always found as a contiguous block featuring the highest distances from the actives. For the set containing 100% dissimilar decoys (Figure 5E), the DOE score is maximal, meaning that the distance from each active to any active is shorter than to any decoy.

The same effect can be seen in the decoy embedding heat maps. The original DEKOIS set is shown in Figure 5F. Every data point in this decoy embedding heat map represents a pair of two actives (A and B) and is color-coded according to the relative number of decoys that are in closer proximity to the active A than the active B. Given the same order of actives in both dimensions, the data points of the diagonal must be zero, because no decoy can be in closer proximity to an active than the active itself. The decoy embedding heat map is not symmetrical with respect to this diagonal (AB $\neq$ BA), because AB relates to the proximity toward A, while BA relates to the proximity toward B. Hence, the number of decoys counted for AB may differ from BA. Figure 5F−J illustrates that upon accumulation of dissimilar decoys, the red areas in the heat maps increase, indicating a growth of areas of insufficient embedding. Likewise, more and more green areas of good embedding are converted to yellow. From the comparison of the depicted heat maps, it becomes visible that the entire heat map seems to shift homogeneously with no particular zones of stronger decay. The decoy embedding heat map for the 100% dissimilar set turns completely red and illustrates a dramatic decrease in decoy embedding quality, which leads to a complete spatial decomposition of actives and decoys.

**Effects on Screening Performance.** We further investigate the effect of decoy set quality on docking enrichment (Figure 6). Therefore, we have constructed decoy sets for six DUD active sets and artificially impair the decoy embedding quality by decoy scrambling as explained above. The decoy set quality decreases from the dark green ROC curve representing the original set to the curves shown in light green, yellow, orange, and red, which refer to the sets containing 25%, 50%, 75%, and 100% dissimilar decoys, respectively. Comparison of all six sets reveals the general trend that docking enrichment enhances with decreasing decoy set quality. A rather remarkable case is the FGFR1 set: there is no docking enrichment for the original decoy set (dark green line in Figure 6) which means that on average the docking program did not score actives any higher than decoys. Thus, for this target under these particular circumstances, the docking program and scoring function appears to be not recommendable at all for an *in silico* screening project, because it is not able to discriminate between actives and decoys. Interestingly, however, the apparent docking performance is artificially enhanced with decreasing decoy set quality. Judging on decoy sets of lower quality, the user will be misled to assume that the program is suitable for screening projects. This emphasizes the importance of using high quality decoy sets for target-dependent benchmarking.

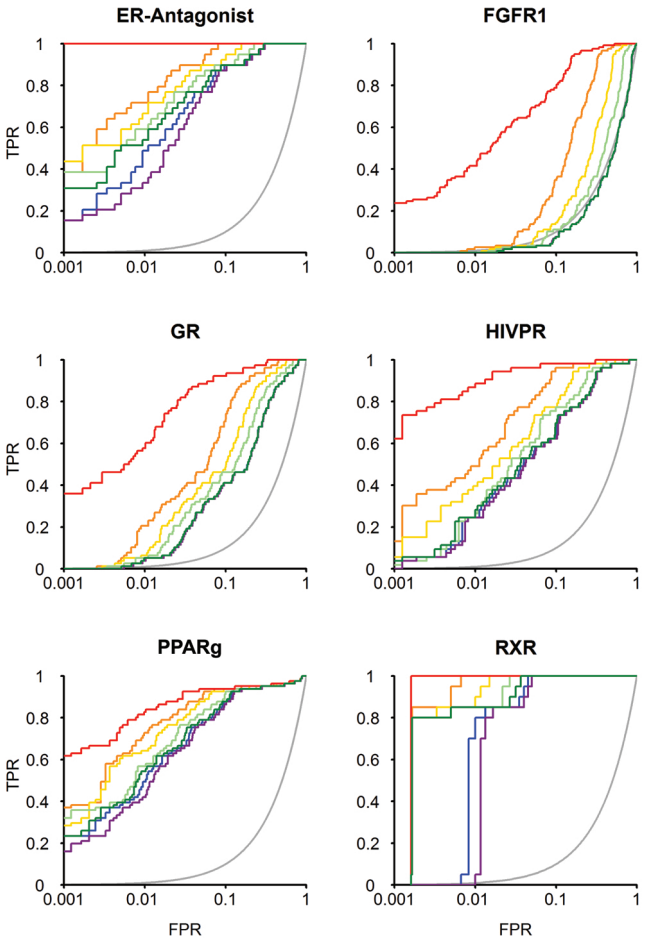



**Figure 6.** pROC plots of GOLD/Chemscore docking experiments with DEKOIS sets for six protein targets. The original good quality set (dark green) is deliberately impaired by replacing 25% (light green), 50% (yellow), 75% (orange), and 100% (red) of the original decoy set with physicochemically dissimilar, low quality decoys. This reduction in decoy embedding quality results in a stepwise improvement of apparent docking enrichment. To illustrate the bias on screening performance induced by LADS (latent actives in the decoy set), some LADS were integrated into the DEKOIS set by substituting decoys with bioactive compounds not yet contained in the active set. The quantity of LADS is 25% (blue) or 50% (purple) the size of the active set. In some cases, the apparent docking enrichment is decreased when LADS are present. The gray line resembles a random docking performance. TPR, true positive rate; FPR, false positive rate.

In accordance to the effects in the decoy embedding heat maps for the ER-antagonist set (Figure 5F–J), in the ROC plots it is obvious that the step from 75% to 100% dissimilar decoys causes a dramatic change in the calculated metrics. While 0% to 75% dissimilar sets follow a continuous trend, the red line representing the 100% dissimilar set clearly stands out. This exceedingly strong change is based upon the loss of the last remaining suitable decoys from the original DEKOIS set.

To further investigate the effect of LADS (latent actives in the decoy set) on docking enrichment, we substituted decoys with reported bioactive compounds that were not included in the initial DUD active sets, thus either increasing the total amount of active compounds by 25% or 50%. These percentages roughly reflect the typical expectancies for the occurrence of LADS according to Table S3 (see the Supporting Information). LADS are classified as decoys but should be scored similarly to the actives. As a consequence, they most probably blend in with the active set at the top of the docking hit list and impair the rank of the classified actives. This is reflected in the ROC plots (Figure 6, blue and purple curves), where the presence of LADS leads to decreased docking enrichments in most cases. The degree of decrease in docking enrichment is more pronounced for ER-antagonist, PPARg, and RXR than for the other sets. The docking enrichment of the FGFR1 set is not decreased when LADS are present. This is well in line with the decoy scrambling results, which indicate that the docking program is not able to discriminate between actives and inactives anyway. Hence, substituting decoys with LADS is not expected to have a negative effect on docking enrichment.

These experiments emphasize the necessity to consider two opposing effects for the construction of decoy sets: (1) poor decoy embedding leads to artificially enhanced docking performance and (2) the presence of LADS leads to artificially impaired docking performance. Not taking care of both effects can yield enrichments that barely represent the real discriminative power of a docking program.

**Influence of DOE Score on pROC.** To elucidate how strongly the decoy embedding quality can influence the docking performance, we have examined the relationship of DOE scores and $pROC^{AUC}$ values in the decoy scrambling experiments (Table 2). Both metrics share an increasing trend with decreasing decoy set quality. As stated before, the values for the 100% dissimilar set represent a fundamental decrease in decoy set quality with most DOE scores approaching the maximum value of 0.5. This indicates an almost complete spatial separation of actives from decoys based on physicochemical properties. Because $pROC^{AUC}$ values benefit most from early recognition, they respond very sensitively to this substantial loss of decoy set quality and increase in a highly disproportionate manner. For the 0%–75% interval, a reasonable correlation ($R^2$ between 0.977 and 0.996) is found between $pROC^{AUC}$ and DOE scores for each single target (Figure 7). The slope of these linear trends is similar for low to moderate docking enrichments but is decreased when $pROC^{AUC}$ values approach the maximal values for each target.

Unsurprisingly, there is no global correlation among different targets as this would mean that docking enrichment is only dependent on the physicochemical properties of the decoy set, which is – according to our data – not true. In Figure 7, ER-antagonist, FGFR1, and HIVPR have similar DOE scores for the original decoy sets (0% dissimilar) ranging from 0.05–0.08, indicating a comparable decoy embedding quality. However, they deviate significantly in $pROC^{AUC}$ values from 0.39 for FGFR1 (pROC value for random enrichment: 0.43), 1.48 for HIVPR, and 2.11 for ER-antagonist. Considering that the quality of these decoy sets is comparable, this finding indicates that the docking performance for the same tool and parameters is strongly target-dependent. This emphasizes the need to assess screening performance of docking programs for the targets of interest with tailor-made decoy sets.

**Comparison of DEKOIS to DUD.** To assess the quality of the DEKOIS as benchmark sets for numerous targets, we compared DEKOIS to the present standard for decoy sets, the Directory of Useful Decoys (DUD).[36] We examined the docking performance with particular focus on the early enrichment and the decoy embedding quality of both sets. Moreover, we investigated the influence of the two scoring functions, Chemscore and Goldscore, on docking enrichment of the benchmark sets. We

Table 2. Decoy Embedding Quality, Measured as DOE Score, and Docking Enrichment, Measured as $pROC^{AUC}$ for ER-Antagonist, FGFR1, GR, HIVPR, PPARg, and RXR Decoy Sets with Stepwise Increased Amounts of Low Quality Decoys

| | ER-antagonist | | FGFR1 | | GR | | HIVPR | | PPARg | | RXR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dissimilar decoys | $pROC^{AUC}$ | DOE score | $pROC^{AUC}$ | DOE score | $pROC^{AUC}$ | DOE score | $pROC^{AUC}$ | DOE score | $pROC^{AUC}$ | DOE score | $pROC^{AUC}$ | DOE score |
| 0% | 2.11 | 0.08 | 0.39 | 0.05 | 0.92 | 0.11 | 1.48 | 0.06 | 2.10 | 0.25 | 2.56 | 0.11 |
| 25% | 2.22 | 0.16 | 0.50 | 0.08 | 1.04 | 0.17 | 1.54 | 0.10 | 2.20 | 0.29 | 2.58 | 0.20 |
| 50% | 2.36 | 0.27 | 0.65 | 0.17 | 1.17 | 0.26 | 1.79 | 0.21 | 2.32 | 0.34 | 2.63 | 0.29 |
| 75% | 2.48 | 0.38 | 0.90 | 0.27 | 1.38 | 0.37 | 2.09 | 0.32 | 2.44 | 0.41 | 2.70 | 0.39 |
| 100% | 3.07 | 0.50 | 1.98 | 0.36 | 2.35 | 0.48 | 2.84 | 0.43 | 2.78 | 0.47 | 2.78 | 0.49 |

will discuss all mutual comparisons of $pROC^{AUC}$ values with respect to the corresponding DOE scores in order to monitor the relationship of these two criteria.

We constructed decoy sets for the 40 DUD targets according to the protocol described in Methods. The DEKOIS active sets were derived from the DUD ligand sets (as discussed in more detail in the Supporting Information).[48] To allow for a better comparability, we docked against the PDB structures specified in the DUD (see Supporting Information Table S1). As a control for the suitability of the used crystal structure and the employed scoring functions, we additionally performed a pose retrieval experiment for each receptor−ligand-complex. All results are shown in Table 3.

We calculated the DOE score for all DEKOIS and DUD sets. Smaller values indicate a better decoy embedding quality. Based on the robustness experiments (Figure S9, Tables S6−S9), we have come to the conclusion that small differences in DOE score and particularly small deviations in $pROC^{AUC}$ values should not be overinterpreted. Thus, we will use a self-imposed safety margin of five and ten percent for differences in DOE score and $pROC^{AUC}$ values, respectively, to consider these to be significant.

For 77.5% of targets the DEKOIS achieved a lower DOE score than the DUD sets, indicating a favorable decoy embedding quality of the DEKOIS. We obtained comparable scores for 12.5% and higher scores for only 10% of the targets. The results for the GOLD/Goldscore benchmark showed a smaller $pROC^{AUC}$ of the DEKOIS for 35%, a comparable $pROC^{AUC}$ for 27.5%, and a higher $pROC^{AUC}$ for 37.5% of all examined targets. The respective GOLD/Chemscore experiments exhibited smaller enrichments of DEKOIS for 37.5%, comparable enrichments for 35%, and higher enrichments for 27.5% of all cases. Consequently, in terms of enrichment performance, the majority of DEKOIS were at least as demanding a docking challenge as the DUD decoy sets for both employed scoring functions.

A closer look at the 40 pROC plots (Figure 8) reveals several interesting characteristics. In general, we observe diverse results for DEKOIS/CS, DEKOIS/GS, DUD/CS, and DUD/GS with moderate to good enrichments. However, a few targets do not exhibit significant differences between the four approaches, despite showing suitable enrichments (ACE, CDK2, COX2, HIVPR, thrombin, and trypsin). Some cases exist where neither for DUD nor DEKOIS any of both scoring functions yields an enrichment performance significantly better than random: EGFR, FGFR1, PDGFRB, and SRC. This either means that (1) the protein structure is not suitable to reflect the true binding situation or the multitude of adaptive binding situations or (2) the scoring function cannot adequately rank the binding pose of bioactive compounds. Various targets show similar pROC curves for DEKOIS and DUD, while the scoring function causes major deviation in the resulting enrichment (e.g., ACHE, FXA, GPB, and INHA). There are also a few examples where the deviation caused by the set exceeds the deviation caused by the scoring function (GART and MR).

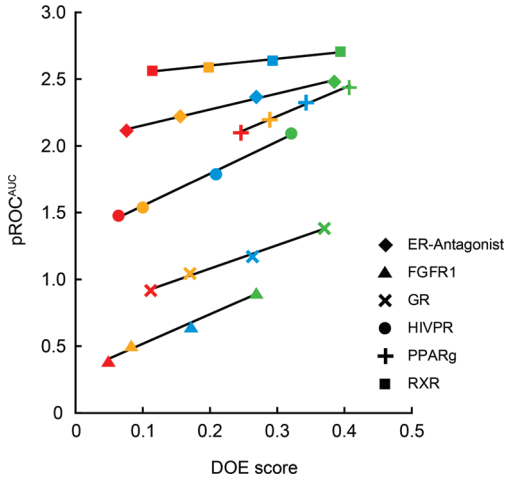



Figure 7. Docking performance − measured as $pROC^{AUC}$ values − in decoy scrambling experiments as a function of decoy embedding − measured as DOE scores. The four data points of each series refer to the sets with 0% (red), 25% (yellow), 50% (blue), and 75% (green) dissimilar decoys. Data points of the 100% dissimilar sets are omitted for all targets, because the loss of decoy embedding leads to a disproportionate response in $pROC^{AUC}$ values.

To elucidate differences between DEKOIS and DUD sets in detail, we plotted delta $pROC^{AUC}$ against delta DOE in Figure 9. For both metrics we subtracted the DUD value from the respective DEKOIS value to yield the final parameter. The high frequency of negative ΔDOE scores indicates a prevalent advantage in decoy embedding quality for the DEKOIS. The $\Delta pROC^{AUC}$ values range from −0.499 to 0.950 revealing no obvious trend toward positive or negative values. We will discuss a selection of representative cases (highlighted spots in yellow) exemplifying characteristics of this plot.

For VEGFR2, the DEKOIS achieved both a lower DOE score and a smaller $pROC^{AUC}$ for the Chemscore docking. The corresponding data point and further examples for this case can be found in the lower left quadrant of Figure 9A. These results are in good agreement with our decoy scrambling experiments, where a lower DOE score correlated with a smaller docking enrichment. These examples emphasize the ambitions to generate benchmarking sets with good decoy embedding quality in order to avoid artificially enhanced screening performance.

In the upper right quadrant of the plot the DUD sets exhibit the preferable characteristics of possessing a smaller DOE and

**Table 3. Comparison of Deviation from Optimal Embedding score (DOE Score) and pROC$^{AUC}$ for Gold- and Chemscore Docking Experiments over 40 Protein Targets**

| | DOE score | | pROC$^{AUC}$ Goldscore | | pROC$^{AUC}$ Chemscore | | rmsd[a] | |
|---|---|---|---|---|---|---|---|---|
| target | DEKOIS | DUD | DEKOIS | DUD | DEKOIS | DUD | Goldscore | Chemscore |
| ACE | 0.04 | 0.16 | 0.97 | 0.84 | 1.08 | 0.98 | 1.1 | 0.8 |
| ACHE | 0.05 | 0.10 | 0.52 | 0.66 | 1.86 | 1.46 | 0.5 | 1.0 |
| ADA | 0.12 | 0.20 | 0.49 | 0.61 | 0.29 | 0.36 | 0.6 | 1.3 |
| ALR2 | 0.04 | 0.17 | 0.80 | 1.16 | 0.47 | 0.82 | 0.9 | 0.7 |
| AMPC | 0.06 | 0.13 | 0.34 | 0.25 | 1.35 | 0.62 | 3.2[d] | 2.4[c] |
| AR | 0.06 | 0.11 | 0.56 | 0.56 | 1.12 | 1.10 | 0.2 | 0.4 |
| CDK2 | 0.07 | 0.18 | 0.83 | 1.02 | 0.88 | 0.76 | 1.6 | 1.8 |
| COMT | 0.08 | 0.20 | 0.74 | 0.96 | 1.33 | 1.03 | 3.1[b] | 3.0[b] |
| COX1 | 0.05 | 0.13 | 0.51 | 0.32 | 0.63 | 0.58 | 0.8 | 1.6 |
| COX2 | 0.02 | 0.13 | 1.21 | 1.41 | 1.41 | 1.33 | 1.5 | 1.2 |
| DHFR | 0.28 | 0.23 | 1.53 | 1.07 | 0.95 | 0.79 | 1.2 | 1.1 |
| EGFR | 0.02 | 0.13 | 0.36 | 0.38 | 0.48 | 0.53 | 5.1[b] | 5.0[b] |
| ER-agonist | 0.09 | 0.14 | 0.97 | 1.15 | 1.70 | 1.86 | 0.5 | 0.6 |
| ER-antagonist | 0.07 | 0.10 | 0.97 | 1.03 | 2.11 | 1.49 | 1.5 | 1.5 |
| FGFR1 | 0.05 | 0.16 | 0.27 | 0.27 | 0.39 | 0.39 | 2.9[b] | 1.2 |
| FXA | 0.28 | 0.26 | 0.42 | 0.44 | 0.90 | 0.87 | 0.6 | 1.7 |
| GART | 0.41 | 0.42 | 1.73 | 1.11 | 1.63 | 1.02 | 2.5[b] | 2.6[d] |
| GPB | 0.21 | 0.20 | 0.49 | 0.58 | 1.43 | 1.64 | 0.3 | 0.2 |
| GR | 0.10 | 0.16 | 0.33 | 0.42 | 0.92 | 1.08 | 0.4 | 0.4 |
| HIVPR | 0.06 | 0.16 | 1.24 | 1.33 | 1.48 | 1.16 | 2.1[d] | 1.3 |
| HIVRT | 0.03 | 0.16 | 0.38 | 0.60 | 0.42 | 0.88 | 1.0 | 1.2 |
| HMGR | 0.06 | 0.14 | 0.37 | 0.68 | 1.24 | 1.52 | 0.5 | 1.5 |
| HSP90 | 0.05 | 0.09 | 1.41 | 1.12 | 0.51 | 0.75 | 0.6 | 0.7 |
| INHA | 0.02 | 0.22 | 0.45 | 0.31 | 1.31 | 1.40 | 1.2 | 1.1 |
| MR | 0.06 | 0.17 | 1.12 | 1.62 | 1.66 | 1.31 | 0.3 | 0.4 |
| NA | 0.14 | 0.11 | 1.52 | 1.20 | 1.41 | 1.48 | 0.7 | 0.5 |
| p38 | 0.03 | 0.14 | 0.43 | 0.52 | 0.77 | 0.79 | 0.8 | 0.9 |
| PARP | 0.06 | 0.11 | 0.36 | 0.36 | 1.59 | 1.16 | 0.8 | 1.0 |
| PDE5 | 0.03 | 0.21 | 0.44 | 0.46 | 0.98 | 0.90 | 3.5[c] | 1.8 |
| PDGFRB | 0.02 | 0.22 | 0.37 | 0.28 | 0.69 | 0.46 | 0.5 | 0.6 |
| PNP | 0.16 | 0.11 | 1.18 | 1.02 | 0.76 | 1.08 | 0.3 | 0.3 |
| PPARg | 0.21 | 0.27 | 1.22 | 0.87 | 2.10 | 1.15 | 2.6[b] | 1.5 |
| PR | 0.08 | 0.18 | 0.38 | 0.42 | 0.71 | 1.09 | 0.7 | 0.7 |
| RXR | 0.09 | 0.26 | 1.77 | 1.57 | 2.57 | 1.88 | 0.3 | 0.5 |
| SAHH | 0.33 | 0.32 | 0.95 | 0.78 | 0.64 | 0.76 | 0.6 | 0.5 |
| SRC | 0.04 | 0.18 | 0.16 | 0.10 | 0.31 | 0.29 | 1.9 | 1.6 |
| thrombin | 0.19 | 0.19 | 0.90 | 0.98 | 1.14 | 0.88 | 1.1 | 6.2[d] |
| TK | 0.08 | 0.10 | 0.72 | 0.50 | 0.98 | 0.83 | 0.3 | 0.7 |
| trypsin | 0.22 | 0.16 | 1.03 | 0.99 | 1.16 | 1.10 | 0.6 | 2.3[c] |
| VEGFR2 | 0.03 | 0.20 | 0.42 | 0.40 | 1.01 | 1.42 | 1.2 | 0.7 |

[a] The quality of pose retrieval in self-docking experiments for the crystal structure ligand is estimated using the in-place root-mean-square deviation (rmsd) of heavy atoms. rmsd relates to the best scoring pose unless specified individually. [b] rmsd ≤2 Å found within rank 2−10. [c] rmsd ≤2 Å found within rank 11−20. [d] No rmsd ≤2 Å found within rank 1−20.

pROC$^{AUC}$ value. The DHFR set represents therefore another case in which the better physicochemical matching between actives and decoys, quantified by the DOE score, resulted in a more demanding benchmark.

FXA represents a case in which both pROC$^{AUC}$ and DOE values differ only slightly between DEKOIS and DUD. The comparable decoy embedding quality for both sets, which implies a comparable level of screening challenge, results in very similar enrichment values. This suggests that both DEKOIS and DUD are equally well suited for this target.

Whenever the employed docking method could not adequately separate actives from decoys, as seen for the random-like enrichment with Goldscore (Figure 8, FXA), potential differences in decoy set quality are unlikely to be detected. Here, only the Chemscore docking approach shows a reasonable enrichment of actives and is therefore suited for comparing the screening

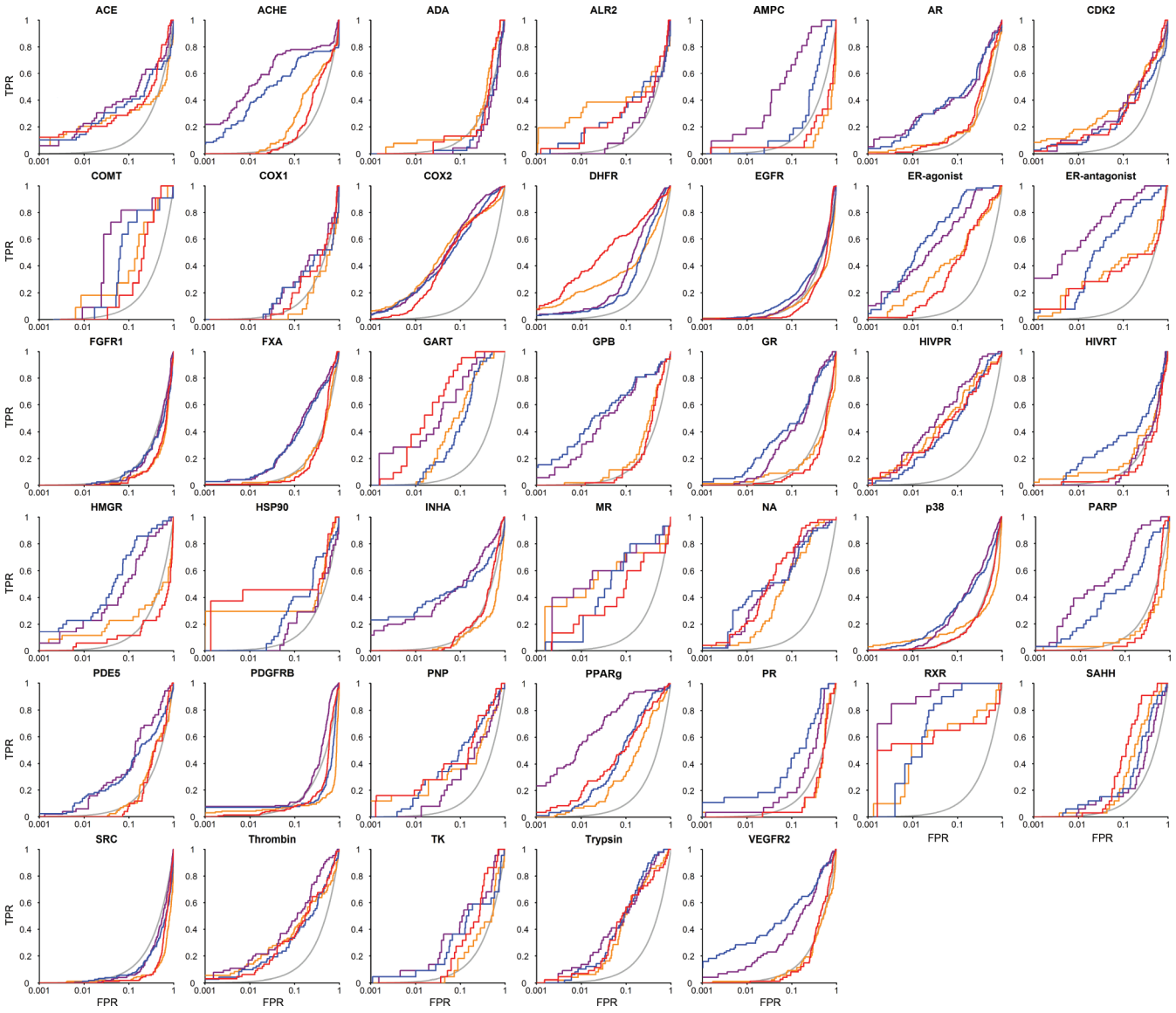



**Figure 8.** pROC plots of docking experiments comparing the screening performance of GOLD for two scoring functions against 40 protein targets using DEKOIS and DUD sets. Goldscore results for DEKOIS and DUD are shown in red and orange, and Chemscore results are shown in purple and blue, respectively. The true positive rate (TPR) is the fraction of recovered actives; the false positive rate (FPR) is the fraction of recovered decoys from a score-ordered list of all decoys. The gray line corresponds to a random screening performance.

performance between the two sets. As Figure 9A shows only differences in $pROC^{AUC}$ values without any information regarding the significance of the enrichments, putative issues with insignificant enrichment are resolved by fading the corresponding data points to gray.

The upper left quadrant in Figure 9A contains results for which DEKOIS obtain higher $pROC^{AUC}$ values despite lower DOE scores, thus surprisingly opposing the expected trend. A prominent example for this case is RXR. We found a possible explanation for this effect by identifying several compounds in the DUD decoy set (Table 1), which display substantial structural similarity to compounds in the active set. In a separate experiment (Figure 6) we successfully demonstrated that deliberate inclusion of LADS in our decoy set decreased enrichment performance for RXR significantly. Thus, we conclude that the presence of LADS constitutes a real hindrance for generating challenging but fair decoy sets, because LADS can artificially impair screening performance. However, it may depend on the applied parameters for molecular recognition of the docking tool, whether LADS exert a strong confounding effect ("high scoring LADS") or a negligible influence ("low scoring LADS") on the enrichment.

Being aware that structural similarity can be an indicator but is surely not a sound proof for comparable bioactivity, we have investigated the occurrence of decoys with high structural similarity to actives in all of the DEKOIS and DUD sets (Supporting Information Table S5). We have generated FCFP_6 fingerprints for all compounds and have compared their similarity by calculating the Tanimoto coefficient between each decoy and active. The highest Tc for mutually exclusive pairs of actives and decoys are used to form an arithmetic mean. This "doppelganger score" provides information about the extent of structural similarity between actives and their most structurally related decoys. Profound differences in this doppelganger score between DEKOIS and DUD might highlight examples where a higher


dx.doi.org/10.1021/ci2001549 |J. Chem. Inf. Model. 2011, 51, 2650–2665

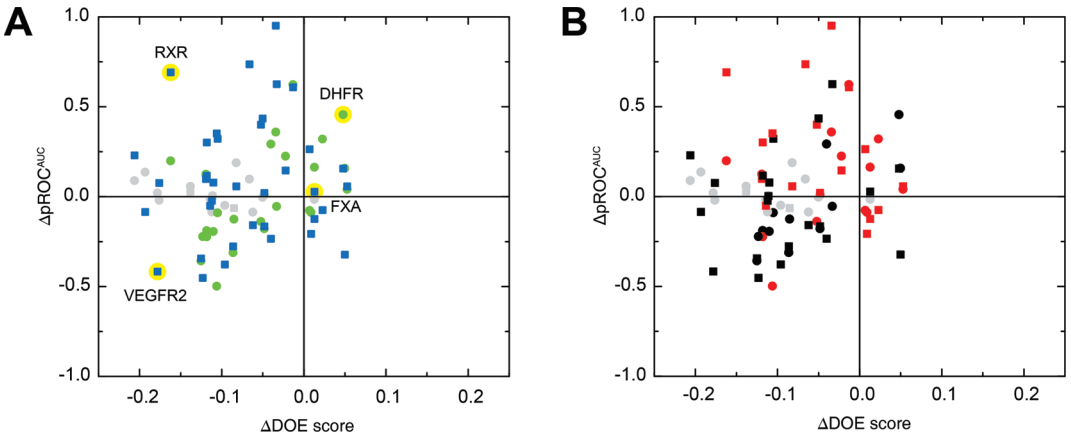



**Figure 9.** The $\Delta pROC^{AUC}$ = $pROC^{AUC}$(DEKOIS) − $pROC^{AUC}$(DUD) is plotted against ΔDOE score = DOE(DEKOIS) − DOE(DUD) for 40 protein targets. Chemscore and Goldscore results are shown as squares and circles, respectively. Gray data points indicate results without sufficient enrichment for DEKOIS and DUD sets, judged by the mean value of both sets. Sufficient enrichment is defined as random performance plus ten percent safety margin: $pROC^{AUC} \geq 0.434 + 0.0434$. (A) Chemscore and Goldscore results are shown in blue and green, respectively. Results for DHFR, FXA, RXR, and VEGFR2 are highlighted in yellow. (B) Black data points refer to a low (<0.1), and red data points refer to a high Δdoppelganger score (≥0.1), calculated as doppelganger score(DUD) − doppelganger score(DEKOIS).

degree of structural similarity can corrupt the enrichment (Table 4). In Figure 9B, we illustrate these cases by coloring data points with high differences in doppelganger score in red (threshold: Δdoppelganger score >0.1). Apparently, quite a few data points in the upper left quadrant (DEKOIS showing favorable embedding but higher screening performance) seem to relate to DUD sets with higher doppelganger scores than the respective DEKOIS. Therefore, an influence of these higher structural similarities on the impairment of the DUD screening performance cannot be ruled out.

Interestingly, when comparing the performances of Goldscore and Chemscore over 40 protein targets, we observe some notable effects. Using the self-imposed safety margin of 10% to consider the docking enrichments to be significantly different, we found that the GOLD/Chemscore combination has yielded higher enrichments for 67.5% of the targets than the GOLD/Goldscore combination. For 17.5% of the targets the two combinations have achieved comparable and for 15% of the targets the GOLD/Chemscore combination has obtained lower enrichments than GOLD/Goldscore. These results indicate that Chemscore performs superior to Goldscore for the majority of targets. This finding could be related to the empirical character of Chemscore, which was parametrized based on 82 ligand−receptor complexes containing structures of DHFR, HIVPR, NA, PNP, thrombin, and trypsin among others.[47,49] Based on the performance on these six particular targets, obviously this parametrization is not the reason for the superior performance of Chemscore. It should be noted that Chemscore was trained to predict ligand binding affinities, and Goldscore was designed to reproduce near-native complex geometries. This might partially explain the differences we have observed. Despite this trend, we generally advocate that the selection of docking tools and scoring functions should always be based on individual benchmarks against particular targets.

## DISCUSSION

In this work, we present a fast, versatile, and robust method to construct *demanding evaluation kits for objective in silico screening* (DEKOIS). This automated process enables a user to construct tailor-made decoy sets for any given sets of bioactive molecules. Each active is accompanied by an individual set of physicochemically similar yet structurally diverse decoys in order to yield a complete physicochemical replica of the active set. The decoys are selected from a precalculated compound database, where each molecule is classified according to its physicochemical properties. This makes the selection process computationally less demanding and enables us to use a reasonably large database (15 million compounds).

Within the process of the decoy selection procedure, the decoy set is not only optimized regarding the physicochemical similarity to the active set but also to avoid latent actives in the decoy set (LADS) as effectively as possible. It is important to fine-tune both (a) the physicochemical similarity and (b) the avoidance of LADS, because both effects have opposed confounding effects on docking enrichment and obscure "real screening performance" as we have shown in the decoy scrambling experiments. We have demonstrated that LADS can artificially impair screening performance and suboptimal decoy embedding can artificially enhance screening performance. Although error compensation between these two effects is possible, the responsiveness of various docking programs and scoring functions toward them should be quite different. Especially the molecular recognition of LADS can strongly depend on the type and parameters of different scoring functions. Error compensation can only occur unsystematically, thus, impeding the comparability of alternative methods.

Among the reported strategies to generate decoy sets, essential differences can be found. We deem it unquestionable that an ideal set of decoys would consist of experimentally validated true actives and true inactives. The MUV data sets[27] utilize a database of compounds with bioactivity status derived from PubChem experimental data. While the availability of such experimental data is desirable, high-quality verification (or falsification) of bioactivity cannot be provided for a multitude of interesting, novel targets and thus is restrictive in terms of a flexible and consistent decoy set construction. Due to the lack of this data, we have employed a strategy for avoiding LADS by deriving the bioactivity status from structural information of known binders.

**Table 4. Arithmetic Mean of the Highest Tanimoto Coefficients for Mutually Exclusive Pairs of Actives and Decoys ("Doppelganger Score") in DEKOIS and DUD Sets**

| target | doppelganger score DEKOIS | doppelganger score DUD | Δdoppelganger score |
|---|---|---|---|
| ACE | 0.23 | 0.41 | 0.19 |
| ACHE | 0.27 | 0.41 | 0.14 |
| ADA | 0.24 | 0.29 | 0.04 |
| ALR2 | 0.23 | 0.32 | 0.08 |
| AMPC | 0.26 | 0.46 | 0.20 |
| AR | 0.20 | 0.38 | 0.18 |
| CDK2 | 0.24 | 0.28 | 0.04 |
| COMT | 0.19 | 0.32 | 0.14 |
| COX1 | 0.25 | 0.36 | 0.11 |
| COX2 | 0.28 | 0.32 | 0.05 |
| DHFR | 0.26 | 0.29 | 0.03 |
| EGFR | 0.25 | 0.30 | 0.05 |
| ER-agonist | 0.23 | 0.32 | 0.08 |
| ER-antagonist | 0.21 | 0.39 | 0.18 |
| FGFR1 | 0.24 | 0.29 | 0.05 |
| FXA | 0.25 | 0.32 | 0.07 |
| GART | 0.24 | 0.50 | 0.26 |
| GPB | 0.25 | 0.43 | 0.18 |
| GR | 0.24 | 0.27 | 0.03 |
| HIVPR | 0.23 | 0.29 | 0.06 |
| HIVRT | 0.25 | 0.30 | 0.05 |
| HMGA | 0.17 | 0.20 | 0.03 |
| HSP90 | 0.21 | 0.28 | 0.07 |
| INHA | 0.25 | 0.29 | 0.04 |
| MR | 0.16 | 0.49 | 0.33 |
| NA | 0.23 | 0.40 | 0.17 |
| p38 | 0.26 | 0.30 | 0.04 |
| PARP | 0.21 | 0.27 | 0.06 |
| PDE5 | 0.24 | 0.27 | 0.03 |
| PDGFRB | 0.25 | 0.31 | 0.06 |
| PNP | 0.22 | 0.26 | 0.04 |
| PPARg | 0.26 | 0.39 | 0.13 |
| PR | 0.22 | 0.27 | 0.05 |
| RXR | 0.17 | 0.37 | 0.20 |
| SAHH | 0.25 | 0.52 | 0.27 |
| SRC | 0.25 | 0.29 | 0.04 |
| thrombin | 0.23 | 0.35 | 0.12 |
| TK | 0.22 | 0.64 | 0.43 |
| trypsin | 0.22 | 0.33 | 0.11 |
| VEGFR2 | 0.24 | 0.28 | 0.04 |

The DUD aims to achieve this by using a fingerprint-based Tanimoto coefficient cutoff of 0.9 for all decoys to any active in the DUD.[36]

One present shortcoming of our method may be the restriction to only five physicochemical properties due to technical reasons. Other studies suggest that the inclusion of further molecular descriptors for this process, like the number of aromatic rings or charge, might increase the quality of generated benchmark sets.[50]

We suggest that the easy availability of tailor-made decoy sets to evaluate and benchmark *in silico* screening methods can contribute to advance the field of structure-based drug discovery in medicinal chemistry. As we have shown in Figure 7, screening performance can be strongly target-dependent. Furthermore, real screening performance can be concealed by the presence of LADS and poor decoy embedding. Our method to individually generate tailor-made, high quality decoy sets helps to benchmark screening tools for a particular target in order to save time and resources.

We describe two complementary ways for efficiently assessing decoy set quality. DOE plots help to visualize positive and negative enrichment solely based on physicochemical dissimilarity. Moreover, they facilitate the quantification of decoy set quality by transforming the DOE curves into a score within a precisely defined range (0−0.5) and meaningful borders (optimal embedding vs no embedding). In addition, we define decoy embedding heat maps based on the number of decoys in close proximity to actives. One significant advantage of these heat maps is the convenient identification of issues with clusters of insufficiently embedded actives allowing for an in-depth analysis of the problem.

Although we utilize the DUD targets to validate our method and compare screening performance and decoy embedding of our DEKOIS to the DUD sets as the present gold-standard, our work does not aim at reinventing the DUD. We rather strive for extending and complementing the collection of publicly available high quality decoy sets toward new target space. To achieve this goal, the method presented herein is well suited as shown in a variety of control experiments. We have established a strategy for decoy embedding that deviates from the conventional distribution based property-matching. By embedding each single active within a cloud of individual decoys, we can ensure a locally optimized embedding in contrast to a globally optimized in the DUD (Figure 10). Based on this approach, we meet or exceed the standards of physicochemical property matching constituted in the DUD. We have demonstrated the importance of actively avoiding the inclusion of latent actives in the decoy set (LADS). Hence, a strong emphasis is put on co-optimization of avoidance of LADS and optimal decoy embedding. However it should be noted that accepting or ignoring the possibility of LADS would produce a higher embedding quality (Figure 2) and pretend a more rigorous reduction of the enrichment, possibly even beyond the "real screening performance" (Figure 6). We conclude that the putative gain in these parameters does not justify accepting the risk of introducing LADS.

However, as not everybody might perceive the risk of introducing LADS in a similar way, the flexibility of our protocol allows us to provide "most rigorous" decoy sets (generated without LADS filter, but with maximal physicochemical matching) upon request.

All present and future DEKOIS data sets will be made accessible through our Web repository at http://www.dekois.com.

## ■ ASSOCIATED CONTENT

**Supporting Information.** A schematic depiction of the binning procedure, tables with docking details and information regarding decoy sets, a protocol for the correction of DUD ligand sets, preparation of the DUD active sets for the construction of DEKOIS, a stochastic approach toward quantification of the number of latent actives in decoy sets, an alternative method to the doppelganger score for describing the most significant structural similarities between actives and decoys, 2D depictions

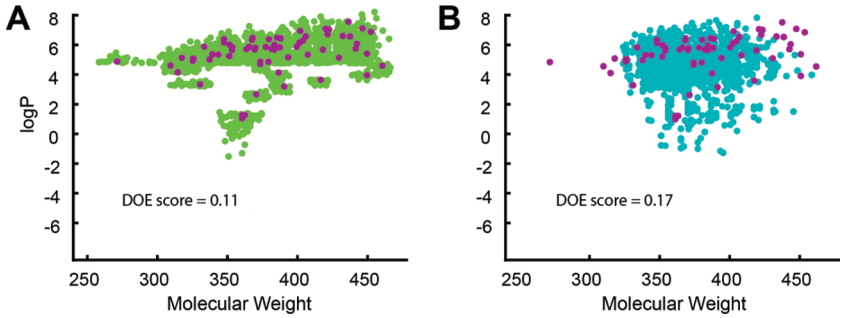



**Figure 10.** Representation of the decoy embedding characteristics in the DEKOIS (A) and DUD set (B), resulting from different decoy selection strategies. Distribution of logP and molecular weight in the GR set is shown as an example in a reduced physicochemical property space.

of DUD decoys that are structurally closely related to DUD actives, effects of decoy scrambling experiments on the physicochemical property distributions, DOE plots and decoy embedding heat maps of the dissimilar decoy sets, an alternative method for conducting decoy scrambling experiments, DOE plots and decoy embedding heat maps of decoy sets for this alternative method to the decoy scrambling experiments, figures and tables with detailed information regarding the robustness experiments, and statistical analysis of the docking enrichment and DOE scores over all 40 targets. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*Phone: +49-7071-2974567. Fax: +49-7071-295637. E-mail: frank.boeckler@uni-tuebingen.de.

### Author Contributions

‡S.M.V. and M.R.B. contributed equally to this work.

## ACKNOWLEDGMENT

We thank Johannes Beifuss for his assistance with docking experiments and Lukas Kusch for his assistance in conducting an alternative method for the decoy scrambling experiments and the Directory of Useful Decoys (DUD) for providing their data without restrictions. We also thank Axel Schüssler, Rainer Wilcken, and Markus Zimmermann for proofreading the manuscript.

## ABBREVIATIONS:

ACE, angiotensin-converting enzyme; ACHE, acetylcholinesterase; ADA, adenosine deaminase; ALR2, aldose reductase; AMPC, AmpC $\beta$-lactamase; AR, androgen receptor; AUC, area under the curve; CDK2, cyclin-dependent kinase 2; COMT, catechol O-methyltransferase; COX1, cyclooxygenase-1; COX2, cyclooxygenase-2; DEKOIS, demanding evaluation kits for *in silico* screening; DHFR, dihydrofolate reductase; DOE, deviation from optimal embedding; DUD, directory of useful decoys; EGFR, epidermal growth factor receptor; ER, estrogen receptor; FGFR1, fibroblast growth factor receptor 1; FPR, false positive rate; FXA, factor Xa; GART, glycinamide ribonucleotide transformylase; GPB, glycogen phosphorylase $\beta$; GR, glucocorticoid receptor; HBA, H-bond acceptor; HBD, H-bond donor; HIVPR, HIV protease; HIVRT, HIV reverse transcriptase; HMGR, hydroxymethylglutaryl-CoA reductase; HSP90, human heat shock protein 90; INHA, enoyl ACP reductase; LADS, latent actives in the decoy set; MR, mineralocorticoid receptor; MW, molecular weight; NA, neuraminidase; p38, P38 mitogen activated protein kinase $\alpha$; PARP, poly(ADP-ribose) polymerase; PDE5, phosphodiesterase 5A; PDGFRB, Beta-type platelet derived growth factor receptor kinase; PNP, purine nucleoside phosphorylase; PPARg, peroxisome proliferator activated receptor $\gamma$; PR, progesterone receptor; RB, rotatable bonds; ROC, receiver operating characteristic; RXR, retinoic X receptor $\alpha$; SAHH, S-adenosyl-homocysteine hydrolase; SRC, tyrosine kinase SRC; Tc, Tanimoto coefficient; TK, thymidine kinase; TPR, true positive rate; VEGFR2, vascular endothelial growth factor receptor 2

## REFERENCES

(1) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273–276.

(2) Clark, D. E. What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discovery* **2008**, *3*, 841–851.

(3) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580–594.

(4) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(5) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 2644–2676.

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

(7) Boeckler, F. M.; Joerger, A. C.; Jaggi, G.; Rutherford, T. J.; Veprintsev, D. B.; Fersht, A. R. Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 10360–10365.

(8) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. *J. Med. Chem.* **2010**, *53*, 8461–8467.

(9) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727–730.

(10) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.

(11) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(12) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7–S26.

(13) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing Scoring Functions for Protein−Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.

(14) Jain, A. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201–212.

(15) Jain, A.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.
(16) Cleves, A.; Jain, A. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 147–159.
(17) Good, A.; Oprea, T. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.
(18) Hawkins, P.; Warren, G.; Skillman, A.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.
(19) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.
(20) Liebeschuetz, J. Evaluating docking programs: keeping the playing field level. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 229–238.
(21) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962–976.
(22) McInnes, C. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494–502.
(23) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
(24) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.
(25) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D. Consideration of Molecular Weight during Compound Selection in Virtual Target-Based Database Screening. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 267–272.
(26) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
(27) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169–184.
(28) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.
(29) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein−Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2005**, *49*, 5856–5868.
(30) McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *J. Med. Chem.* **2003**, *46*, 2895–2907.
(31) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
(32) 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*, 704–718.
(33) Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical Comparison of Virtual Screening Methods against the MUV Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 2168–2178.
(34) Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H. An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **2010**, *38*, D255–D266.
(35) Wallach, I.; Lilien, R. Virtual Decoy Sets for Molecular Docking Benchmarks. *J. Chem. Inf. Model.* **2011**, *51*, 196–202.
(36) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(37) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 529–536.
(38) Good, A. C.; Hermsmeier, M. A. Measuring CAMD Technique Performance. 2. How "Druglike" Are Drugs? Implications of Random Test Set Selection Exemplified Using Druglikeness Classification Models. *J. Chem. Inf. Model.* **2006**, *47*, 110–114.
(39) Clark, R.; Webster-Clark, D. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141–146.
(40) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2004**, *45*, 177–182.
(41) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
(42) Molecular Operating Environment (MOE), version 2009.10; Chemical Computing Group Inc.: Montreal, Canada, 2009.
(43) DUD:Errata - DISI. http://wiki.bkslab.org/index.php/DUD:Errata (accessed Dec 01, 2011).
(44) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(45) *GOLD*, version 3.2; The Cambridge Crystallographic Data Centre (CCDC): Cambridge, UK, 2008.
(46) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
(47) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
(48) ten Brink, T.; Exner, T. E. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein−Ligand Docking Results. *J. Chem. Inf. Model.* **2009**, *49*, 1535–1546.
(49) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503–519.
(50) Irwin, J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.