JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

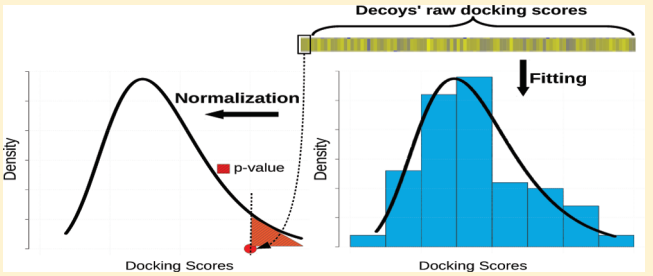
# Normalizing Molecular Docking Rankings using Virtually Generated Decoys

Izhar Wallach,[*,†,‡] Navdeep Jaitly,[†] Kong Nguyen,[§] Matthieu Schapira,[§,∥] and Ryan Lilien[*,†,‡,⊥]

[†]Department of Computer Science, [‡]Donnelly Centre for Cellular and Biomolecular Research, [§]Structural Genomics Consortium, [∥]Department of Pharmacology and Toxicology, and [⊥]Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada M5S 3G4

Supporting Information

**ABSTRACT:** Drug discovery research often relies on the use of virtual screening via molecular docking to identify active hits in compound libraries. An area for improvement among many state-of-the-art docking methods is the accuracy of the scoring functions used to differentiate active from nonactive ligands. Many contemporary scoring functions are influenced by the physical properties of the docked molecule. This bias can cause molecules with certain physical properties to incorrectly score better than others. Since variation in physical properties is inevitable in large screening libraries, it is desirable to account for this bias. In this paper, we present a method of normalizing docking scores using virtually generated decoy sets with matched physical properties. First, our method generates a set of property-matched decoys for every molecule in the screening library. Each library molecule and its decoy set are docked using a state-of-the-art method, producing a set of raw docking scores. Next, the raw docking score of each library molecule is normalized against the scores of its decoys. The normalized score represents the probability that the raw docking score was drawn from the background distribution of nonactive property-matched decoys. Assuming that the distribution of scores of active molecules differs from the nonactive score distribution, we expect that the score of an active compound will have a low probability of having been drawn from the nonactive score distribution. In addition to the use of decoys in normalizing docking scores, we suggest that decoy sets may be a useful tool to evaluate, improve, or develop scoring functions. We show that by analyzing docking scores of library molecules with respect to the docking scores of their virtually generated property-matched decoys, one can gain insight into the advantages, limitations, and reliability of scoring functions.




## INTRODUCTION

Docking-based virtual screening of compound libraries is a method applied in drug discovery to identify active molecules. Using this technique, a large library of candidate molecules can be screened against a target receptor prior to wet-laboratory experimentation. The accurate prediction of binding affinities continues to be a challenge as the scoring functions utilized by currently available docking algorithms only approximate the underlying physical forces of molecular interaction.[1] When the goal of virtual screening is only to identify candidate binders, the absolute precision of a scoring function becomes less critical. In this situation, the prediction of absolute binding affinities is only important insofar as it can be utilized to generate an enriched set of candidate molecules. It is therefore common to evaluate docking algorithms not by their ability to produce accurate absolute binding energies, but by their ability to enrich the fraction of binders in a set of candidate ligands. In most cases, the *enrichment* is computed as the improvement in fraction of active compounds between the starting library and an output set. The frequency of active ligands present within the top $k$% of ranked molecules can be compared across different virtual screening techniques. The higher the frequency of actives among top-ranked compounds, the more successful the docking method. For practical reasons one is typically interested in frequencies when $k$ is small because only a small fraction of the library will eventually be followed up by wet-laboratory validations.

Docking methods are often evaluated on their ability to rank compounds in a manually crafted benchmark data set. These sets often consist of a small number of known binders and for each binder a set of decoys (or nonbinders). To obtain a fair assessment of compound enrichment, it is important to pay close attention to the physical properties of the ligands in the decoy set. To maintain consistency with previous literature, we define *physical similarity* to be similarity among several *physical properties* (e.g., molecular weight, number of rotational bonds, calculated log $P$ (cLogP), and number of hydrogen bond donors and acceptors) and *chemical similarity* to be similarity in chemical structure (e.g., molecular topology and functional groups) as measured using molecular fingerprints (e.g., MACCS[2]) and a Tanimoto[3] similarity score. The ligands contained in any

**Received:** April 18, 2011
**Published:** June 24, 2011

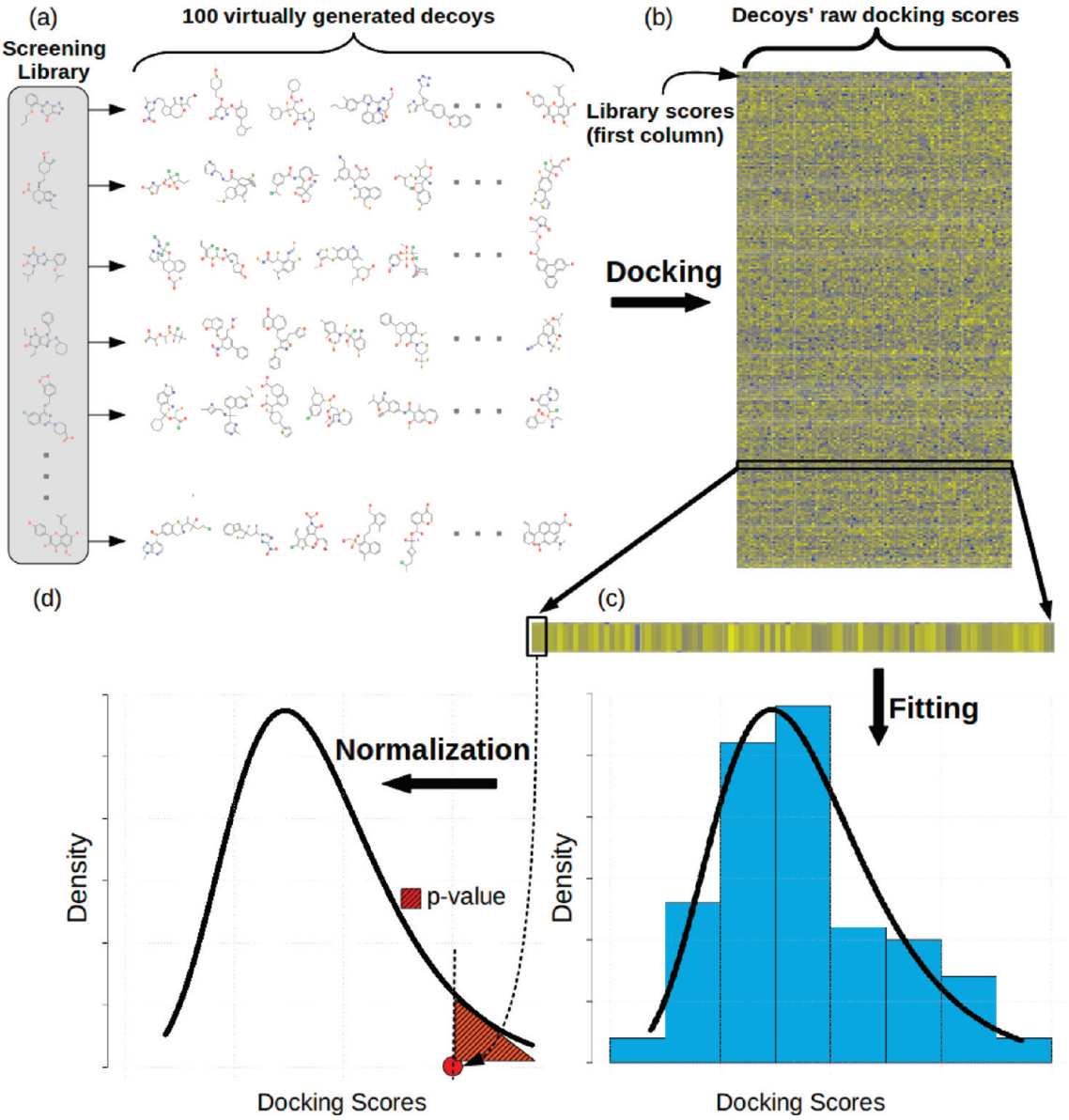



**Figure 1.** Illustration of the decoy-based ranking process. (a) Given a screening library, for each molecule in the library a corresponding set of virtual property-matched decoys is generated. (b) The library molecules and the virtual decoy sets are docked into the receptor, and raw docking scores are obtained. (c) Each set of raw docking scores is considered to be a random sample from a background distribution of scores of nonbinders corresponding to the library molecule for which the decoy set was generated. The docking scores are fit to a mixture model of extreme-value and uniform distributions (EVD-U). Alternatively, the empirical cumulative density function of the sample is computed. (d) The raw docking score of the library molecule is projected onto that distribution, and a corresponding *p*-value is calculated.

reasonable screening library have differences in both physical properties and chemical structures.

Unfortunately the scoring functions used in docking are often undesirably influenced by the physical properties of the ligand resulting in the possibility of bias in the screening results. Huang et al.[4] demonstrated screening results with respect to four different decoy sets over 12 different receptors. For each receptor, the same set of active ligands was separately merged with each one of the four different decoy sets. Each merged set was then screened against the protein target, and the ability to separate the actives from the decoys was measured. It was shown that separating a set of actives from physically dissimilar decoys is more achievable than it was to separate the actives from physically similar yet chemically dissimilar compounds. This result suggests that physical features rather than chemical interactions were playing a large role in the ranking process. In other words, the docking method may have simply learned to rank compounds on the basis of physical similarity rather than the strength of their chemical interaction with the target. We recently demonstrated a similar effect.[5] We generated different decoy sets for the same collection of active ligands such that every decoy set had the same distribution of physical properties as the active ligands except for a single difference. For example, in one experiment, decoys were generated with a number of hydrogen-bond acceptors (HBA) that was offset by −2, −1, +1, or +2 from the active ligands. We were able to arbitrarily increase or decrease the enrichment for these sets just by controlling for the number of HBA in the decoys which implied that enrichments

may be biased when the physical properties of decoys do not match the active compounds. Force-field and empirical based scoring functions are particularly prone to such biases when there is no correction for the additivity of scoring components. For example, an empirical scoring function that considers the explicit number of hydrogen bonds may be biased toward large ligands or deep receptors merely because the likelihood of forming such bonds increases.

An ideal docking algorithm and scoring function should be immune to these biasing phenomenon. New methods for unbiasing docking scores hold promise to normalize variations in screening libraries. Current methods attempt to reduce bias by docking the same ligand into multiple receptors and analyzing the distribution of scores. The first method taking this approach was the multiple active site correction (MASC) presented by Vigers and Rizzi.[6] The authors selected a set of $k$ auxiliary receptors in addition to the target receptor and docked every ligand to both the target and auxiliary receptors. For every ligand they obtained a sampling of $k$ scores, one from each receptor. The score distribution was used to normalize the interaction with the actual target into a calculated $z$-score. The underlying assumption behind this method is that the score of a true interaction for the given ligand would have a score significantly different from a background distribution of nonbinding scores for the same ligand. The nonbinding scores are expected to follow a normal distribution with a mean and variance reflecting the inherent ligand scoring bias. In other words, the interplay of the ligand's physical properties and the specifics of the scoring function will set the characteristics of the background distribution of nonbinding interactions. The binding score of the ligand in the target receptor can be judged in the context of this background distribution. A variant of this normalization was proposed by Jacobsson and Karlén.[7] They normalize the scores from a specific target prior to the normalization across multiple targets. First, they transform the raw docking scores to $z$-scores using the docking score distribution for every single target. Then, they correct the scores across targets by subtracting the mean (similar to MASC, but without dividing by the standard deviation). Their MASC variant outperformed the MASC method in most cases, though only moderately.

The MASC approach has several limitations. First, there is an additional computational effort because each docking experiment must be accompanied by a second set of experiments run against a number of additional receptors. Second, it is not clear how many and which additional receptors should be used to reliably estimate the background distribution. While Vigers and Rizzi made an attempt to estimate the number, the results presented by Jacobsson suggest that this number may vary and is probably larger than the 7−9 suggested by Vigers and Rizzi or the 12 additional receptors used by the FRED[8] docking package. Fukunishi et al.[9] showed that the screening results are sensitive to the selection of the additional targets primarily due to the large range of observed scores. They suggested the use of ligand rank across the $k + 1$ different receptors (i.e., a score between 1 and $k + 1$) rather than raw interaction score to reduce the sensitivity to large score variations. Unfortunately, the proposed strategy does not address the underlying problem of how to select the set of receptors. We hypothesize that these issues contribute to the poor ranking performance of the MASC-based approaches reported by Jacobsson and Karlén.[7] Recently, Swann et al.[10] introduced a probabilistic method that combines structure- and ligand-based screening to predict bioactivity. In their method, a normal distribution is fit to the docking scores of both the library ligands and the decoy set. The normalization allows them to quantify the significance of a docking score; however, as a single normal distribution is fit to all docking scores, the normalization does not induce a rank order different from that obtained with the raw docking scores. Because their decoy set is drawn from compounds in the ZINC database,[11] which is the collection of commercially available compounds, it can be difficult to find enough physical property-matched decoys for certain library molecules. In previous work, we have shown that libraries of mismatched physical property distributions can result in dissimilar binding scores and subsequent score distributions.[5] In this work we show that it may be possible to avoid such biases if for every library molecule we consider a separate score distribution generated using physical property-matched decoys.

We present a framework for normalizing docking scores using the distribution of scores obtained from property-matched virtual decoy sets (VDS). For every query molecule in a screening library a set of 100 virtual decoys is generated (Figure 1a). Each query molecule and its matched decoys are then docked into the target receptor (Figure 1b). The scores of the 100 property matched decoys are considered random samples from a distribution of nonactive scores. We parametrize this distribution (Figure 1c) and use it to produce a $p$-value for the query molecule (Figure 1d). We hypothesize that the score of an active query molecule will differ from the scores of nonactive ligands and will therefore lie in the sparse extreme regions of the distribution. In other words, we compute a $p$-value corresponding to the likelihood that the score of the query molecule was drawn from the distribution of nonactive decoy scores.

Our method utilizes two classes of distributions to fit the raw docking scores of the decoy molecules. The first was a parametric mixture of extreme value and uniform distributions (EVD-U), and the other was the nonparametric empirical cumulative density generated from the sample (ECDF). The extreme value distribution has the most intuitive interpretation in the context of molecular docking since a docking algorithm returns the best (or extreme) score obtained across all searched docking poses. Therefore, the docking scores computed for a library of compounds are expected to follow an extreme value distribution. In our case, the docking scores of a set of decoy molecules are expected to follow an extreme value distribution that characterizes all decoy molecules sharing the physical properties of the query molecule. Our use of a uniform distribution in conjunction with EVD aims to reduce the sensitivity of the maximum likelihood estimator to outliers. While the EVD-U model is justifiable given the problem domain, our secondary consideration of the ECDF, a nonparametric distribution, is aimed to minimize assumptions of distribution structure.

The primary advantage of ligand-based normalization using decoys over receptor-based normalization (e.g., MASC) is the fact that we can control and minimize the variability in physical and chemical properties across the decoy ligands, whereas this variability is less manageable across receptors. Our results demonstrate that by controlling the decoys' physical and chemical properties, we can improve the enrichment for diverse screening libraries. Our approach can also be used to identify potential biases in a scoring function. One disadvantage of our decoy-based approach is the increase in computational effort required to generate and test the virtual decoy set. In Discussion we discuss this computational limitation and possible ways to mitigate it.

## RESULTS

We performed a series of experiments to benchmark the VDS-based normalizations against the traditional approach of ranking ligands based on raw docking scores. Our data set consists of a set of receptors and for each receptor a set of query ligands. The query ligands include both active and nonactive compounds. Because structurally similar active ligands may bias the evaluation of ranking,[12] we used a set of 13 DUD protein targets each of which have more than 15 diverse active ligands (based on Andrew Good's DUD clustering[13] (http://dud.docking.org/clusters), June 13, 2011). For each protein target, we treated the *DUD active ligands* as our *active compounds*. We next selected a random set of *nonactive compounds* from the target's corresponding *DUD decoys*. The number of selected nonactives is 20 times the number of actives. It is worth noting that although the DUD decoys are not likely to be active, there may in fact be some false negatives; that is, some of the DUD decoys may actually bind the target receptor. For each query compound (both active and nonactive) we generated a set of physical property-matched virtual decoys as detailed in Methods. In short, our decoy generation approach utilizes Tripos's EA-Inventor (http://tripos.com, June 13, 2011) which itself employs a genetic algorithm to generate, denovo, a population of small molecules. EA-Inventor is guided by user-specified fitness and penalty functions. In our case, the fitness function rewards similarity to a reference query molecule with respect to seven physical properties: molecular weight, number of rotational bonds, number of hydrogen donor/acceptor, cLogP, formal charge, and topological polar surface area (TPSA). The penalty function attempts to prevent the simultaneous inclusion of two compounds whose Tanimoto coefficient[3] (TC) is greater than a specified threshold (using the MACCS fingerprints[2]). The fitness function ensures that the decoys have similar physical properties to the query ligand and the penalty function ensures that the decoys are chemically diverse. Finally, we docked the complete data sets (i.e., the actives, nonactives, and all virtual decoys) using the Glide[14] and the eHiTS[15] docking algorithms. In the remainder of this section we compare the results of the traditional (unnormalized) approach to both the EVD-U and ECDF normalization procedures. In the docking experiments, we of course hide the active/nonactive labels for each compound and determine if the ranking algorithm can produce an enriched output. We compare the methods using two criteria. The AUC, which is the area under the receiver operator curve (ROC), and the retrieval rate, which is the ratio between the observed number of actives molecules at some *k*% of the data set and total number of actives in the data set.

**Experiments Using a 0.8 Tanimoto Coefficient Threshold.** In our initial set of experiments we generated decoys using a Tanimoto coefficient penalty threshold of 0.8 (using MACCS fingerprints). Because the genetic algorithm used by EA-Inventor only penalizes the fitness of molecules lying above the TC threshold but does not enforce a strict cutoff, the generated set of compounds may contain molecules with TC > 0.8. In practice, specifying a threshold of 0.8 typically eliminates all compounds with a TC > 0.9 to any other included ligand (Supporting Information Figure 2). Previous work[4,5] considered a 0.9 MACCS (or CACTVS type 2 fingerprints) threshold (roughly equivalent to 0.7 using the Daylight fingerprints[16]) to sufficiently enforce chemical dissimilarity between reference molecules and their property-matched decoys. Enforcing a large chemical dissimilarity, as measured by the MACCS fingerprints, increases the likelihood that the decoys are nonactive.[17] For each active and nonactive query molecule, 100 property-matched virtual decoys are generated. Supporting Information Figure 1 illustrates the excellent

**Table 1. Summary of the Ranking Performance between the EVD-U and ECDF Normalization Methods to the Naive Ranking[a]**

| | | Glide | | | | | eHiTS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| target | | naive | EVD-U | | ECDF | | naive | EVD-U | | ECDF | |
| InhA | 3% | 0.14 | 0.18 | 0.18 | 0.23 | 0.14 | 0.05 | 0.00 | 0.09 | 0.09 | 0.09 |
| | 20% | 0.18 | 0.55 | 0.41 | 0.50 | 0.45 | 0.32 | 0.36 | 0.45 | 0.36 | 0.45 |
| | AUC | 0.39 | 0.78 | 0.71 | 0.72 | 0.74 | 0.67 | 0.72 | 0.73 | 0.71 | 0.74 |
| P38 MAP | 3% | 0.10 | 0.10 | 0.15 | 0.10 | 0.15 | 0.10 | 0.10 | 0.05 | 0.10 | 0.00 |
| | 20% | 0.25 | 0.30 | 0.50 | 0.25 | 0.50 | 0.20 | 0.20 | 0.20 | 0.20 | 0.25 |
| | AUC | 0.69 | 0.68 | 0.77 | 0.64 | 0.72 | 0.39 | 0.42 | 0.48 | 0.42 | 0.49 |
| AChE | 3% | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 | 0.18 | 0.06 | 0.06 | 0.18 | 0.00 |
| | 20% | 0.06 | 0.24 | 0.35 | 0.35 | 0.18 | 0.53 | 0.59 | 0.53 | 0.59 | 0.65 |
| | AUC | 0.42 | 0.60 | 0.60 | 0.61 | 0.60 | 0.76 | 0.72 | 0.72 | 0.72 | 0.73 |
| PDGFrb | 3% | 0.10 | 0.10 | 0.05 | 0.05 | 0.05 | 0.05 | 0.14 | 0.05 | 0.10 | 0.00 |
| | 20% | 0.15 | 0.20 | 0.25 | 0.25 | 0.25 | 0.29 | 0.52 | 0.52 | 0.43 | 0.57 |
| | AUC | 0.46 | 0.62 | 0.56 | 0.58 | 0.58 | 0.57 | 0.74 | 0.76 | 0.67 | 0.72 |
| FXa | 3% | 0.37 | 0.26 | 0.21 | 0.26 | 0.16 | 0.05 | 0.11 | 0.16 | 0.05 | 0.05 |
| | 20% | 0.58 | 0.58 | 0.63 | 0.58 | 0.53 | 0.74 | 0.42 | 0.63 | 0.53 | 0.63 |
| | AUC | 0.81 | 0.80 | 0.79 | 0.80 | 0.79 | 0.77 | 0.74 | 0.75 | 0.75 | 0.74 |
| VEGFr2 | 3% | 0.29 | 0.29 | 0.35 | 0.26 | 0.35 | 0.03 | 0.03 | 0.06 | 0.00 | 0.03 |
| | 20% | 0.68 | 0.58 | 0.55 | 0.65 | 0.61 | 0.23 | 0.26 | 0.45 | 0.16 | 0.29 |
| | AUC | 0.83 | 0.82 | 0.78 | 0.81 | 0.82 | 0.52 | 0.59 | 0.69 | 0.59 | 0.63 |
| CDK2 | 3% | 0.31 | 0.31 | 0.28 | 0.28 | 0.34 | 0.31 | 0.19 | 0.22 | 0.19 | 0.25 |
| | 20% | 0.66 | 0.62 | 0.62 | 0.59 | 0.66 | 0.69 | 0.47 | 0.78 | 0.59 | 0.75 |
| | AUC | 0.75 | 0.75 | 0.73 | 0.74 | 0.75 | 0.87 | 0.81 | 0.86 | 0.83 | 0.86 |
| EGFr | 3% | 0.30 | 0.25 | 0.20 | 0.20 | 0.22 | 0.22 | 0.05 | 0.15 | 0.05 | 0.18 |
| | 20% | 0.72 | 0.60 | 0.65 | 0.55 | 0.70 | 0.60 | 0.52 | 0.57 | 0.50 | 0.72 |
| | AUC | 0.83 | 0.76 | 0.82 | 0.74 | 0.82 | 0.80 | 0.74 | 0.80 | 0.76 | 0.83 |
| HIVRT | 3% | 0.33 | 0.33 | 0.35 | 0.33 | 0.29 | 0.12 | 0.12 | 0.24 | 0.18 | 0.24 |
| | 20% | 0.67 | 0.67 | 0.65 | 0.67 | 0.71 | 0.41 | 0.65 | 0.47 | 0.65 | 0.53 |
| | AUC | 0.83 | 0.79 | 0.79 | 0.80 | 0.81 | 0.76 | 0.81 | 0.78 | 0.81 | 0.78 |
| SRC | 3% | 0.24 | 0.19 | 0.33 | 0.24 | 0.29 | 0.29 | 0.14 | 0.33 | 0.14 | 0.38 |
| | 20% | 0.76 | 0.71 | 0.81 | 0.76 | 0.86 | 0.90 | 0.67 | 0.90 | 0.71 | 0.90 |
| | AUC | 0.88 | 0.87 | 0.91 | 0.88 | 0.90 | 0.92 | 0.84 | 0.91 | 0.84 | 0.91 |
| COX-2 | 3% | 0.29 | 0.19 | 0.29 | 0.19 | 0.33 | 0.20 | 0.11 | 0.20 | 0.11 | 0.18 |
| | 20% | 0.71 | 0.62 | 0.62 | 0.71 | 0.76 | 0.57 | 0.32 | 0.55 | 0.30 | 0.55 |
| | AUC | 0.87 | 0.80 | 0.84 | 0.80 | 0.82 | 0.72 | 0.51 | 0.70 | 0.52 | 0.68 |
| PDE5 | 3% | 0.18 | 0.18 | 0.32 | 0.27 | 0.32 | 0.14 | 0.23 | 0.41 | 0.23 | 0.36 |
| | 20% | 0.73 | 0.64 | 0.68 | 0.68 | 0.68 | 0.64 | 0.82 | 0.77 | 0.82 | 0.73 |
| | AUC | 0.78 | 0.80 | 0.82 | 0.81 | 0.83 | 0.76 | 0.89 | 0.91 | 0.91 | 0.91 |
| ACE | 3% | 0.06 | 0.11 | 0.00 | 0.11 | 0.06 | 0.28 | 0.06 | 0.28 | 0.17 | 0.28 |
| | 20% | 0.22 | 0.28 | 0.28 | 0.22 | 0.28 | 0.56 | 0.33 | 0.50 | 0.39 | 0.44 |
| | AUC | 0.58 | 0.56 | 0.59 | 0.47 | 0.54 | 0.62 | 0.64 | 0.65 | 0.62 | 0.65 |
| mean | 3% | 0.21 | 0.19 | 0.21 | 0.18 | 0.22 | 0.16 | 0.10 | 0.18 | 0.11 | 0.15 |
| | 20% | 0.49 | 0.51 | 0.54 | 0.52 | 0.55 | 0.51 | 0.47 | 0.56 | 0.48 | 0.57 |
| | AUC | 0.70 | 0.74 | 0.75 | 0.72 | 0.75 | 0.70 | 0.71 | 0.75 | 0.70 | 0.74 |

[a] Values for 3% and 20% indicate the fraction of the actives retrieved after ranking the corresponding percentage of the database. AUC denotes the area under the ROC curve. $TC_{0.8}$ and $TC_{0.5}$ indicate results obtained using VDS generated with a 0.8 and a 0.5 TC threshold, respectively.

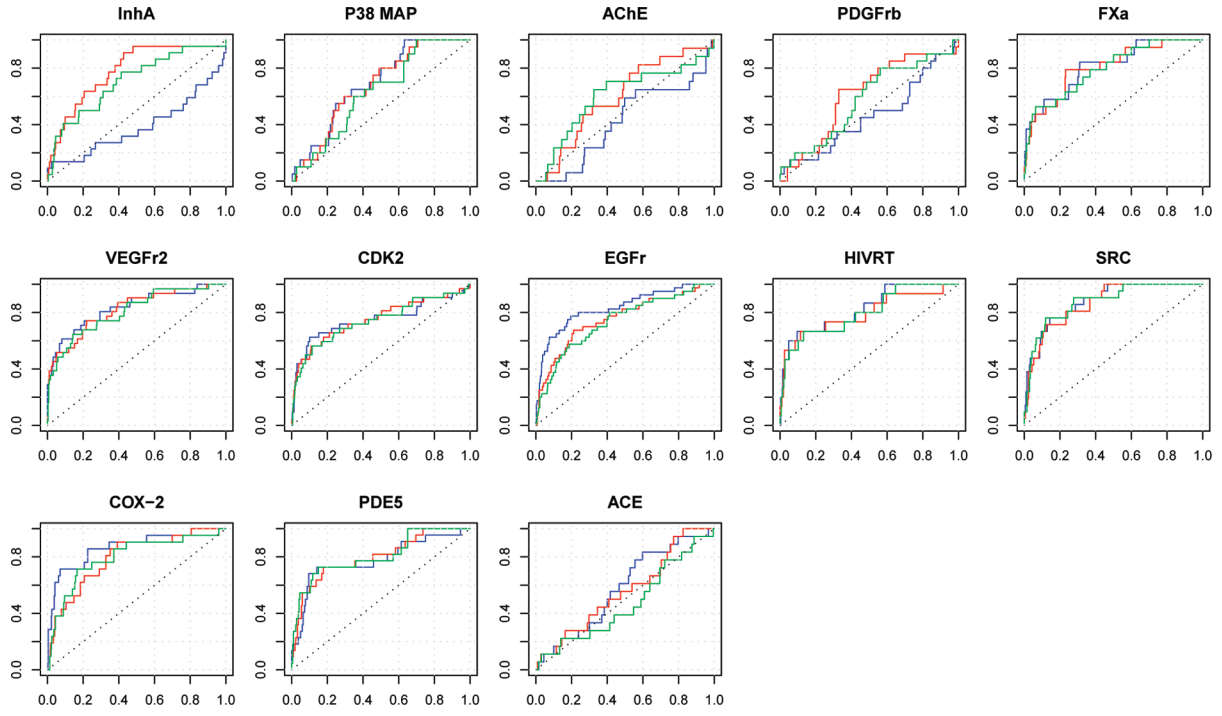



**Figure 2.** Docking results using Glide and a 0.8 TC similarity threshold. The fraction of the docking ranked database (*x*-axis) is plotted against the fraction of the active molecules (*y*-axis). Red lines represent ranking results using the EVD-U normalization, green lines represent ranking results using the ECDF normalization, and blue lines represent the naive ranking results.

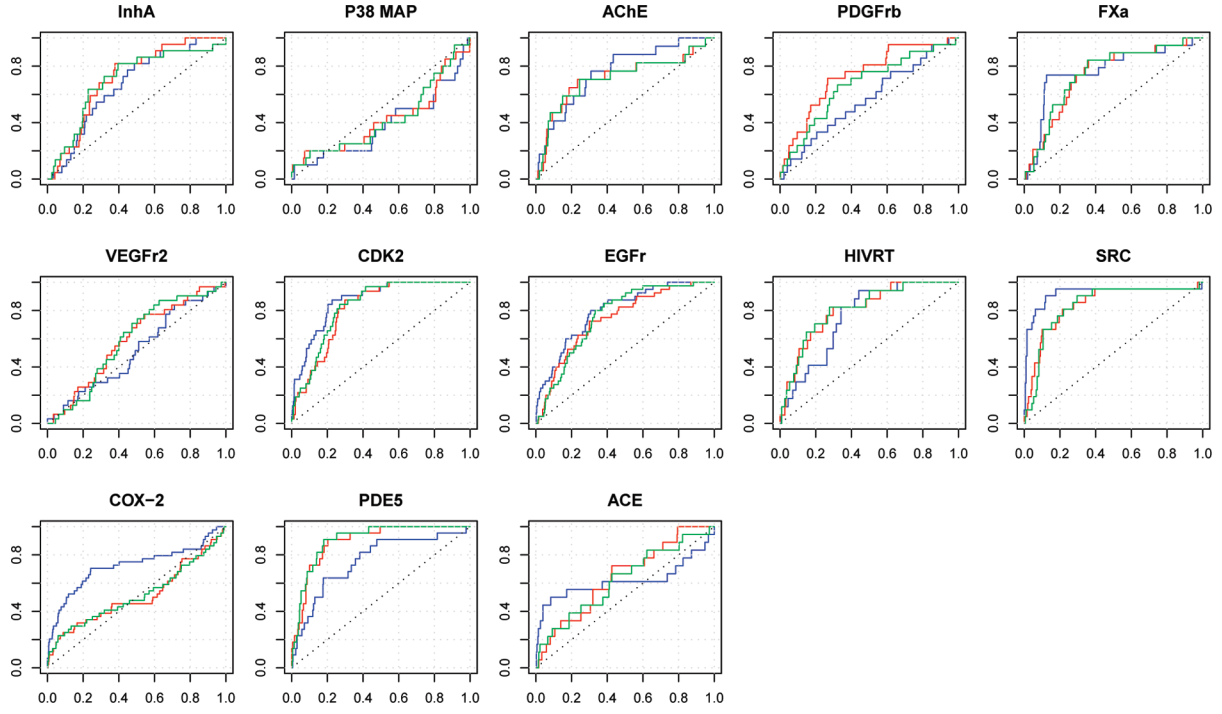



**Figure 3.** Docking results using eHiTS and a 0.8 TC similarity threshold. The fraction of the docking ranked database (*x*-axis) is plotted against the fraction of the active molecules (*y*-axis). Red lines represent ranking results using the EVD-U normalization, green lines represent ranking results using the ECDF normalization, and blue lines represent the naive ranking results.

agreement between the reference molecules and the VDS over the seven matching physical properties. We docked the generated decoys along with their corresponding active and nonactive molecules into the 13 target receptors and obtained their raw docking scores. We then normalized the raw docking scores of each active and nonactive molecule using the EVD-U and ECDF

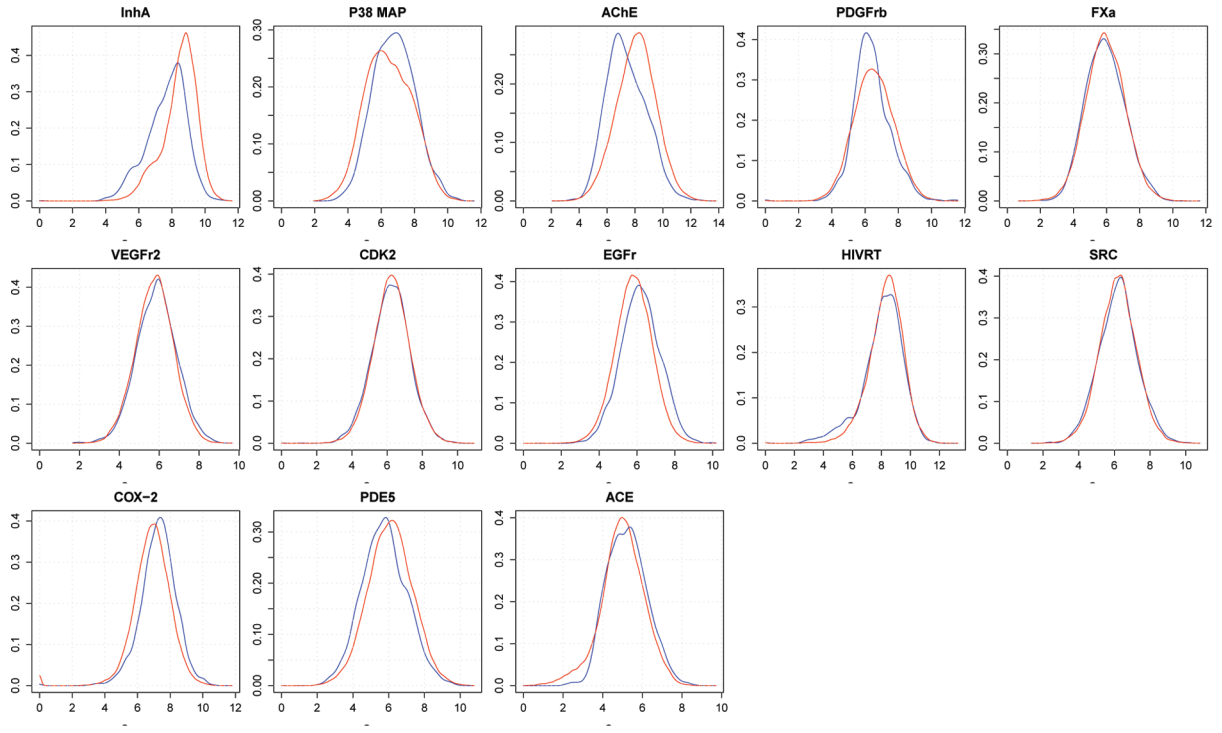



**Figure 4.** Comparison of score distributions between decoys generated for active-VDS molecules (blue) and decoys generated for nonactive-VDS molecules (red) using Glide and a 0.8 TC similarity threshold. The *x*-axis shows the distribution of scores, and the *y*-axis shows the density. Active-VDS are the virtual decoys generated for a set of actives, and nonactive-VDS are the set of virtual decoys generated for a set of nonactive molecules. The docking scores of the two sets are expected to have similar distributions because of the following: (i) the active and nonactive molecules have similar physical properties due to the DUD generation process, and (ii) both active- and nonactive-VDS are generated with matching physical properties (to each corresponding library molecule) but dissimilar chemical properties (via the TC thresholds). However, in many cases, the distributions are different.

normalization methods (see Methods) and obtained the probability that each query molecule was drawn from the distribution of decoys. We ranked the query molecules on the basis of their *p*-values, where a low *p*-value indicates a low probability of being nonactive (i.e., there is a low-probability that the score of the query molecule was drawn from the set of scores of nonbinding decoy compounds). The ranking results are summarized in Table 1 and illustrated using ROC plots in Figure 2 and Figure 3 for Glide and eHiTS, respectively. Because of the small size of our data set, we measured the fraction of actives discovered at 3 and 20% of the screening set and the overall AUC. We separately compared normalization using the two background distributions to the non-normalized naive ranking. We found that when using a TC threshold of 0.8, the non-normalized scores occasionally yield better ranking than the normalized ones for both docking algorithms. However, in almost all cases where the naive approach yielded better ranking, the normalized methods still find a substantial number of actives. For example, using normalization on eHiTS' docking results decreases the success rate for the COX2 and ACE targets (Figure 3). On the other hand, we observed a substantial improvement in ranking Glide's docking results for the InhA target (Figure 2) when applying normalization.

To understand the mixed performance of the normalization methods, we analyzed the raw docking scores of the generated decoys. For clarity, we refer to the virtual decoys generated for a set of active and a set of nonactive molecules as active-VDS and nonactive-VDS, respectively. We compared the distributions of the raw docking scores for both the active-VDS and nonactive-VDS sets. The set of active and nonactive molecules, which were all drawn from the DUD, have similar physical properties due to the underlying DUD generation process.[4] The active-VDS and nonactive-VDS each have physical properties similar to the library molecule for which they were generated. Because the corresponding VDSs are generated with matching physical properties but dissimilar chemical properties (via the TC thresholds described above), we expected similar score distributions for both the active-VDS and the nonactive-VDS. In other words, we expected the active, the nonactive, the active-VDS, and the nonactive-VDS compounds to have similar physical property distributions and thus similar docking scores. Figure 4 and Figure 5 illustrate the score distributions of the two sets for each of the 13 targets using Glide and eHiTS. Blue and red lines indicate the distributions of scores for the active-VDS and nonactive-VDS sets, respectively. Ideally the score distributions for the active-VDS and nonactive-VDS should be the same; however, in many cases, the distributions are different. For example, in Figure 4 for the InhA and AChE receptors, the distribution of nonactive-VDS is shifted toward higher docking scores compared to the active-VDS. Dissimilar docking score distributions for the active-VDS and nonactive-VDS could indicate that the VDS decoys were too chemically similar to the active or nonactive molecule for which they were generated. That is, the active-VDS could contain compounds that were similar enough to the known actives that the decoys themselves might be active. In the cases where the active-VDS and active score distributions were different than the nonactive-VDS and nonactive score distributions, our normalization may have improved or worsened the ranking for the wrong reason. For example, consider the case of

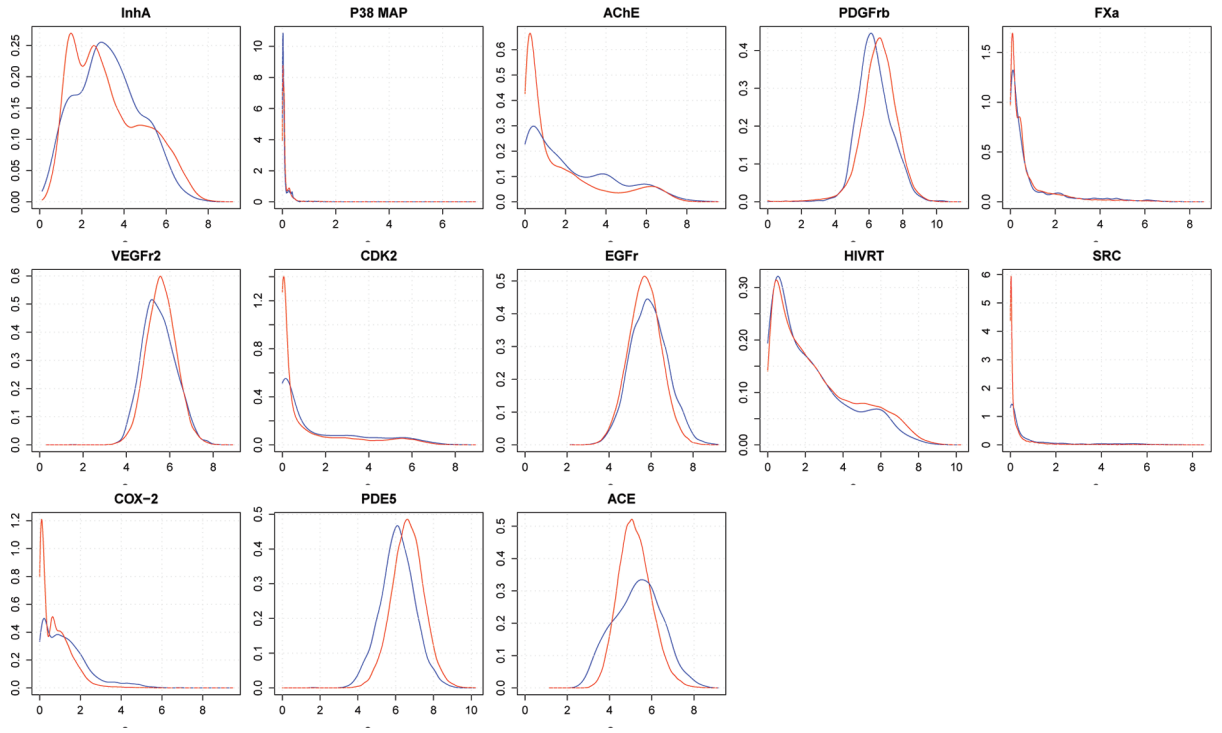



**Figure 5.** Comparison of score distributions between decoys generated for active molecules (blue) and decoys generated for nonactive-VDS molecules (red) using eHiTS and a 0.8 TC similarity threshold.

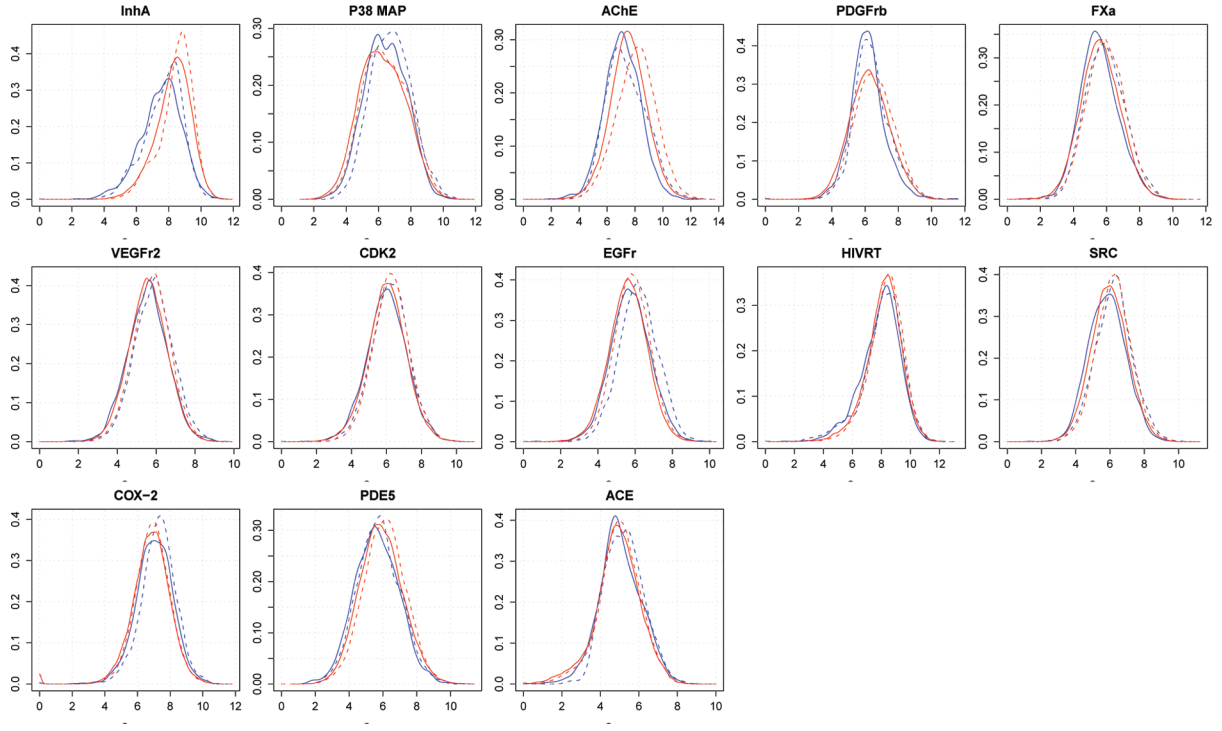



**Figure 6.** Comparison of score distributions between decoys generated for active molecules (blue) and decoys generated for nonactive-VDS molecules (red) using Glide and either a 0.8 (dashed) or a 0.5 (solid) TC similarity threshold.

InhA. The active-VDS molecules received worse docking scores than the nonactive-VDS molecules. Therefore, the normalized ranking is likely better than naive. This is indeed the case. Our normalization method increases the AUC by 0.39 (from 0.39 to 0.78). On the other hand, the COX-2 and ACE receptor scores demonstrate the opposite phenomenon. The active-VDS molecules score higher than the nonactive-VDS molecules. In this case, normalization reduces the ranking accuracy. We hypothesized that


dx.doi.org/10.1021/ci200175h |*J. Chem. Inf. Model.* 2011, 51, 1817–1830

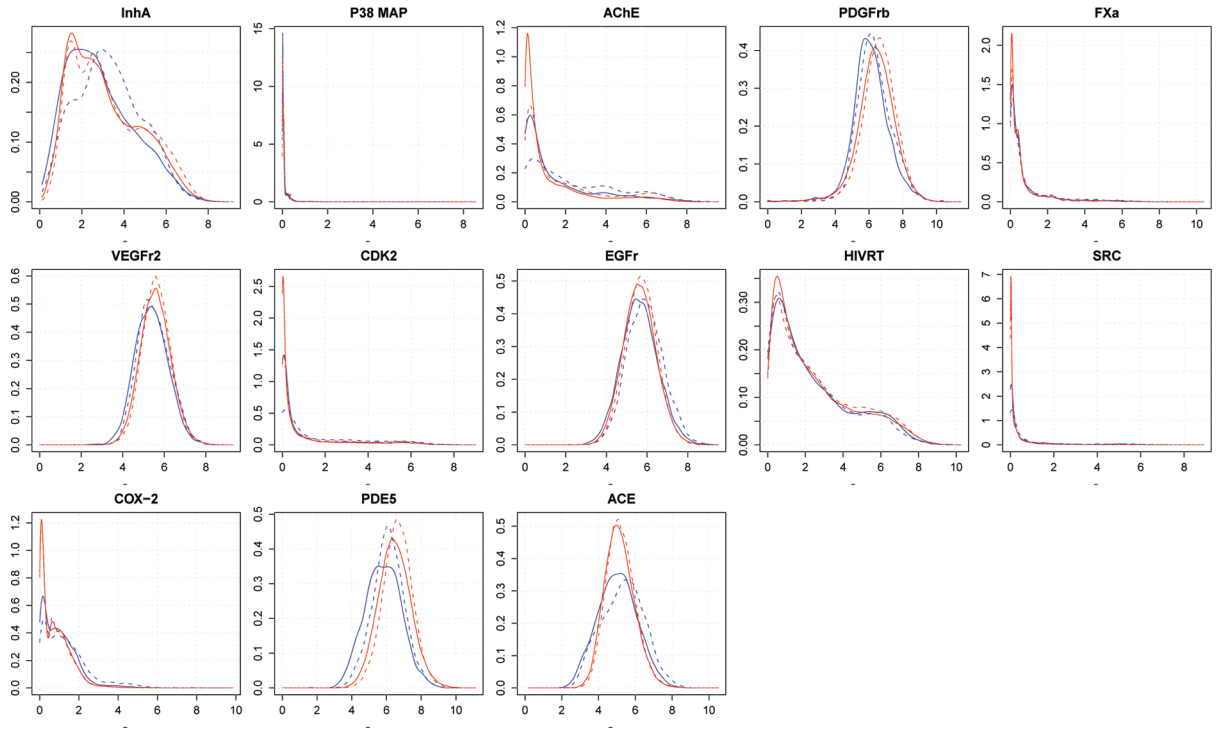



**Figure 7.** Comparison of score distributions between decoys generated for active molecules (blue) and decoys generated for nonactive-VDS molecules (red) using eHiTS and either a 0.8 (dashed) or a 0.5 (solid) TC similarity threshold.

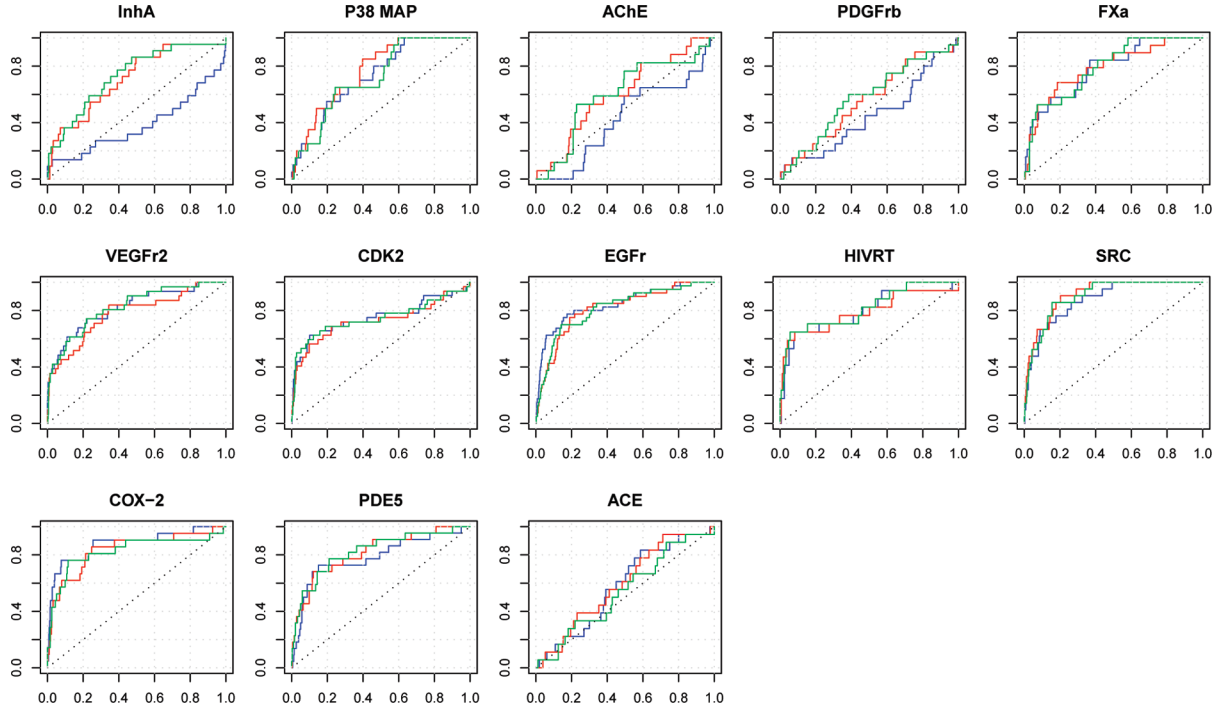



**Figure 8.** Docking results using Glide and a 0.5 TC similarity threshold. The fraction of the docking ranked database (*x*-axis) is plotted against the fraction of the active molecules (*y*-axis). Red lines represent ranking results using the EVD-U normalization, green lines represent ranking results using the ECDF normalization, and blue lines represent the naive ranking results.

the observed distribution differences could be corrected by modifying the TC threshold to reduce the chemical similarity between the decoy set and the original library molecule (see next section).

**Experiments Using a 0.5 Tanimoto Coefficient Threshold.** To reduce the difference between the distributions of scores obtained for the active- and nonactive-VDS, we generated a second set of decoys using a more stringent penalty term of

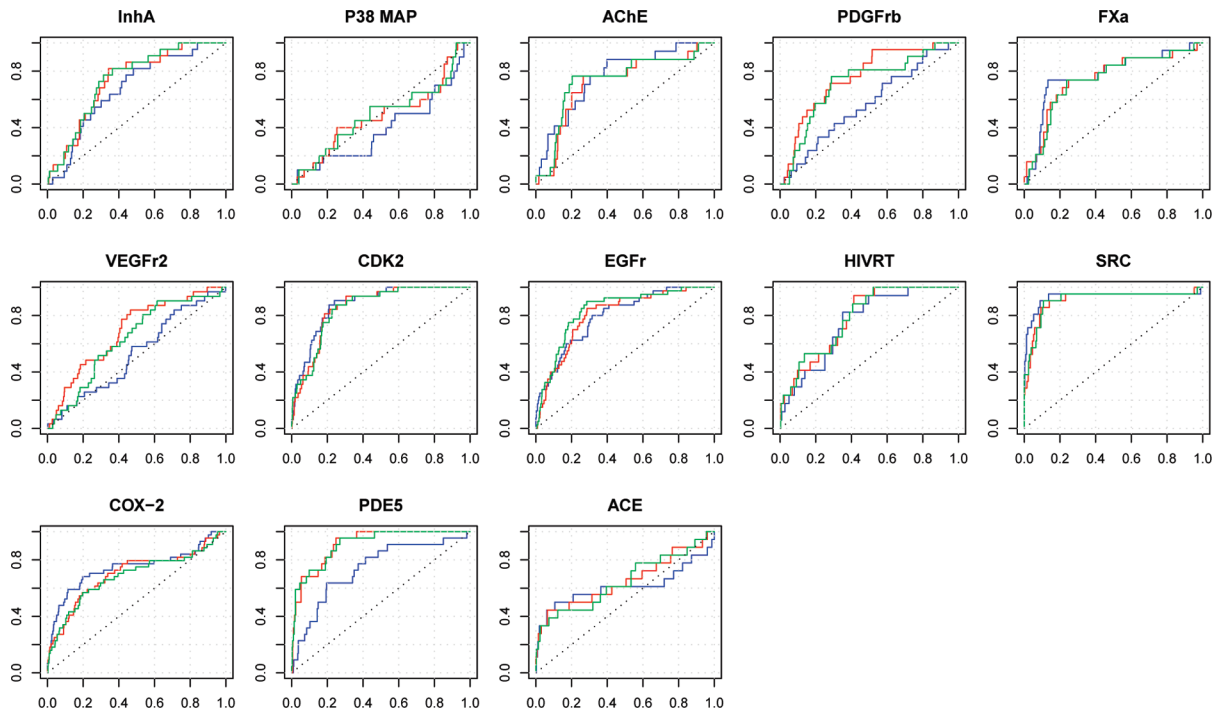



**Figure 9.** Docking results using eHiTS and a 0.5 TC similarity threshold. The fraction of the docking ranked database (*a*-axis) is plotted against the fraction of the active molecules (*y*-axis). Red lines represent ranking results using the EVD-U normalization, green lines represent ranking results using the ECDF normalization, and blue lines represent the naive ranking results.

0.5 TC. We performed the docking experiments and obtained raw docking scores as described in the previous sections. Figure 6 and Figure 7 illustrate the differences between the distributions of raw docking scores using the two different TC penalty terms (0.5 and 0.8) for Glide and eHiTS, respectively. Using a tighter 0.5 threshold (solid lines) yields densities that are more similar than densities obtained using the initial 0.8 threshold (reproduced as dashed lines). We measured the difference between the distributions using the Kullback−Leibler divergence (KL; see Methods). The KL values for the distributions of scores obtained with Glide and eHiTS appear in Supporting Information Tables 1 and 2, respectively. There is a statistically significant reduction in the KL values from sets generated with TC = 0.8 to sets generated with TC = 0.5 ($p$-values of 0.003 and 0.01 for Glide and eHiTS, using one-side paired $t$-test).

Using the new virtual decoy sets generated using TC = 0.5 we normalized the docking scores using EVD-U and ECDF. The results are shown in Figure 8 and Figure 9 for Glide and eHiTS, respectively. We observed that the normalization is generally helpful in that the AUCs are either improved or remained the same when shifting from 0.8 to 0.5 penalty threshold. More important was the result that in cases where the previous performance was poor, such as COX2 (Glide and eHiTS) and ACE (eHiTS), the improvement was significant resulting in ranking comparable to the naive. For COX2, using the EVD-U normalization, the retrieval rate of active molecules at 3% of the database increased from 0.19 to 0.29 for Glide and from 0.11 to 0.20 for eHiTS. The retrieval rate for ACE increased from 0.06 to 0.28 for eHiTS. Overall, the average retrieval rate was increased from 0.19 to 0.21 for Glide and from 0.10 to 0.18 for eHiTS, yielding slightly better ranking than the naive approach. The complete results are summarized in Table 1.

**Results with Two Validated Systems.** In the previous section, we showed that changing the TC threshold from 0.8 to 0.5 improved the chemical diversity of compounds while maintaining a decoy set with matched physical properties. However, even when a TC threshold of 0.5 was used some differences remained between the distribution of naive scores for the active-VDS and the nonactive-VDS. There are many potential causes of this phenomenon, including the possibility that many scoring functions have been inadvertently optimized for those protein systems commonly used in training. In other words, there are only so many protein ligand systems that we can use to build and train scoring functions, many of which are contained in the DUD. We next chose to evaluate our method on two systems not included in the DUD and for which both the active and nonactive molecule sets have been experimentally validated. The actives are confirmed active and the nonactives are confirmed nonactive. These data are publicly provided by the Shoichet group (http://shoichetlab.compbio.ucsf.edu/take-away.php, June 13, 2011). These sets are particularly challenging since many of the nonactive molecules were experimentally tested only because there was some preliminary indication of activity. Particularly, the nonactives were initially thought to be potential binders and were only assigned to be nonbinders after wet-laboratory experimentation. Similar to Mysinger and Shoichet,[18] we tested two receptors, the T4-lysozyme with the Leu99Ala substitution (L99A) and the cytochrome-*c*-peroxidase (CCP) with the Trp191Gly substitution (W191G). The receptors and ligands were prepared in the same manner as in our previous experiments. We generated decoy sets for the T4-lysozyme and CPP proteins using a TC penalty term of 0.5. The ranking results using Glide and eHiTS are illustrated in Figure 10a,b. The results show that the VDS-based normalization moderately improves the Glide-based ranking for both targets compared to the naive

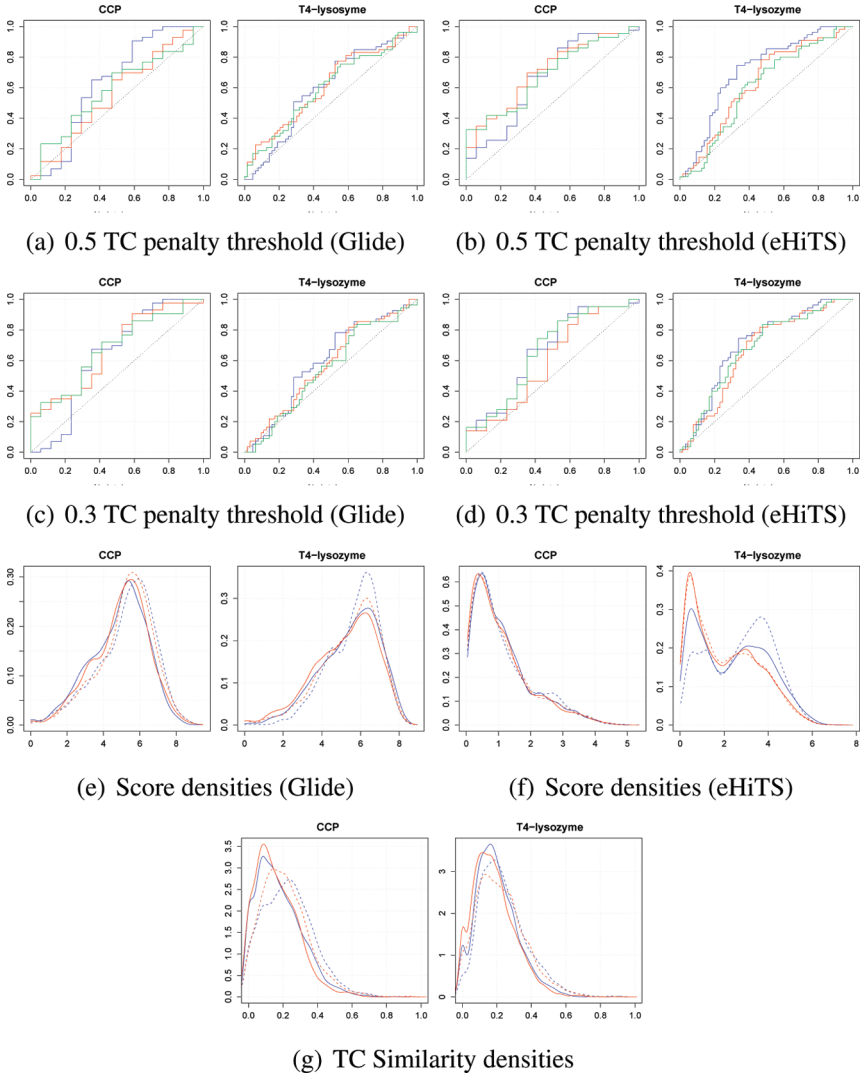



**Figure 10.** Analysis of normalization results for the CPP and T4-lysozyme validated systems. (a, b) Docking results using 0.5 and 0.3 TC similarity threshold for Glide and eHiTS. The fraction of the docking ranked database (*a*-axis) is plotted against the fraction of the active molecules (*y*-axis). Three ranking methods are shown, EVD-U normalization (red), ECDF normalization (green), and raw ranking results (blue). (e, f) Comparison of score distributions between the active-VDS molecules (blue) and the nonactive-VDS molecules (red) using either a 0.5 (dashed) or a 0.3 (solid) TC similarity threshold for Glide and eHiTS, respectively. (g) Comparison of the similarity distributions between active-VDS molecules (blue) and nonactive-VDS molecules (red) using either a 0.5 (dashed) or a 0.3 (solid) TC similarity threshold. Similarity is measured using the Tanimoto coefficient of the MACCS fingerprints between VDS decoys to their corresponding library molecule.

ranking. The use of eHiTS-based rankings provides a substantial improvement in early enrichment for the CPP data set, and no improvement for the T4-lysozyme data set compared to the naive ranking. The density plot of the raw docking scores (Figure 10e,f, dashed lines) demonstrates that a difference between the densities of scores for the T4-lysozyme docking scores using eHiTS. This provides a possible explanation for the eHiTS T4-lysozyme performance.

As a final test, we generated a second set of virtual decoys using a penalty term of 0.3 TC. The ranking results for this set are illustrated in Figure 10c,d. The more stringent penalty makes the active- and nonactive-VDS T4-lysozyme score distributions more similar although there are still differences (Figure 10f, solid lines). The TC = 0.3 generated decoys caused a substantial improvement in the ranking of the CCP ligands using Glide. On the other hand, the ranking performance for the T4-lysozyme using Glide and CCP using eHiTS decreased. This may suggest that optimal parametrization to generate decoy sets may vary between different docking algorithms and different targets.

**Experiments Using Randomly Selected Druglike Libraries.** In our first two experiments, decoys from the DUD were used as nonactive molecules in the screening sets. Although physical dissimilarity may appear between active ligands; for each active ligand, the active and its corresponding DUD decoys (i.e., the nonactives in our library) share physical similarity. This partial physical property matching is due to the process by which the DUD set was originally selected. Because the overall physical similarities between actives and nonactives were somewhat similar, the benefits provided by our normalization were less pronounced. To demonstrate the effects of our normalization in a less balanced scenario, we performed additional screening experiments using randomly selected nonactive molecules that

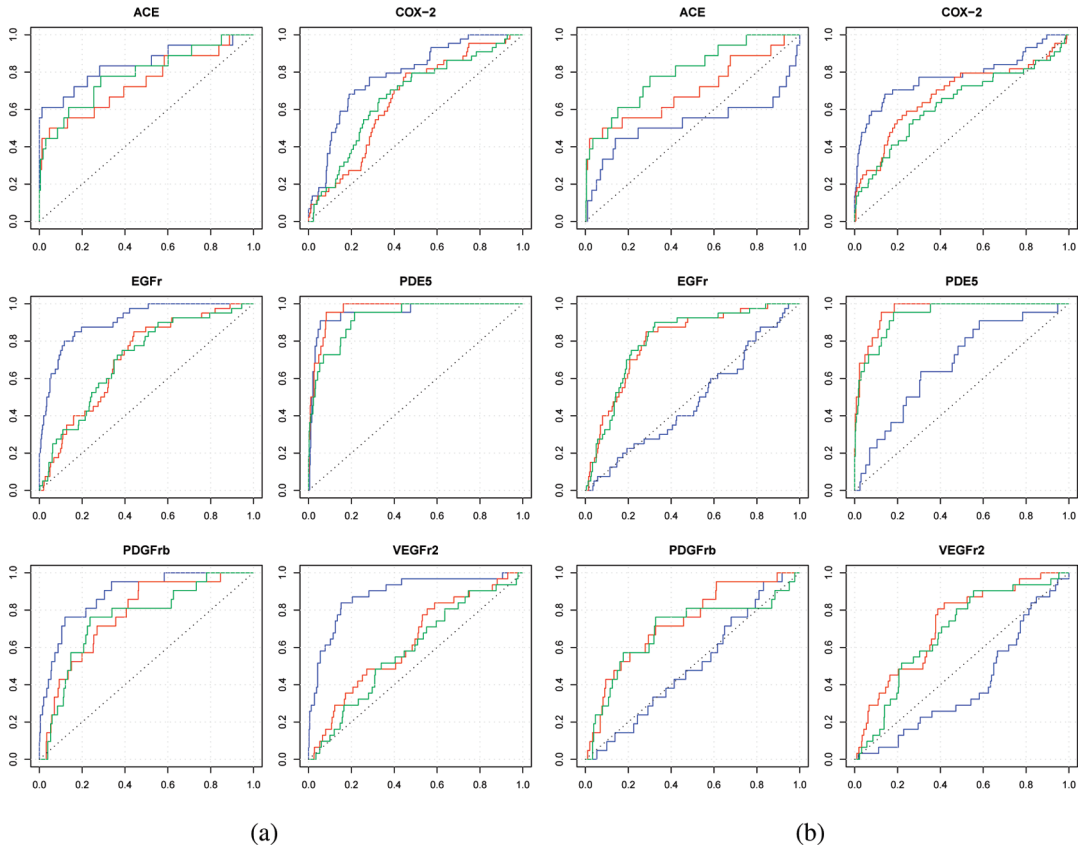



**Figure 11.** Docking results using two randomly selected libraries of 400 druglike molecules coupled with DUD's active ligands (average molecular weight of ~350 Da). (a) Library with an average molecular weight of ~250 Da. (b) Library with an average molecular weight of ~450 Da. Docking was performed using eHiTS and virtual decoys were generated using a 0.5 TC similarity threshold. The fraction of the docking ranked database (*x*-axis) is plotted against the fraction of the active molecules (*y*-axis). Ranking results using EVD-U normalization (red), ranking results using ECDF normalization (green), and naive ranking results (blue).

were not controlled for physical similarity. In these experiments, we used the same active ligands as previously described for each of our 13 receptors. We created a new set of nonactive compounds by randomly selecting a nonredundant set of 400 compounds from the ZINC clean-drug-like subset (see Methods). An analysis of the physical properties of the randomly selected library revealed that indeed there was a significant difference between the average molecular weights of the actives and nonactives. While the actives averaged ~350 Da, the nonactives (by random chance) averaged ~250 Da. To explore the potential impact of this difference on the ranking, we randomly selected another nonredundant set of 400 nonactive compounds but this time we controlled for an average molecular weight of 450 Da (see Methods). The ranking results of the low and high molecular weight libraries are illustrated in Supporting Information Figures 3 and 4, respectively. The comparison of the ranking results from the low and high molecular weight libraries suggests significant differences for many protein targets and demonstrates the effect that molecular weight variation can have on the ranking.

Building from these results, we applied our normalization method to a subset of six receptors for which we observed substantial ranking differences between the low and high molecular weight libraries. The normalized ranking results for the low and high molecular weight libraries are illustrated in Figure 11a,b, respectively. It is apparent that the normalized ranking is more consistent than the naive ranking and is less sensitive to the selection of the screening library. This is not surprising because the normalization effectively minimizes the physical variation within any screening library. These ranking experiments also demonstrate a subtle yet important point regarding the evaluation of virtual docking results. It is often unclear if observed ranking results of docking algorithms are right for the right reasons and, therefore, how consistent the docking results might be. For example, considering the difference of the naive ranking for the EGFr receptor using the two libraries, it is difficult to infer how successful the screening might be for any arbitrary library. By factoring out possible physical effects, our normalization method allows one to evaluate screening results in a consistent manner independent of the library in used.

## DISCUSSION

In this work we demonstrated the application of virtual decoys for normalizing scoring function bias with respect to the task of ranking. We encountered unexpected differences between the distribution of docking scores obtained for physical property matched decoys generated using active and nonactive molecules. Following Huang et al.,[4] we assumed that selection of a property-matched decoy set with pairwise chemical similarities below a TC of 0.9 (using MACCS fingerprints) would generate sufficient diversity of chemical features for benchmarking docking algorithms. Our docking experiments suggest that this is not the case.

We showed that distribution of docking scores of two chemically diverse sets of decoys, generated for active and nonactive molecules, differ across different receptors. Furthermore, we showed that improved ranking performance and more similar distributions of scores came from reducing the TC threshold. It is possible that because the DUD employed a cutoff of 0.9, it includes decoys that are harder to separate from the actives due to high chemical similarity. In that case, using the 0.9 TC threshold may result in generating decoys that are too chemically similar to the active data set used to fit the scoring functions and thus these decoy compounds may be scored favorably by these functions. This phenomenon could be further complicated by any scoring function overfitting. Clearly, this should not be surprising because scoring functions will almost inevitably be overfit, over time, to any popular benchmark just due to the constant attempt to improve against that benchmark. These results demonstrate a limitation of the current docking-based virtual screening benchmarks and the need for diversification. That is, the potential reuse of commonly available binding data can influence the measurement of docking performance. In previous work[5] we suggested that using virtual decoys may mitigate the common problem of overfitting to a single benchmark. Unfortunately, this approach only addresses half the bias; it can diversify the decoy set, but the set of binders remains the same. Getting new diverse binders for benchmarks is clearly a hard task, but, currently, it is very likely that scoring functions are trained and tested over the same (or similar) set of target systems and active ligands.

One of the major limitations of the VDS approach is the additional computational effort required to generate and dock additional decoys for each library molecule. Although we utilized 100 decoy molecules for each query molecule, we have not established this as the appropriate size in all cases. Although not discussed in this paper, we retrospectively analyzed our target systems with varying sized decoy sets (data not shown) and observed that ranking results comparable to the ones demonstrated in this paper may be achieved using smaller sets. A possible approach to deal with such variation is to start with a modestly sized decoy set and gradually increase the size until the uncertainty in the distribution parameters is below some desired value. Additionally, redundancies in the overall set of decoys may be utilized to reduce computational effort. If some library molecules share similar physical properties, the same set of decoys may be used for both library molecules.

Finally, we note that while our method does not always improve the ranking, it seldom worsens it. In 9 of 13 cases the AUC computed using our method is within ±10% of the AUC computed using the naive approach. In one case our method has a AUC more than 10% lower than naive, and in three cases our method has an AUC more than 10% higher than naive (Table 1). These results suggest that while our method decreases the chance of failure (Supporting Information Figure 5), additional work is required.

## CONCLUSION

We introduced a method to improve the results of virtual docking by normalizing each ligand's docking score against the score distribution of automatically generated property-matched decoys. An interesting and welcome finding of our work is the result that the "naive" raw score of current docking algorithms is tough to beat. In 70% of the cases (9 of 13 systems) the naive approach performs well with or without normalization. In an additional 23% of the cases our normalization improves ranking performance and in only 7% of the cases (1 of 13) does normalization hurt performance. Furthermore, when screening libraries without controlling for physical variation, our normalization method shows great consistency compared to the naive ranking. This consistency is achieved because the normalization effectively minimizes the physical variation within the screening library and thereby cancels possible physical effects on scoring functions. We propose that our normalization method should become part of the normal docking protocol at least until scoring functions have advanced to the point where normalization is no longer necessary. Unfortunately, it is not always clear when normalization is necessary. In addition to presenting an approach to improve ranking and enrichment, we also suggest that virtual decoy sets can serve as a tool to develop, study, and improve scoring functions. For example, the VDS generation process provides great control over the physical and chemical property match between the library molecule and generated decoys. By varying a single physical or chemical property, one can quantify its effect on scoring. Many areas of science benefit from normalization. Although there has been significant progress in the accuracy of docking algorithms, normalization may provide at least a short-term patch to correct some of the underlying scoring biases.

## METHODS

**Generation of Data Sets.** We use a set of 13 DUD targets each containing at least 15 different active ligands as per Good's DUD clustering. For each target, our active set was the DUD actives and our inactive set was a sampling of the DUD inactive decoys. We selected 20 inactive DUD decoy compounds for each active. We used Tripos's EA-Inventor to generate a set of 100 property-matched virtual decoys (VDS) for all active and nonactive library molecules. The EA-Inventor uses a set of initial chemical building-blocks and 34 chemical operator (e.g., add/delete atom, form/break ring, mutate atom number/charge, and mutate bond type/stereochemistry) to produce a set of molecules that minimize the following fitness function:

$$
\begin{aligned}
F(VDS, m^r) = & \sum_{i=1}^{|VDS|} (s_{\mathrm{mw}}(m^{\mathrm{r}}_{\mathrm{mw}} - m^{i}_{\mathrm{mw}}))^4 \\
& + \sum_{i=1}^{|VDS|} (m^{\mathrm{r}}_{\mathrm{rot.}} - m^{i}_{\mathrm{rot.}})^4 \\
& + \sum_{i=1}^{|VDS|} (m^{\mathrm{r}}_{\mathrm{hbd}} - m^{i}_{\mathrm{hbd}})^4 \\
& + \sum_{i=1}^{|VDS|} (m^{r}_{\mathrm{hba}} - m^{i}_{\mathrm{hba}})^4 \\
& + \sum_{i=1}^{|VDS|} (m^{\mathrm{r}}_{clogp} - m^{i}_{clogp})^4 \\
& + \sum_{i=1}^{|VDS|} (m^{\mathrm{r}}_{\mathrm{chg}} - m^{i}_{\mathrm{chg}})^4 \\
& + \sum_{i=1}^{|VDS|} (s_{\mathrm{tpsa}}(m^{\mathrm{r}}_{\mathrm{tpsa}} - m^{i}_{\mathrm{tpsa}}))^4
\end{aligned}
$$

where VDS = $\{m^1 ... m^{100}\}$ is the current population of property-matched decoys; $m^r$ is a reference molecule for which the property-matched decoys are generated; $m_{mw}$, $m_{rot}$, $m_{hbd}$, $m_{hba}$, $m_{clogp}$, $m_{chg}$, and $m_{tpsa}$ are the molecular weight, number of rotational bonds, hydrogen bond donors, hydrogen bond acceptors, calculated log $P$, total charge, and topological polar surface area. $s_{mw}$ and $s_{tpsa}$ are constants equal to 0.025 and 0.05, respectively. To enforce the generation of chemically diverse sets, molecules having similarity greater than 0.8 (or 0.5) Tanimoto coefficient (MACCS fingerprints) were penalized according to EA-Inventor's internal fitness function.

*Randomly Selected Druglike Libraries.* The library with an average molecular weight of ~250 Da was generated by randomly choosing 400 compounds from the ZINC clean-drug-like set (set no. 13) with 0.8 similarity threshold (CACTVS type 2 fingerprints). Because ZINC's subsets of druglike compounds are limited to molecules with mass less than 500 Da, the library with an average molecular weight of ~450 Da was randomly selected from the all-purchasable set (set no. 6) while using the same filters of drug-likeliness as used in the druglike set but with molecular weight between 450 and 650 Da. Redundant compounds were removed using 0.8 similarity threshold (MACCS fingerprints).

**Docking.** We used two docking packages, Glide (version 201007) and eHiTS (version 2009.1).

*Receptor Preparation.* eHiTS receptors were prepared using the Dock Prep utility in Chimera[19] with default parameters (delete solvent, add hydrogens, add charges, and replace incomplete side-chains using Dunbrack rotamer library[20]). Glide receptors were prepared using prepwizard (http://www.schrodinger.com, June 13, 2011) with the OPLS2005 force field and the "-mse -s" flags in addition to the default parameters.

*Ligand Preparation.* The ligand preparation process for both Glide and eHiTS starts with the following two steps. First, ligands (nongenerated ones) were converted to SMILES strings using OpenBabel.[21] Second, we used ChemAxon's MolConvert (http://www.chemaxon.com, June 13, 2011) to add hydrogens, to convert compounds back to 3D, and to perform energy minimization. Then, compounds used with eHiTS had their charge states determined (at pH 7.0) using OpenBabel. Compounds used with Glide were processed with ligprep to determine the charge state (at pH 7.0) and to perform a brief energy minimization using the OPLS2005 forcefield.

*Docking.* Docking was performed with the fast mode ("1") for eHiTS and "SP" mode for Glide using a clipping box of 10 Å.

**Fitting and Normalization.** *Mixture Model of Extreme-Value and Uniform Distributions.* Given a set of $n$ docking scores, $S = \{s_1, s_2, ..., s_n\}$, obtained for a set of $n$ decoys, we maximaize the following log-likelihood function:

$$LL(k, \lambda, \alpha_1, \alpha_2 | S) = \sum_{i=1}^{n} \log\left[\alpha_1 \frac{k}{\lambda}\left(\frac{s_i}{\lambda}\right)^{k-1} e^{-(s_i/\lambda)^k} + \alpha_2 \frac{1}{\max(S)}\right]$$

where $k > 0$ and $\lambda > 0$ are the shape and scale parameters of the Weibull distribution and $\alpha_1$ and $\alpha_2$, the proportion of the mixture, are both positive and sum to 1.

Then, given a corresponding reference score, $s_r$, the $p$-value is calculated as the following:

$$p = 1 - \left[\alpha_1\left(1 - e^{-(s_r/\lambda)^k}\right) + \alpha_2 \min\left(\frac{s_r}{\max(S)}, 1\right)\right]$$

*Empirical Cumulative Density Function.* Given a set of $n$ docking scores, $S = \{s_1, s_2, ..., s_n\}$, obtained for a set of $n$ decoys, the empirical cumulative distribution function (ECDF) is a step function that jumps by $1/n$ at each of the $n$ data points in $S$. Given a corresponding reference score, $s_r$, the $p$-value is calculated as the following:

$$p = 1 - \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $x_i$ is an indicator variable such that

$$x_i = \begin{cases} 1 & \text{if } s_i \le s_r \\ 0 & \text{if } s_i > s_r \end{cases}$$

All fitting and normalization procedures were implemented in the R programming environment (http://www.r-project.org, June 13, 2011).

**Kullback–Leibler Divergence.** The Kullback–Leibler divergence between two distributions, $P$ and $Q$ is defined as

$$P_{KL}(P, Q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)}$$

Because the KL is not symmetric and hence not a metric, we use the average KL over the two directional divergence measures:

$$P_{KL}(P, Q) = \frac{1}{2}\left[\int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} + \int_{-\infty}^{+\infty} q(x) \log \frac{q(x)}{p(x)}\right]$$

## ■ ASSOCIATED CONTENT

**Supporting Information.** Figures showing distribution comparison of seven physical properties, comparison of similarity distributions for decoys, docking results using a randomly selected library, and ROC curves and tables listing distances between distribution of docking scores of decoys using Glide and eHiTS. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

**Corresponding Author**

*E-mail: izharw@cs.toronto.edu (I.W.); lilien@cs.toronto.edu (R.L.).

## ■ ACKNOWLEDGMENT

We thank Simulated Biomolecular Systems (Toronto, Ontario) for providing access to their eHiTS High Throughput Screening software. R.L. is supported by a Discovery grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

## ■ REFERENCES

(1) Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for this Class of Proteins? *J. Chem. Inf. Model.* **2009**, *49*, 1568–1580.

(2) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(3) Tanimoto, T. T. An Elementary Mathematical Theory of Classification and Prediction. *IBM Internal Report*, New York, 1958.
(4) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(5) Wallach, I.; Lilien, R. Virtual Decoy Sets for Molecular Docking Benchmarks. *J Chem. Inf. Model.* **2011**, *51*, 196–202.
(6) Vigers, G. P.; Rizzi, J. P. Multiple Active Site Corrections for Docking and Virtual Screening. *J. Med. Chem.* **2004**, *47*, 80–89.
(7) Jacobsson, M.; Karlén, A. Ligand Bias of Scoring Functions in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 1334–1343.
(8) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolymers* **2003**, *68*, 76–90.
(9) Fukunishi, Y.; Mikami, Y.; Nakamura, H. Similarities among Receptor Pockets and among Compounds: Analysis and Application to in Silico Ligand Screening. *J. Mol. Graphics Modell.* **2005**, *24*, 34–45.
(10) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening. *J. Med. Chem.* **2011**, *54*, 1223–1232.
(11) Irwin, J. J.; Shoichet, B. K. ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(12) Brooijmans, N.; Cross, J. B.; Humblet, C. Biased Retrieval of Chemical Series in Receptor-Based Virtual Screening. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 1053–1062.
(13) Good, A. C.; Oprea, T. I. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.
(14) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.
(15) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: A New Fast, Exhaustive Flexible Ligand Docking System. *J. Mol. Graphics Modell.* **2007**, *26*, 198–212.
(16) James, A. C.; Weininger, D.; Delany, J. *Daylight Theory Manual-Daylight 4.71*; Daylight Chemical Information Systems: Laguna Niguel, CA, 2000.
(17) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.
(18) Mysinger, M. M.; Shoichet, B. K. Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561–1573.
(19) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.
(20) Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431–440.
(21) Guha, R.; Howard, M.; Hutchison, G.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. The Blue Obelisk-Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.