JOURNAL OF CHEMICAL INFORMATION AND MODELING

Cite This: J. Chem. Inf. Model. 2018, 58, 2343−2354

Article

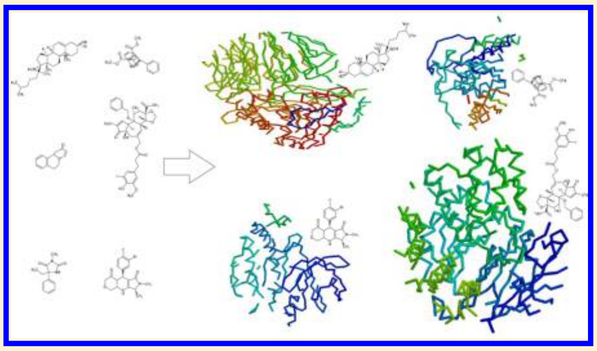
pubs.acs.org/jcim

FINDSITE$^{comb2.0}$: A New Approach for Virtual Ligand Screening of Proteins and Virtual Target Screening of Biomolecules

Hongyi Zhou, Hongnan Cao, and Jeffrey Skolnick*

Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, 950 Atlantic Drive, NW, Atlanta, Georgia 30332-2000, United States

Supporting Information



ABSTRACT: Computational approaches for predicting protein−ligand interactions can facilitate drug lead discovery and drug target determination. We have previously developed a threading/structural-based approach, FINDSITE^{comb}, for the virtual ligand screening of proteins that has been extensively experimentally validated. Even when low resolution predicted protein structures are employed, FINDSITE^{comb} has the advantage of being faster and more accurate than traditional high-resolution structure-based docking methods. It also overcomes the limitations of traditional QSAR methods that require a known set of seed ligands that bind to the given protein target. Here, we further improve FINDSITE^{comb} by enhancing its template ligand selection from the PDB/DrugBank/ChEMBL libraries of known protein−ligand interactions by (1) parsing the template proteins and their corresponding binding ligands in the DrugBank and ChEMBL libraries into domains so that the ligands with falsely matched domains to the targets will not be selected as template ligands; (2) applying various thresholds to filter out falsely matched template structures in the structure comparison process and thus their corresponding ligands for template ligand selection. With a sequence identity cutoff of 30% of target to templates and modeled target structures, FINDSITE$^{comb2.0}$ is shown to significantly improve upon FINDSITE^{comb} on the DUD-E benchmark set by increasing the 1% enrichment factor from 16.7 to 22.1, with a p-value of 4.3×10^{-3} by the Student t-test. With an 80% sequence identity cutoff of target to templates for the DUD-E set and modeled target structures, FINDSITE$^{comb2.0}$, having a 1% ROC enrichment factor of 52.39, also outperforms state-of-the-art methods that employ machine learning such as a deep convolutional neural network, CNN, with an enrichment of 29.65. Thus, FINDSITE$^{comb2.0}$ represents a significant improvement in the state-of-the-art. The FINDSITE$^{comb2.0}$ web service is freely available for academic users at http://pwp.gatech.edu/cssb/FINDSITE-COMB-2.

INTRODUCTION

The goal of drug discovery is to find effective and safe drugs that bind to a given protein target.[1] Since drugs often have serious side effects that result in FDA disapproval or drug withdrawal, the ability to derisk drugs could have an enormous economic impact on pharmaceutical companies.[2] To find a drug and understand its side effects or to avoid serious side effects, the first step is to identify all potential protein targets of the drug. Experimental discovery of drugs binding to a given protein target and all of their off-target human proteins is both costly and time-consuming.[3] Virtual ligand screening (VLS) is a computational tool for predicting protein−ligand interactions that has been widely employed in modern drug discovery for lead identification[4] and is more cost-effective than experimental high-throughput ligand screening for proteins or protein target screening for drugs.

To date, there are three broad categories of virtual ligand screening approaches: (1) high-resolution structure-based docking methods,[5−8] (2) ligand-based QSAR methods,[9−11] and (3) threading/structure-based methods.[12−19] The advantage of high-resolution structure-based docking approaches is that they employ physics principles and can potentially discover novel active binders. Their limitation and drawbacks are the availability of high-resolution structures, which currently cover around 1/3 of all human proteins,[20] their computational expense and lower accuracy than ligand-based methods.[19] Although ligand-based methods have better accuracy and are much faster than docking methods, they require a prior set of ligands known to bind to the protein target of interest. For the majority of human protein targets, such known sets of ligands are unavailable.[21]

Threading/structure-based methods address the limitations of the lack of availability of high-resolution structures required by docking approaches and the absence of known binders required by ligand-based approaches, while having comparable efficiency and accuracy as ligand-based approaches. Our recently developed FINDSITE^{comb} is a salient example of a threading/structure-based method that has been extensively experimentally validated.[19,22] FINDSITE^{comb} has also been

Received: May 21, 2018
Published: October 2, 2018

DOI: 10.1021/acs.jcim.8b00309

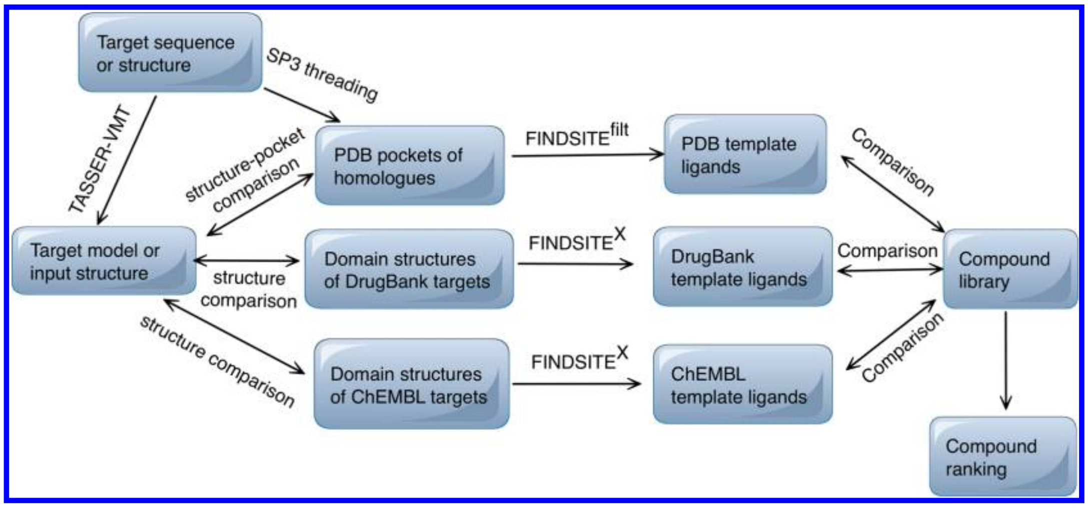


Figure 1. Flowchart of FINDSITE$^{comb2.0}$.

demonstrated to outperform state-of-the-art docking methods such as DOCK6[5] and AUTODOCK-vina[8] and shows comparable accuracy when low resolution predicted or experimental structures are used.

Recently, the emerging new convolutional neural network (CNN) or deep CNN[23,24] technology has been applied to virtual ligand screening.[23,24] The new protein−ligand scoring with CNN boosts the accuracy of high-resolution structure-based docking methods to the level of ligand-based and threading/structure-based approaches.

Methods for ligand virtual screening can, in principle, be employed for virtual target screening (VTS) where the goal is to predict for a ligand/small molecule, all of its possible protein targets in a given exome.[20,25] For traditional structure-based docking, virtual target screening requires high-resolution structures of all proteins in a given exome. At present, this is not feasible even if computational cost were not an issue, which it is.[20] For ligand-based approaches, a prior set of known binding ligands for all proteins in a exome is also not known. However, both issues disappear in the threading/structure-based method FINDSITE^{comb}. In a recent work,[26] we applied FINDSITE^{comb} to predict possible human targets of all FDA approved drugs as well as experimental drugs from the DrugBank database.[27] FINDSITE^{comb} covers 97% of human proteins (i.e., there are modeled structures for at least one domain of 97% of the sequences in the human proteome). The predicted drug targets allowed us to infer drug side effects and to repurpose FDA approved drugs to treat rare diseases.[26]

To predict the interaction of a given protein target to a ligand, FINDSITE^{comb} starts with the target protein sequence and then uses the SP3[28] threading method to find its homologous structures in the Protein Data Bank (PDB).[29] If there is no user provided target protein structure, FINDSITE^{comb} predicts one using TASSER-VMT[30] from threading identified PDB template protein structures (we call the proteins in the structural database "template proteins"). Subsequently, to find potential binders ("template ligands"), the target structure or model is compared to (a) ligand binding pockets from the subset of its homologous PDB template structures having ligand complexes from the PDB and (b) precomputed template modeled holo protein models of the proteins in the ChEMBL[31] and DrugBank[27] databases to find template ligands for the protein of interest. Then, using a similarity measure defined from a Tanimoto coefficient based fingerprint (mTC),[32] the ligand in a screened library is compared to each set of template ligands from the PDB, ChEMBL, and DrugBank. This mTC similarity score measures the likelihood that a given ligand binds to a particular protein.

In practice, the accuracy of FINDSITE^{comb} strongly depends on the accuracy of the selected template ligands that serve the same role as known binders for the protein target in ligand-based approaches. Here, to improve performance, we focus on the elimination of falsely selected, (*viz.,* false positive) template ligands. In the current work, we demonstrate that the improved version, FINDSITE$^{comb2.0}$, is significantly better than the original version and outperforms the state-of-the-art deep learning CNN methods.[23,24] For the 102 DUD-E benchmark set[33] and using predicted protein structures whose allowed templates have a sequence identity to the target protein <30%, the average area under the receiver operating characteristic (ROC) curve (AUC) and the top 1% enrichment factor improves from 0.745 and 16.74 for FINDSITE^{comb} to 0.784 and 22.06 for FINDSITE$^{comb2.0}$. On DUD-E, with the same 80% sequence identity cutoff of target to templates/training targets as used in the CNN scoring method,[23] the average AUC and ROC 1% enrichment factor of FINDSITE$^{comb2.0}$ using modeled target structures are 0.876 and 52.39, respectively, compared to 0.868 and 29.65 for the CNN scoring method[23] using experimental structures. Also with the same 80% sequence identity cutoff and modeled structures, FINDSITE$^{comb2.0}$ has 59/102 (57.8%) targets having an AUC greater than 0.9 compared to the same 59/102 (57.8%) targets by AtomNet that includes also the training set.[24] However, in a more realistic comparison to its 30 target testing set, AtomNet's number is reduced to 14/30 or 46.7%[24] whereas for a random subset of 30 targets, FINDSITE$^{comb2.0}$, has 16/30 (53.3%) targets having an AUC greater than 0.9. Using experimental structures and an 80% sequence identity cutoff for template ligand selection, FINDSITE$^{comb2.0}$ performs even better with an average AUC and ROC 1% enrichment factor of 0.886 and 56.27, respectively, and the number of targets having an AUC greater than 0.9 is 64 or 62.7%. Thus, by these measures, FINDSITE$^{comb2.0}$ is better than the most recent state-of-the-art alternative methods.

METHODS

Overview. The flowchart of FINDSITE$^{comb2.0}$ is shown in Figure 1. The differences between FINDSITE^{comb} and FINDSITE$^{comb2.0}$ are (1) how the protein−ligand libraries of DrugBank and ChEMBL are utilized and (2) the procedure for selecting template ligands. Given a protein target sequence or structure, both FINDSITE^{comb} and FINDSITE$^{comb2.0}$ employ the SP3 threading method[28] to find homologous PDB template structures. Ligand binding pockets are excised from a subset of these template structures having bound ligands (PDB pockets). If the input is just the protein's sequence, then TASSER-VMT[30] is used to build a structural model of the protein target. Then, template ligands are extracted from template proteins in the PDB/DrugBank/ChEMBL libraries that have similar pockets/structures to the target. These template proteins are obtained by comparing the modeled/input structure of the target to (a) the above excised PDB pockets, (b) precomputed structure models of the domains of the templates in DrugBank, and (c) precomputed structure models of the domains of the templates in ChEMBL. Template ligands from each of the three databases (PDB, DrugBank & ChEMBL) will be compared with each molecule in a screening compound library and a similarity score will be used to rank the molecules. The three rankings are combined by using the largest score of each compound among the three. Details concerning the selection of template ligands and improvement measures are provided next.

Template Ligands from the PDB. The procedure for selecting template ligands from the PDB and predicting binding sites is done by a program called FINDSITE^{filt}[19] that was previously developed. PDB structure templates identified by the SP3 threading algorithm[28] as having a threading Z-score (score in standard deviation units relative to the mean of the structure template library) > 4.0 are collected. The threading Z-score cutoff of 4.0 is an empirical value that ensures that there are a sufficient number of ligand bound templates (in practice, about 100) even for hard targets identified by the SP3 method as having a threading Z-score < 4.5. Then, the model or input structure of the target is compared to the ligand binding pockets from a subset of the PDB templates having bound ligands to determine the potential ligand binding sites and bound ligands of the target using a structure−pocket alignment algorithm. The details of the structure−pocket alignment algorithm can be found in the original FINDSITE^{comb} method.[19] A ligand binding pocket structure consists of the following: (1) the C_α atoms of the template residues, any of whose backbone and/or side chain heavy atoms are within 4.5 Å of the bound ligand's heavy atoms; (2) the C_α atoms of the template residues that are within 8 Å of the bound ligand's heavy atoms. These C_α atoms are usually scattered along the protein's sequence. We then apply the structure−pocket alignment algorithm to compare the target structure to each pocket and rank the pockets according to a similarity score.

In FINDSITE^{comb},[19] up to 100 ligands from top ranked pockets are selected as template ligands regardless of the alignment coverage of the target to the pockets. In this improved version, the number of template ligands will be optimized to reduce falsely selected ligands (ligands that are less likely to bind to the targets). To do so, we shall apply a filter to the ligand binding pockets. We require that the alignment length between the structure of the target and template's pocket must satisfy the following:

$$N_{ali}/N_{ali}^{max} > p_1 \text{ or } N_{ali}/N_{pocket} > p_2 \quad (1)$$

where N_{ali} is the alignment length of the pockets, N_{ali}^{max} is the maximal alignment length of all the pockets to the target pocket, N_{pocket} is the number of C_α atoms of the given pocket, and p_1, p_2 are two cutoffs. Here, empirical values $p_1 = 0.7$, $p_2 = 0.5$ are used.

Binding Site Prediction. To predict the binding sites of the target protein, we superimpose the above selected PDB template ligands onto the target structure using the rotational and translational matrix from the structure−pocket alignment. We then cluster the ligands according to their spatial proximity: (a) starting from the top ranked unclustered ligand L_i, going through all the other unclustered ligands L_j, if the distance of the center of mass of heavy atoms between L_i and L_j is less than 3 Å, then L_j is put into the cluster represented by L_i; (b) repeat process a until all ligands are clustered.

Next, we define a contact between a residue of the target protein and a ligand if the distance between any heavy atom of the residue and the ligand is less than r_{cut} + r_{vdW}(ligand atom) + r_{vdW}(residue atom), where r_{cut} is an empirical cutoff, and r_{vdW} is the van de Waals radius of the heavy atom obtained from the CHARMM force field parameters.[34] A residue of the target is defined as a binding site specific to a cluster of ligands if the residue is in contact with > p_{cut} fraction of ligands in that cluster. We optimized $p_{cut} = 0.34$, $r_{cut} = 0.7$ Å using the binding site prediction targets of the eighth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP8). Benchmarking carried out for the binding site prediction targets of CASP9 gives a Matthew's correlation coefficient (MCC) of 0.71 between the predicted and observed binding residues. This performance is indistinguishable from the best human prediction in CASP9.[35] The binding sites of each cluster of PDB template ligands define a predicted pocket of the target.

Parsing the Template Proteins into Domains and Clustering Their Ligands. While the PDB database has protein−ligand complex structures that allow us to determine the pocket a ligand binds to, the DrugBank and ChEMBL databases do not have the structures of the protein−ligand complex. In FINDSITE^{comb},[19] known binders of a protein target in the DrugBank and ChEMBL databases are assumed to bind to all domains of the protein because we did not have the specific binding position/pocket of the ligands. Based on this assumption, if an unknown target has a close similarity score to one domain of a library protein in the DrugBank or ChEMBL database, all ligands of that library protein will be transferred to the unknown target as template ligands, even though they might not actually bind to the domain that is similar to the unknown target. Obviously, this assumption will result in some falsely selected template ligands.

Another issue associated with the FINDSITE^{comb} approach is that some of the protein targets in the ChEMBL database have too many known ligands (exceeding a few thousand). If all of them are used for template ligands, they slow down the algorithm and often provide little added value, as their information is redundant. To increase efficiency, we cluster the ligands using the Tanimoto Coefficient (TC)[36] similarity measure defined by its Fingerprints.[37] Throughout this work, we use the FP2 fingerprint generated by the Open Babel (https://openbabel.org/wiki/Tutorial:Fingerprints) program with default options enabled. We then use a variable number of TC cutoffs, TC_{cut}, for clustering according to the total number of known ligands of the given target, N_{lig}:

$$TC_{cut} = \begin{cases} 0.95 \text{ if } N_{lig} < 200 \\ 0.85 \text{ if } 200 \leq N_{lig} \leq 1000 \\ 0.80 \text{ if } 1000 < N_{lig} \leq 2000 \\ 0.70 \text{ if } 2000 < N_{lig} \leq 5000 \\ 0.60 \text{ if } N_{lig} > 5000 \end{cases} \quad (2)$$

A greedy algorithm is employed to cluster the ligands: (1) pairwise TCs are calculated; (2) for each ligand, we count the number of neighboring ligands having > TC_{cut} to it; (3) starting from the ligand with largest number of neighbors, we use it as a cluster representative and its neighbors as members to define a cluster; (4) remove the ligands belonging to the cluster and repeat steps 1−3 until all ligands are assigned to a cluster.

To avoid false template ligands, we will partition the above known representative ligands of a protein in the DrugBank and ChEMBL databases into their possible binding domains. We shall use the same DrugBank and ChEMBL data sets as utilized in $FINDSITE^{comb}$.[19] The DrugBank database has 3833 protein targets and ChEMBL has 2453 protein targets. The overview of the ligand partitioning/target protein domain assignment procedure is shown in Figure 2. For each protein template in the DrugBank or ChEMBL database, we use PFAM[38] to partition its sequence into domains. 2407 out of the 3833 DrugBank templates and 2061 out of the 2453 ChEMBL templates have multiple domains. We then employ TASSER-VMT[30] to model each database protein domain. We then run $FINDSITE^{filt}$ on the each predicted ChEMBL or DrugBank domain structure to assign where the ChEMBL or DrugBank ligand binds on the basis of the maximal TC (called maxiTC) score between the to be assigned ligand and all the PDB template ligands of the given domain. In very rare cases, (e.g., among the total 13 004 ligand assignments for DrugBank, only 582 or 4.5% have multiple domain assignments) where the database protein target has multiple identical or very close domains, and the maxiTC scores are tied among these domains, then, the ligand will be assigned to all of the tied domains. In the end, for each database protein, we have structure models and the binding sites of its domains with each domain assigned a subset (cluster representatives) of the known binding ligands of the protein.

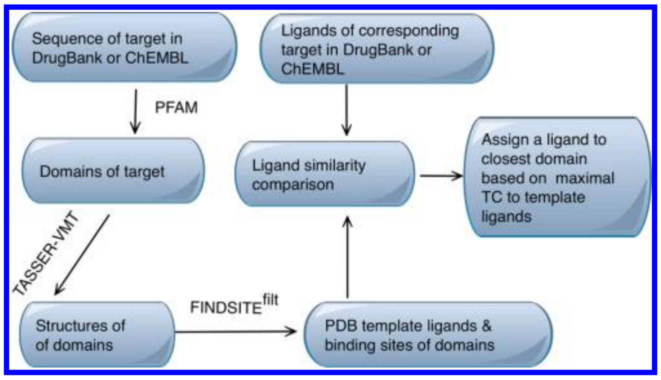

Sequence of target in DrugBank or ChEMBL
Ligands of corresponding target in DrugBank or ChEMBL
PFAM
Domains of target
Ligand similarity comparison
Assign a ligand to closest domain based on maximal TC to template ligands
TASSER-VMT
FINDSITE^filt
Structures of of domains
PDB template ligands & binding sites of domains

Figure 2. Overview of ligand partitioning/domain assignment process for proteins in the DrugBank and ChEMBL databases.

Template Ligands from DrugBank and ChEMBL. The structure model of an input protein target will be aligned to the precomputed model of each protein domain in a database (DrugBank or ChEMBL) using a modified version of Fr-TM-align.[18,39] First, Fr-TM-align is used to align the two structures, and then, instead of using the TM-score[40] that is purely based on a structural similarity measure of the two structures, we include a sequence similarity score based on evolutionary similarity. The output score is the summation of the BLOSUM62 substitution matrix[41] value over the aligned residues provided by Fr-TM-align and normalized by the target length. In other words, Fr-TM-align is used to build the equivalent sequence alignment, and then, BLOSUM62 is used to calculate the sequence alignment score (without gap penalties and normalized by target length). The alignment score will be used to rank database domains. We assume that the larger the score, the closer the domain's function to the target.

In the original $FINDSITE^{comb}$ approach, we used the ligands of the top first ranked protein target in the database as the template ligands of the input target, regardless of the closeness of structure similarity to its template. This simple implementation has some drawbacks. First, if the top first ranked template structure and the target structure are distant, their function might not be similar, and thus, the template ligands from the template structure likely do not bind to the target protein. Second, the target structure could have been aligned to a region of the template that does not bind to the selected ligands. Third, if two templates have very close similarity to the target (one of them is ranked first) but have very different known sets of ligands, a slight inaccuracy in the model of one template could give an unpredictable rank order switch and thus result in quite different template ligands. This gives an unstable result.

In this new version of $FINDSITE^{comb2.0}$, to improve the performance we made the following modifications to reduce falsely selected template domains, and consequently, template ligands:

(1) A global structure similarity threshold TM_{cut} measured by the TM-score[40] of the input target to the domains in DrugBank or ChEMBL is used. Structural alignment is done by Fr-TM-align.[39] This avoids the use of template ligands of likely unrelated structures.

(2) To ensure that the global alignment of the template domain to the target covers a reasonable part of the binding site of the template domain (functional part), we apply the following alignment overlap threshold: the structurally aligned region of the domain and the predicted binding region of the domain must have $>OV_{cut}$ overlap. This filter ensures that the target has structure similarity to the region of the template domain that binds to its known set of ligands.

(3) To increase tolerance to structure modeling inaccuracy, we select the top N_{dm} (≥ 1) rather than only top first ranked domains for template ligand inference.

The above thresholds of TM-score[40] TM_{cut}, overlap OV_{cut}, and top domains N_{dm} as well as the number of PDB pockets selected N_{pkt} for PDB template ligands will be optimized. The selected template ligands will be employed for virtual ligand screening of the input target using eq 3.

Ligand Similarity Comparison. Once template ligands are obtained by the above procedure, they are employed to

DOI: 10.1021/acs.jcim.8b00309
J. Chem. Inf. Model. 2018, 58, 2343−2354

search for actives of the input target in a compound library by the following similarity score:

$$\mathrm{mTC} = w\frac{\sum_{l=1}^{N_{\mathrm{lg}}} \mathrm{TC}(L_l, L_{\mathrm{lib}})}{N_{\mathrm{lg}}} + (1 - w) \max_{l\in(1,\ldots,N_{\mathrm{lg}})} (\mathrm{TC}(L_l, L_{\mathrm{lib}})) \tag{3}$$

where TC is the Tanimoto Coefficient;[36] N_{lg} is the number of template ligands from the putative evolutionarily related proteins; L_l and L_{lib} stand for the template ligand and the ligand in the compound library, respectively; w is a weight parameter. $w = 1$ gives the average TC in the original FINDSITE screening score.[16] The second term is the maximal TC between a given compound and all the template ligands. Here, we empirically choose $w = 0.1$ to give more weight to the second term so that when the template ligands are true ligands of the target, they will be favored.

Assessment Criteria. To assess the virtual ligand screening results, we employed the AUC (the area under the ROC curve) and the enrichment factor of the top x% of molecules defined by

$$\mathrm{EF}_x = \frac{\text{number of actives in the top } x\%}{xN_{\mathrm{active}}/100} \tag{4}$$

where N_{active} is the total number of actives in the entire screened molecule set. Ideally, if all actives are within the top 1% of screened molecules, EF_1 should be 100. However, since for the DUD-E set,[33] the ratio of decoys/actives is on average ∼60, the number of actives is greater than 1% of the total number of molecules. Thus, the top 1% of screened molecules can only accommodate 60% of the actives. As a consequence, the maximal possible EF_1 will be only, on average, ∼60. Other measures we employ are the precision and coverage:

$$\text{precision} = \frac{\text{number of actives with mTC} \geq \text{cutoff or within } \Delta\text{mTC}}{\text{total number of molecules with mTC} \geq \text{cutoff or within } \Delta\text{mTC}} \tag{5a}$$

$$\text{coverage} = \frac{\text{number of actives with mTC} \geq \text{cutoff}}{\text{total number of actives}} \tag{5b}$$

Optimization of Thresholds. We performed a grid-based optimization of the four parameters on the DUD-E set[33] with a 30% sequence identity cutoff. The grid search spaces are (0.2, 0.4, 0.6, 0.8) for $\mathrm{TM}_{\mathrm{cut}}$, (20%, 40%, 60%, 80%) for the overlap cutoff $\mathrm{OV}_{\mathrm{cut}}$, (1, 5, 10, 20, 30) for N_{dm}, and (50, 100, 200, 300) for N_{pkt}, respectively. The parameters that give the best average AUC of 0.7878 are ($\mathrm{TM}_{\mathrm{cut}}$, $\mathrm{OV}_{\mathrm{cut}}$, N_{dm}, N_{pkt}) = (0.6, 20%, 20, 100). The parameters for best average EF_1 of 22.905 are (0.6, 80%, 20, 100) that will be used as the default for future applications. It is interesting to note that the parameters that have the worst average AUC of 0.7456 and worst EF_1 of 13.70 are (0.8, 40%, 1, 300).

Experimental Materials and Methods. Human erythrocyte carbonic anhydrase 1 hCA1 and porcine heart malate dehydrogenase pMDH were obtained from Sigma, USA. *Pseudomonas aeruginosa* aminoglycoside *O*-phosphotransferase APH(3″)-Ib (gb|ABK33456.1|) with an N-terminal His-tag followed by TEV protease cleavable linker was overexpressed in *E. coli* and purified using standard nickel affinity chromatography. The detailed experimental procedures are documented in the Supporting Information.

Thermofluor or thermal shift assays were performed using previously reported protocols.[22,42] Briefly, thermal denaturation of the target protein was performed in 96-well plates using a RealPlex quantitative PCR instrument from Eppendorf (Eppendorf, NY, USA). The system was precalibrated before each run. A 20 μL portion of protein sample, at a fixed concentration in the range of 1−5 μM, was aliquoted into the bottom of the wells, incubated with the compounds of interest for 10−30 min at room temperature and mixed with SYPRO orange (serially diluted to a final concentration of 5× from a 5000× stock solution). The same reaction buffer containing 50 mM HEPES pH 7.3, and 100 mM NaCl was used to screen different compounds at a final concentration of 0.5 mM unless otherwise stated. A heating ramp of 1 °C/min from 20 to 95 °C was used, and one data point was acquired for each degree increment. The excitation and emission wavelengths were 465 and 580 nm, respectively. The compound/buffer control melting curve was subtracted from the protein containing samples. In this study, the compounds themselves typically yielded minimal fluorescence signal compared to the protein's melting curve. The negative first derivative of relative fluorescence in arbitrary units, −d(FAU)/dt, was plotted against T with Excel software to determine the melting temperature T_{m}. Only one negative peak in the plot was observed for each sample, indicating a concerted melting of the enzyme following a quasi-two-state model. The thermal shift $\Delta T_{\mathrm{m}} = T_{\mathrm{m}}(\text{protein with drug}) - T_{\mathrm{m}}(\text{protein})$. Each condition was run at least in duplicate and gave a standard deviation <1 °C, unless otherwise stated. While a $\Delta T_{\mathrm{m}} > 1$ °C (considering experimental errors) indicates binding, an appropriate estimation of the absolute thermal dynamics properties of traditional binding affinity K_{d} requires a "titration" experiment, i.e., the measurement and fitting of a set of T_{m}s from multiple melting curves under conditions of at least 2−3 different ligand concentrations.[22,43]

Web Service of FINDSITE$^{\mathrm{comb2.0}}$ for VLS and VTS. We have implemented an online web service of FINDSITE$^{\mathrm{comb2.0}}$ for VLS and VTS that is freely available for academic users at http://pwp.gatech.edu/cssb/FINDSITE-COMB-2. For VLS, the user inputs a protein sequence or PDB structure and selects a screening library of compounds from DrugBank, or the NCI diversity set or user selected molecules in a SMILES string,[44] mol2 or sdf format. The output will be the rank order of the screened molecules along with their predicted binding precision. For VTS, the user inputs **single** molecule in a SMILES string,[44] mol2 or sdf format. Currently, we only screen the molecule against the human exome. The output will be the rank order of all human proteins (including isoforms) that are predicted to bind the molecule. The precision of the binding prediction is also provided. Since the computation is not instantaneous, the user needs to supply a valid email address. Once screening is completed, the user will receive an email and a link to download the results or the results will be attached to the email.

■ RESULTS

Comparison to FINDSITE$^{\mathrm{comb}}$ for Virtual Ligand Screening. We tested FINDSITE$^{\mathrm{comb2.0}}$ on the benchmarking set of the Directory of Useful Decoys, Enhanced (DUD-E) for virtual ligand screening[33] and compared it to FINDSITE$^{\mathrm{comb}}$. Since FINDSITE$^{\mathrm{comb2.0}}$ has four free parameters to choose, in order to have a fair comparison, we employ a jack-knife test. For each testing target, we optimized its parameters on the

DUD-E set by excluding the testing target and use EF_1 as the objective function. DUD-E has 102 targets. On average, the ratio of the number of decoys to actives is around 60, with the average numbers of actives of ~224 and decoys of 13 696.

In practice, we merge the screening results of all targets and calculate a single precision and coverage for the whole data set, whereas AUC and EF_x are calculated for each target. In Table 1, we compare the performance of FINDSITE$^{comb2.0}$ to FINDSITE^{comb} with a sequence identity cutoff 30%, i.e. no template with known binders from the PDB, DrugBank, or ChEMBL having a sequence identity >30% to the input target are used for template ligand inference. We compare these approaches using both modeled structures generated using TASSER-VMT with a 30% sequence identity cutoff for modeling templates and experimental structures. For modeled structures, FINDSITE$^{comb2.0}$ has a mean AUC and a mean top 1% enrichment factor of 0.785 and 22.06, respectively, as compared to 0.745 and 16.74 for FINDSITE^{comb}. FINDSITE$^{comb2.0}$ has 88 targets with better than random selection (EF_1 > 1) compared to 81 for FINDSITE^{comb}. FINDSITE$^{comb2.0}$ is shown to significantly improve FINDSITE^{comb} on the DUD-E benchmark set with a p-value of 6.0×10^{-4} for AUC and 4.3×10^{-3} for a 1% enrichment factor by the Student t-test. With an mTC cutoff of 0.9, FINDSITE$^{comb2.0}$ has a prediction precision of 83.8% whereas FINDSITE^{comb} has a prediction precision of 75.0%. In Figure 3, using modeled structures, we show the scatter plot of EF_1 for the two compared methods. FINDSITE$^{comb2.0}$ has 59 targets having a larger EF_1 than FINDSITE^{comb}, whereas FINDSITE^{comb} has only 28 targets with a larger EF_1 than FINDSITE$^{comb2.0}$. Using experimental structures, both methods show around an ~15% increase in EF_1, but their relative performance does not change.

Table 1. Comparison between FINDSITE$^{comb2.0}$ and FINDSITE^{comb} on the DUD-E Set

	AUC[a]	EF_1[b]	precision (mTC ≥ 0.9)	coverage (mTC ≥ 0.9)
	Modeled Target Structures			
FINDSITE^{comb}	0.745 (21)	16.74 (81)	75.0%	3.35%
FINDSITE$^{comb2.0}$	0.785 (30)	22.06 (88)	83.8%	6.01%
FINDSITE$^{comb2.0}$-nofilter	0.755(24)	16.76(89)	71.7%	3.06%
	Experimental Target Structures			
FINDSITE^{comb}	0.779 (26)	19.15 (90)	74.6%	3.35%
FINDSITE$^{comb2.0}$	0.811 (36)	25.21 (93)	83.3%	6.27%
FINDSITE$^{comb2.0}$-nofilter	0.779(30)	19.73(91)	75.1%	3.35%

[a] Numbers in parentheses are the number of targets having an AUC > 0.9. [b] Numbers in parentheses are the number of targets having an EF_1 > 1.

To examine the effects of the two major improvements of FINDSITE$^{comb2.0}$ over FINDSITE^{comb}: (1) parsing library templates into domains and (2) applying filters to template ligand selection, we tested an alternative approach called FINDSITE$^{comb2.0}$-nofilter that does not apply filters to template ligand selection. With modeled target structures, although on average FINDSITE$^{comb2.0}$-nofilter has only slightly better AUC and EF_1 and slightly worse precision and coverage than those of FINDSITE^{comb}, it has more targets with better than random ligand selection (EF_1 > 1) and an AUC > 0.9. FINDSITE$^{comb2.0}$-nofilter has a better EF_1 than FINDSITE^{comb} for 48 targets, whereas FINDSITE^{comb} is better for 32 targets. With experimental target structures, FINDSITE$^{comb2.0}$-nofilter still performs slightly better than FINDSITE^{comb}, but it is significantly worse than FINDSITE$^{comb2.0}$. We also note that by using only the PDB library for template ligands and modeled target structure, FINDSITE$^{comb2.0}$ (FINDSITE$^{filt2.0}$) has an average EF_1 of 14.20 compared to 22.06 of full FINDSITE$^{comb2.0}$. The number of targets having an EF_1 > 1 by FINDSITE$^{filt2.0}$ is 76 whereas by FINDSITE$^{comb2.0}$, it is 88. This indicates that adding the templates from the DrugBank and ChEMBL libraries significantly improves the performance of FINDSITE$^{comb2.0}$.

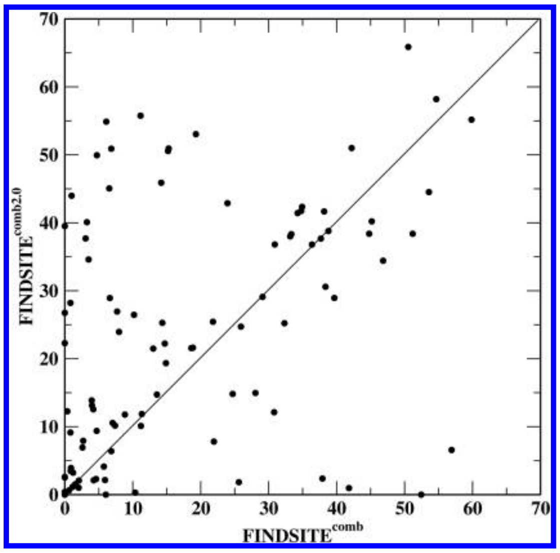


Figure 3. For modeled target protein structures, a scatter plot of EF_1 between FINDSITE^{comb} and FINDSITE$^{comb2.0}$.

Here, we show two examples of improved targets. For target *Ada*, using modeled structures, FINDSITE$^{comb2.0}$ has an EF_1 of 40.09 whereas FINDSITE^{comb} only has an EF_1 of 3.25, while FINDSITE$^{comb2.0}$-nofilter has an EF_1 of 7.58. This indicates that the domain parsing of templates in the DrugBank and ChEMBL databases improves its EF_1 from 3.25 to 7.58. After implementing FINDSITE$^{comb2.0}$ that applies filters and increases the number of templates for the DrugBank and ChEMBL libraries, the EF_1 greatly improves to 40.09. The performance increase in this case is due to the elimination of false positive ligand templates. Specifically, the number of template ligands in the PDB library is reduced from 100 to 10, in DrugBank from 1 to 0, and in ChEMBL from 104 to 0, respectively.

Another example is target *rock1*. FINDSITE^{comb} has an EF_1 of 0.9995, while FINDSITE$^{comb2.0}$-nofilter increases the EF_1 to 2.999. More significantly, FINDSITE$^{comb2.0}$ further improves EF_1 to 43.98. In this case, FINDSITE$^{comb2.0}$ does not filter out any PDB template ligands. Besides *protein kinase C iota type* that is selected by FINDSITE^{comb}, FINDSITE$^{comb2.0}$ selects these additional template targets for template ligands from DrugBank: *Beta-adrenergic receptor kinase 2; Beta-adrenergic receptor kinase 1; 3-phosphoinositide-dependent protein kinase 1; MAP kinase-activated protein kinase 2; Serine/threonine-protein kinase 12; Calcium/calmodulin-dependent protein kinase type IV; Calcium/calmodulin-dependent protein kinase type 1D; Serine/threonine-protein kinase 17B.* FINDSITE$^{comb2.0}$ also selects 9 more protein templates from the ChEMBL library than FINDSITE^{comb}.

Comparison to Deep CNN Methods for Virtual Ligand Screening. Recently, two groups[23,24] have developed virtual ligand screening approaches using state-of-the-art deep convolutional neural networks.[45] Though they still require high resolution protein structures, their performance in virtual ligand screening is much better than traditional docking methods such as AUTODOCK-vina.[8] Thus, it is interesting to know how FINDSITE$^{\text{comb2.0}}$ compares to these deep CNN methods. AtomNet[24] used randomly selected 72 DUD-E targets for training and the remaining 30 targets for testing. The CNN scoring method[23] clustered the 102 DUD-E targets with an 80% sequence identity threshold and divided the clusters into three sets to do three-fold cross-validation tests. For fair comparison, we apply an 80% sequence cutoff to templates for template ligand selection: i.e. ligands from any template in the PDB, DrugBank and ChEMBL having a sequence identity >80% will be ignored. Furthermore, for each testing target, we only optimize its four free parameters on the DUD-E targets having sequence identity <80% to it. We tested FINDSITE$^{\text{comb2.0}}$ using both modeled structures (with 30% sequence identity cutoff) and experimental structures. The CNN scoring approach[23] used the ROC enrichment factor (ROC-EF) to assess their results. The ROC-EF is slightly different from eq 4 and is defined as

$$\text{ROC-EF}_x = \frac{\text{number of actives at } x\text{\% false positive rate}}{\text{number of decoys at } x\text{\% false positive rate}} \tag{6}$$

Table 2 compares the performance of FINDSITE$^{\text{comb2.0}}$ to deep CNN methods along with FINDSITE$^{\text{comb}}$ and the traditional docking method AUTODOCK-vina[8] on the DUD-E set. AtomNet has the best mean AUC of 0.895 only when it includes the training targets. Even though the deep CNN scoring method has a better AUC than FINDSITE$^{\text{comb}}$, it has a worse ROC-EF$_1$. In practice, a better ROC-EF$_1$ or EF$_1$ is more useful because one only needs to test the top few percent of the ranked ligands. FINDSITE$^{\text{comb2.0}}$ with modeled structures has a better mean AUC of 0.876 and a better mean ROC-EF$_1$ of 52.39 as compared to 0.868 and 29.65 of the CNN scoring method.[23] FINDSITE$^{\text{comb2.0}}$ also has more targets, 59(57.8%), with an AUC > 0.9 than the 49(48.0%) of the CNN scoring method.[23] For a randomly selected 30 target subset, FINDSITE$^{\text{comb2.0}}$ has a mean AUC of 0.880 and 16(53.3%) targets having an AUC > 0.9 whereas AtomNet has 0.855 and 14(46.7%), respectively.[24] The traditional docking method AUTODOCK-vina has the worst performance by all measures. With experimental structures, FINDSITE$^{\text{comb2.0}}$ performs even better, with a mean AUC of 0.886, ROC-EF$_1$ of 56.27, and 64 or 62.7% of targets having an AUC > 0.9. The ROC-EF$_1$ difference between modeled and experimental structures is less than 10%. Thus, FINDSITE$^{\text{comb2.0}}$ outperforms the state-of-the-art deep learning based methods and offers the further advantage that it can use predicted as well as experimental structures.

Table 2. Comparison of FINDSITE$^{\text{comb2.0}}$ with CNN Methods on the DUD-E Set

	AUC	no. of AUC > 0.9 (%)	ROC-EF$_1$
AtomNet[24] (DUD-E-30)	0.855	14 (46.7%)	
AtomNet[24] (DUD-E-102)[a]	0.895	59 (57.8%)	
CNN scoring[23]	0.868	49 (48.0%)	29.65
AUTODOCK-vina[23]	0.716	2 (2.0%)	7.32
FINDSITE$^{\text{comb}}$	0.829	39 (38.2%)	37.26
FINDSITE$^{\text{comb2.0}}$ (modeled structures)	0.876	59 (57.8%)	52.39
FINDSITE$^{\text{comb2.0}}$ (modeled structures, DUD-E-30)[b]	0.880	16 (53.3%)	53.55
FINDSITE$^{\text{comb2.0}}$ (experimental structures)	0.886	64 (62.7%)	56.27

[a]Includes results for the 72 training protein targets. [b]30 random targets from DUD-E set.

Comparison to Ligand-Based Methods for Virtual Ligand Screening. FINDSITE$^{\text{comb2.0}}$ distinguishes itself from ligand-based QSAR methods by not using any information about known binding ligands to the target protein. The only requirement from the target protein is the amino acid sequence. Since the DUD-E data set favors ligand-based methods,[46] we shall employ an unbiased data set called ULS/UDS developed in ref 46 to compare FINDSITE$^{\text{comb2.0}}$ to ligand-based methods. The ULS/UDS (Unbiased Ligand Set/Unbiased Decoy Set) set was designed not to bias toward ligand-based virtual screening methods. It has 17 GPCR targets with an average ~600 decoys and ~15 ligands. With a sequence identity cutoff of 80% (i.e., information from any templates having sequence identity >80% to the target will not be used in ligand homology modeling), FINDSITE$^{\text{comb2.0}}$ achieves an average 0.862 of AUC of the ROC curve in comparison to 0.675 by the FCFP_6 fingerprint based method.[46] Even with a much stricter sequence identity cutoff of 30%, FINDSITE$^{\text{comb2.0}}$ still has an average AUC of 0.713. Thus, for an unbiased data set like ULS/UDS, we conclude that FINDSITE$^{\text{comb2.0}}$ performs better than ligand-based methods. Results for individual targets can be found in Table S2 in the Supporting Information.

Comparison to Docking-Based Method for Ligand Diversity. A good performing ligand virtual screening method should not only have a good enrichment factor, but to increase the chance of finding new classes of ligand hits, it should also show significant diversity of the top ranked ligands. Since FINDSITE$^{\text{comb2.0}}$ uses a similarity search as its last step, it will be informative to know if its recalled ligands within the top 1% of ranked molecules are more diverse than traditional docking methods. Hence, we compare FINDSITE$^{\text{comb2.0}}$ to AUTODOCK-vina[8] on the DUD-E data set. We cluster the active ligands within the top 1% of ranked screened molecules using a TC cutoff of 0.8. The average number of clusters per protein target of all ligands is 123.9. AUTODOCK-vina recalls on average 13.01 clusters within the top 1% whereas FINDSITE$^{\text{comb2.0}}$ recalls on average 48.98 clusters with an 80% sequence identity cutoff. Thus, in terms of recalled ligand diversity, FINDSITE$^{\text{comb2.0}}$ is much better than docking methods such as AUTODOCK-vina.

Comparison to FINDSITE$^{\text{comb}}$ for Virtual Target Screening. In our earlier work, we created a database and service that provides predicted protein targets of drugs, drugs for a given protein target, and associated diseases and side effects of drugs, called DR. PRODIS.[26] There we applied FINDSITE$^{\text{comb}}$ to predict the binding targets and possible side effects of all small molecules in DrugBank.[26] In contrast to virtual ligand screening, virtual target screening screens a given molecule against the entire human genome. We have modeled and generated data for virtual target screening for 97% of all human protein sequences (including isoforms).[26] Within 97% of modeled sequences, 85.6% of sequences have at least one domain with a predicted TM-score to native >0.4. We have shown in FINDSITE^{comb} that, on average, a target will have better than random VLS results if its predicted TM-score to native >0.4. Those modeled targets are ready to be used by FINDSITE$^{comb2.0}$ for virtual target screening. To the best of our knowledge, this is the first academic approach that could screen almost the entire human genome for drug target discovery. Other sequence-based methods for drug−target interaction prediction rely on knowledge of existing drug−target interactions since their inference-based approaches require known drug−drug and/or protein−protein similarity.[47−52] In addition, methods that require high resolution protein structures are limited by the availability of such structures and have a significant computational expense.[20]

To test FINDSITE$^{comb2.0}$ in the context of virtual target screening, i.e. predicting the human targets of a given compound, we compiled a set of 540 small molecule drugs from DrugBank version 5.09[27] that are not present in the earlier version 3.0 employed by FINDSITE$^{comb2.0}$ and FINDSITE^{comb} as the knowledge library. The results as assessed by the AUC and top 1% enrichment factor are compiled in Table 3. FINDSITE$^{comb2.0}$ achieves an average AUC = 0.881 and EF_1=59.96 compared to an AUC = 0.851 and EF_1=52.34, respectively, for FINDSITE^{comb}. Thus, FINDSITE$^{comb2.0}$ has a significant improvement over FINDSITE^{comb} for virtual target screening as indicated by a p-value of 4.2×10^{-6} and 1.7×10^{-8} with a Student t test for AUC and EF_1, respectively. Again, we examined how FINDSITE$^{comb2.0}$-nofilter performs when only domain parsing of the DrugBank and ChEMBL libraries are employed. FINDSITE$^{comb2.0}$-nofilter has AUC of 0.861 and EF_1 of 55.92, both having obvious improvement over FINDSITE^{comb}.

Table 3. Comparison between FINDSITE$^{comb2.0}$ and FINDSITE^{comb} for Drug Target Prediction

	AUC	EF_1
FINDSITE^{comb}	0.851	52.34
FINDSITE$^{comb2.0}$	0.881	59.96
FINDSITE$^{comb2.0}$ (53 drugs)[a]	0.866	59.88
FINDSITE$^{comb2.0}$I-nofilter	0.861	55.92
PROBE$_x$[b]	0.81	31.4

[a] Randomly selected from the 540 protein set. [b] Tested on 53 (whose identity is not available from the literature) drugs.

Although there are no similar academic methods published, we did notice a commercially available method called PROBE$_x$ by CYCLICA (https://static1.squarespace.com/static/54b9178ae4b09cb81d821314/t/5872da0fc534a5d5acbf55d5/1483921941137/Cyclica_ValidationNote_ROKT.pdf). PROBE$_x$ was tested on 53 drugs from DrugBank for target prediction with result of AUC = 0.81 and EF_1=31.4. Since we do not know which 53 drugs were assessed in their study, we randomly evaluate 53 drugs from among our 540 drugs. FINDSITE$^{comb2.0}$ has AUC = 0.866 and EF_1=59.88 for this 53 drug set. Thus, FINDSITE$^{comb2.0}$ performs significantly better than PROBE$_x$.

Experimental Validation of FINDSITE$^{comb2.0}$ Predictions. Using the DUD-E set, we examine the dependence of the precision by FINDSITE$^{comb2.0}$ on the mTC score cutoff in Figure 4. Under the very stringent condition that no template has a sequence identity >30% to the input target in the PDB, DrugBank and ChEMBL libraries, with an mTC cutoff of 0.9, the precision is 84%, whereas with an mTC cutoff of 0.8, it drops rapidly to 41%.

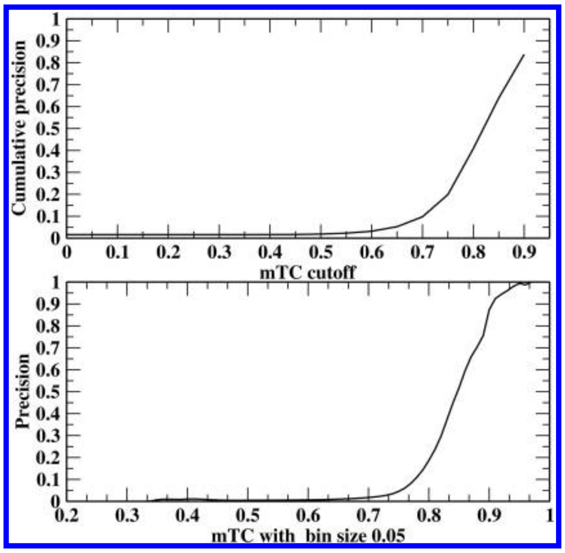


Figure 4. Dependence of protein−ligand interaction predicted cumulative precision based on the mTC cutoff (upper) and the precision with a given mTC within a bin size of 0.05 (lower) by FINDSITE$^{comb2.0}$ with a 30% sequence identity cutoff and modeled target structures. Data are derived from the 102 target DUD-E set.

Next, we experimentally tested three protein targets to assess the predicted benchmark precision of FINDSITE$^{comb2.0}$. The detailed experimental procedures are summarized in the Supporting Information. In practice, we applied FINDSITE$^{comb2.0}$ to ligand binding prediction for three protein targets−human erythrocyte carbonic anhydrase 1 hCA1 (Sigma, USA), porcine heart malate dehydrogenase pMDH (Sigma, USA), and *Pseudomonas aeruginosa* aminoglycoside *O*-phosphotransferase APH(3″)-Ib (in-house overexpressed and purified from *E. coli*) to screen against the molecules from the National Cancer Institute (NCI) diversity set (https://dtp.cancer.gov/organization/dscb/obtaining/available_plates.htm). A sequence identity cutoff of 30% for protein templates was applied for pMDH (sp|P11708|) and APH(3″)-Ib (WP|084929469.1|) and a control test with no cutoff was applied to hCA1 (sp|P00915|). The reason for using NCI molecules is that they are easy to obtain (http://dtp.nci.nih.gov/branches/dscb/repo_open.html). The NCI diversity set consists of 1597 molecules from Diversity Set III, 97 molecules from the Approved Oncology Drugs Set IV, and 118 molecules from the Natural Product Set II (total 1812 NCI molecules). Molecules with an mTC > 0.8 are selected for Thermofluor assay, a common and sensitive fluorescence-based thermal shift experimental test of ligand binding to the protein target, using our previously reported protocols.[20,48] We found that the experimentally observed precision (overall hit rate of true binders defined as compounds displaying a thermal shift over 1 °C at 0.5 mM final concentration) was generally consistent with the average expected precision of FINDSITE$^{comb2.0}$ predictions for the DUD-E benchmark set (see Table 4).

As shown in Table 4, for pMDH the observed and expected precision values are 0.5 and 0.76 respectively. Whereas for APH(3″)-Ib the observed and expected precision are 1.0/0.7 respectively. In particular, we identified the novel binding of the FDA-approved anticancer drug Sunitinib (Drug Bank ID: DB01268) to pMDH with an estimated K_d of ~5.4 μM based

Table 4. List of Thermal Shift ΔT_m Values versus mTC Scores of Different Proteins and Drugs[a]

pMDH			
drug	ΔT_m (°C)	mTC	precision
Imatinib	0	0.930	0.961
Sorafenib	**2**[c]	0.926	0.961
Sunitinib	**10**[c]	0.925	0.961
Lapatinib	nd	0.925	0.961
Dasatinib	nd	0.924	0.944
Erlotinib	0	0.922	0.944
Pazopanib	−1	0.922	0.944
Fludarabine	**1.5**[c]	0.854	0.518
NSC151252	**4.5**[c]	0.804	0.187
Gefitinib	0.5	0.801	0.187
precision observed/expected: 0.50/0.76			

APH(3″)-Ib							
empirical parameters				optimized parameters			
drug	ΔT_m (°C)	mTC	precision	drug	ΔT_m (°C)	mTC	precision
Lapatinib	nd[b]	0.927	0.961	NSC36398	**2**[c]	0.900	0.874
Dasatinib	1	0.926	0.961	NSC34875	**3**[c]	0.845	0.518
Sunitinib	nd	0.926	0.961				
Erlotinib	−3	0.924	0.944				
Pazopanib	−2	0.924	0.944				
NSC36398	**2**[c]	0.900	0.874				
NSC76988	**3**[c]	0.861	0.596				
NSC34875	**3**[c]	0.845	0.518				
NSC26744	**4**[c]	0.827	0.373				
Gefitinib	−1	0.804	0.187				
precision observed/expected: 0.50/0.73				precision observed/expected: 1.0/0.70			

hCA1			
drug	m	mTC	precision
Celecoxib	**2.5**[c]	0.921	0.962
NSC107679	**3.5**[c]	0.861	0.694
NSC263220	**2.5**[c]	0.837	0.540
NSC17129	**3.5**[c]	0.828	0.456
NSC133195	0	0.818	0.371
precision observed/expected:[d] 0.8/0.60			

[a]The experiment was done using predictions from an empirical choice of the four parameters before they were optimized. The predictions for hCA1 and APH(3″)-Ib are the same by both parameter sets; whereas the empirical parametrization yielded a significant number of false positives that are kinase inhibitors that are absent in the parameter set optimized on DUD-E. [b]nd means that the fluorescence signal is significantly dampened or no melting transition observed. These ligands are ignored in the calculation of positive hit rate. [c]The ΔT_m values of true binders ($\Delta T_m > 1$ °C) are indicated in bold. Positive hit rate (%) = $N_{\text{true binder}}/N_{\text{total tested with observable melting curves}}$. It is 80% for hCA1, 50% for pMDH, and 50% for APH(3″)-Ib, respectively. [d]The observed precision is calculated the same way as the experimental positive hit rate described above. The expected precision is calculated based on average precision statistics of FINDSITE$^{\text{comb2.0}}$ predictions for the 102 DUD-E benchmark sets at specific template/target sequence identity cutoff and mTC score with ΔmTC of 0.05. In the experimental validation sets, the mTC score cutoff was 0.8 for all compounds tested. There the sequence identity cutoff was set to be 30% for pMDH and APH(3″)-Ib, and there was no cutoff sequence identity cutoff for hCA1 as a control set.

on the thermal shift assay (Figure 5). Further tests confirmed that the thermal shift was due to the receptor tyrosine kinase inhibitor component of Sunitnib (free base form) and not the malate ingredient in the Sunitinib malate drug formulation. The mode of potential off-target interaction of the anticancer drug Sunitinib with malate dehydrogenase needs to be elucidated. Interestingly, it has been recently reported that cytosolic malate dehydrogenase activity supports glycolysis in actively proliferating cells and cancer.[53] Kinase inhibitors are known to often show promiscuity and are an important repertoire for drug repurposing for the treatment of infectious diseases.[54,55] Finally, the average observed precision is hCA1 is 0.8, as compared to the expected precision of 0.6.

The two binders identified for APH(3″)-Ib belong to the flavonoid family with ΔT_m in the range of 2−4 °C (Table 4, Figure 6, Table S1). A BLAST search inferred that APH(3″)-Ib is a streptomycin kinase. We confirmed that APH(3″)-Ib shows substrate specificity toward streptomycin and is inactive with kanamycin or gentamycin (Figure S1). This is consistent with a previous study where the flavanol quercetin (NSC36398) inhibits a related kanamycin active kinase APH(2″)-Iva by occupying the ATP binding site as evidenced by its crystal structure.[56]

On the other hand, when no sequence cut off was applied, hCA1 shows a relatively high observed precision of 80% which is roughly consistent with the average expected precision of 70% (Table 4), with 4 out of 5 compounds tested, all belonging to sulfonamide family known to inhibit carbonic anhydrases. These ligands have a positive ΔT_m in the range of 2.5−3.5 °C. The relatively high hit rate of the hCA1 control

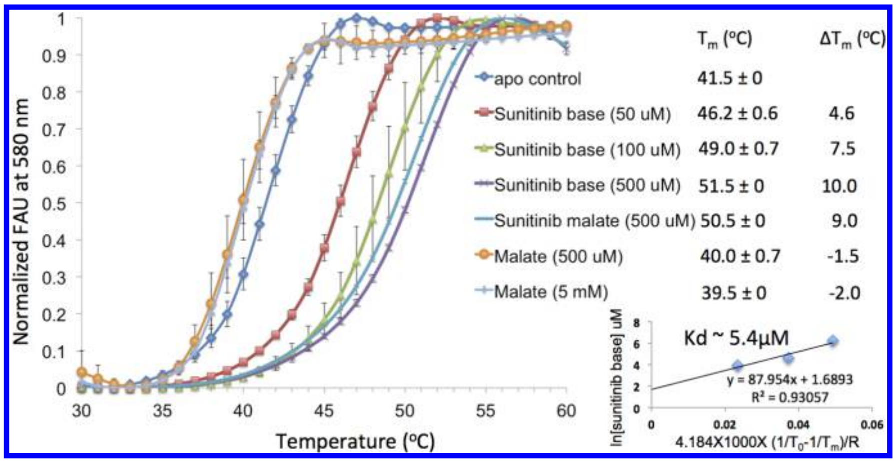


Figure 5. Thermal shifts of pMDH at different concentrations of Sunitinib either in the free base form or with malate. A malate control also tested. (inset) Linear fitting of ln[ligand] vs $4.184 \times 1000(1/T_0 - 1/T_m)/R$ where T_0 is melting temperature of the apo protein in the absence of ligand of interest, and T_m is melting temperature at a given ligand concentration. The ligand affinity to the protein K_d (μM) ≈ $e^{y\text{-intercept}}$, an approximation under conditions of [ligand] ≫ [protein] during the thermal transition based on previously derived binding thermodynamics equations.[22,43]

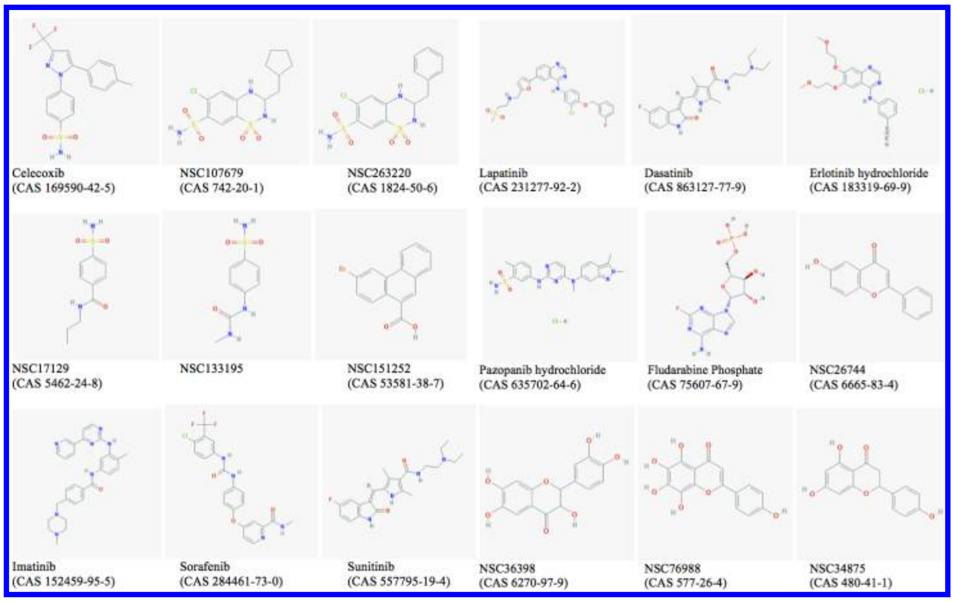


Figure 6. Chemical structures of different compounds/drugs tested in the thermal shift assay. 2D structures were directly obtained from PubChem, and the available CAS number is in parentheses.

test where no sequence identity cutoff was applied is due to the existing crystal structure of hCA1 complexes with sulfonamide inhibitors available from the PDB that provide both protein and ligand templates. For example, Celecoxib (Drug Bank ID: DB00482), an FDA approved nonsteroidal anti-inflammatory drug (NSAID) known to inhibit prostaglandin-endoperoxide synthase COX-2 (but not COX-1) and hCA2 proteins according to DrugBank (www.drugbank.ca), was identified to be a true binder for hCA1 (sequence identity between hCA1 and hCA2 is 60%). These results corroborate the prediction power of FINDSITE$^{\text{comb2.0}}$ as a VLS approach and its ability to accelerate experimental drug lead discovery by *in silico* ligand prescreening.

■ DISCUSSION

Previously, using the DUD benchmarking set,[57] FINDSITE$^{\text{comb}}$ was demonstrated to perform better than traditional docking methods for virtual ligand screening[19] and was also employed for drug target and side effect predictions for all DrugBank drugs.[26] To the best of our knowledge, FINDSITE$^{\text{comb}}$ is the only method applied to the entire human proteome. Recently developed deep learning CNN scoring boosts the performance of high-resolution structure based docking methods in terms of AUC and enrichment factor.[23,24] Although CNN methods have a better AUC than FINDSITE$^{\text{comb}}$, they have worse performance than FINDSITE$^{\text{comb}}$ in terms of the enrichment factor. The reason could be due to fact that the AUC is mostly determined by actives ranking beyond the top 1−5%, whereas the enrichment factor is determined by actives ranking at the very top. Machine learning-based methods are good at ranking actives beyond the top 1−5% range. But as a practical matter, this is not the most relevant region.

Here, based upon FINDSITE$^{\text{comb}}$, we developed the FINDSITE$^{\text{comb2.0}}$ approach that significantly improves over FINDSITE$^{\text{comb}}$ in both virtual ligand and virtual target screening. Even when modeled structures are used, FINDSITE$^{\text{comb2.0}}$ performs better than deep CNN methods that use experimental target structures, not only for the enrichment factor, but also for their AUC. The improvement is mostly attributable to filters added to the template ligand selection. Parsing the template protein targets and ligands in the DrugBank and ChEMBL databases to domains also contributes to this improvement, especially for the number of targets that show better than random ligand selection. For three proteins and a small set (<10) of predicted binding ligands, we experimentally demonstrated significant agreement between the average observed precision of 60% and the average calculated expected precision of 70% of FINDSITE$^{comb2.0}$ prediction for protein−ligand interactions when mTC > 0.80 with a sequence identity cutoff of 30% (Table 4). Using Figure 4, we can predict the likely precision of a given set of ligands, thereby suggesting when it makes sense to experimentally test FINDSITE$^{comb2.0}$'s predictions. In other words, it is the first algorithm that can suggest under what conditions it makes sense to experimentally test the VTS or VLS predictions. Since only a handful of ligands (<50) need to be screened to identify novel hits and it can employ predicted as well as experimental protein structures, FINDSITE$^{comb2.0}$ is a very powerful VLS and VTS tool that can greatly assist in accelerating drug discovery and side effect predictions.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00309.

Additional information for Materials and Methods, Figure S1, and Tables S1−S2 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: 404-407-8975. Fax: 404-385-7478. E-mail: skolnick@gatech.edu.

ORCID

Jeffrey Skolnick: 0000-0002-1877-4958

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by grant No. 1R35GM-118039 of the Division of General Medical Sciences of the National Institutes of Health. The authors would like to thank Dr. Bartosz Ilkowski for managing the cluster on which this work was conducted and Jessica Forness for assistance in preparation of the manuscript.

■ REFERENCES

(1) Settleman, J.; Cohen, R. Communication in Drug Development: "Translating" Scientific Discovery. *Cell* **2016**, *164*, 1101−04.

(2) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The Price of Innovation: New Estimates of Drug Development Costs. *Journal of Health Economics* **2003**, *22*, 151−185.

(3) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011**, *10*, 188−195.

(4) Reddy, A. S.; Pati, S. P.; Kumar, P. P.; Pradeep, H. N.; Sastry, G. N. Virtual Screening in Drug Discovery − A Computational Perspective. *Curr. Protein Pept. Sci.* **2007**, *8*, 329−351.

(5) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(6) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(7) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(8) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* **2009**, *31*, 455−461.

(9) Stahura, F.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189−1202.

(10) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity − a Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(11) Glen, R. C.; Adams, S. E. Similarity Metrics and Descriptor Spaces - Which Combinations to Choose? *QSAR Comb. Sci.* **2006**, *25*, 1133−1142.

(12) Brylinski, M.; Skolnick, J. Q-Dock: Low-resolution flexible ligand docking with pocket-specific threading restraints. *J. Comput. Chem.* **2008**, *29*, 1574−88.

(13) Brylinski, M.; Skolnick, J. Q-Dock$^{LHM:}$ Low-resolution refinement for ligand comparative modeling. *J. Comput. Chem.* **2009**, *31*, 1093−105.

(14) Lee, H. S.; Zhang, Y. BSP-SLIM: A blind low-resolution ligand-protein docking approach using predicted protein structures. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 93−110.

(15) Roy, A.; Zhang, Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* **2012**, *20*, 987−997.

(16) Brylinski, M.; Skolnick, J. FINDSITE: A threading-based method for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 129−134.

(17) Brylinski, M.; Skolnick, J. FINDSITE^{LHM}: a threading-based approach to ligand homology modeling. *PLoS Comput. Biol.* **2009**, *5*, e1000405.

(18) Zhou, H.; Skolnick, J. FINDSITE^{X}: A Structure-Based, Small Molecule Virtual Screening Approach with Application to All Identified Human GPCRs. *Mol. Pharmaceutics* **2012**, *9*, 1775−1784.

(19) Zhou, H.; Skolnick, J. FINDSITE^{comb}: A Threading/Structure-Based, Proteomic-Scale Virtual Ligand Screening Approach. *J. Chem. Inf. Model.* **2013**, *53*, 230−240.

(20) Reardon, S. Project ranks billions of drug interactions. *Nature* **2013**, *503*, 449.

(21) Brylinski, M.; Skolnick, J. Comprehensive Structural and Functional Characterization of the Human Kinome by Protein Structure Modeling and Ligand Virtual Screening. *J. Chem. Inf. Model.* **2010**, *50*, 1839−1854.

(22) Srinivasan, B.; Zhou, H.; Kubanek, J.; Skolnick, J. Experimental validation of FINDSITEcomb virtual ligand screening results for eight proteins yields novel nanomolar and picomolar binders. *J. Cheminf.* **2014**, *6*, 16−29.

(23) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein−Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−57.

(24) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. *arXiv.org* **2015**, 1510.02855.

(25) Chen, Y.; Zhi, D. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 217−226.

(26) Zhou, H.; Gao, M.; Skolnick, J. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Sci. Rep.* **2015**, *5*, 11090.

(27) Wishart, D.; Knox, C.; Guo, A.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−72.

(28) Zhou, H.; Zhou, Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 321−328.

(29) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(30) Zhou, H.; Skolnick, J. Template-based protein structure modeling using TASSER$^{\mathrm{VMT}}$. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 352−361.

(31) Gaulton, A.; Bellis, L.; Bento, A.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−07.

(32) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.

(33) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.

(34) Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(35) Schmidt, T.; Haas, J.; Cassarino, T. G.; Schwede, T. Assessment of ligand binding residue predictions in CASP9. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 126.

(36) Tanimoto, T. T. *An elementary mathematical theory of classification and prediction*. IBM Internal Report, 1958.

(37) Anonymous. Daylight Chemical Information Systems, Inc, Aliso Viejo, CA, 2007.

(38) Bateman, A.; Birney, E.; Cerruti, L.; Durbin, R.; Etwiller, L.; Eddy, S.; Griffiths-Jones, S.; Howe, K. L.; Marshall, M.; Sonnhammer, E. L. L. The Pfam Protein Families Database. *Nucleic Acids Res.* **2002**, *30*, 276−280.

(39) Pandit, S.; Skolnick, J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinf.* **2008**, *9*, 531.

(40) Zhang, Y.; Skolnick, J. A scoring function for the automated assessment of protein structure template quality. *Proteins: Struct., Funct., Genet.* **2004**, *57*, 702−710.

(41) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 10915−10919.

(42) Cao, H.; Walton, J. D.; Brumm, P.; Phillips, G. N. J. Structure and Substrate Specificity of a Eukaryotic Fucosidase from Fusarium Graminearum. *J. Biol. Chem.* **2014**, *289*, 25624−25638.

(43) Lo, M.; Aulabaugh, A.; Jin, G.; Cowling, R.; Bard, J.; Malamas, M.; Ellestad, G. Evaluation of Fluorescence-Based Thermal Shift Assays for Hit Identification in Drug Discovery. *Anal. Biochem.* **2004**, *332*, 153−159.

(44) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97−101.

(45) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.

(46) Xia, J.; Jin, H.; Liu, Z.; Zhang, L.; Wang, X. S. An Unbiased Method To Build Benchmarking Sets for Ligand-Based Virtual Screening and its Application To GPCRs. *J. Chem. Inf. Model.* **2014**, *54*, 1433−1450.

(47) van Laarhoven, T.; Marchiori, E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One* **2013**, *8*, e66952.

(48) Lounkine, E.; Keiser, M.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A.; Cote, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361−368.

(49) Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Predicting of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **2012**, *8*, e1002503.

(50) Cobanoglu, M. C.; Liu, C.; Hu, F.; Oltvai, Z. N.; Bahar, I. Predicting Drug−Target Interactions Using Probabilistic Matrix Factorization. *J. Chem. Inf. Model.* **2013**, *53*, 3399−3409.

(51) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, i232−i240.

(52) Yamanishi, Y.; Pauwels, E.; Kotera, M. Drug side-effect prediction based on the integration of chemical and biological spaces. *J. Chem. Inf. Model.* **2012**, *52*, 3284−92.

(53) Hanse, E. A.; Ruan, C.; Kachman, M.; Wang, D.; Lowman, X. H.; Kelekar, A. Cytosolic Malate Dehydrogenase Activity Helps Support Glycolysis in Actively Proliferating Cells and Cancer. *Oncogene* **2017**, *36*, 3915−3924.

(54) Walsh, C. T.; Fischbach, M. A. Repurposing Libraries of Eukaryotic Protein Kinase Inhibitors for Antibiotic Discovery. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 1689−1690.

(55) Dichiara, M.; Marrazzo, A.; Prezzavento, O.; Collina, S.; Rescifina, A.; Amata, E. Repurposing of Human Kinase Inhibitors in Neglected Protozoan Diseases. *ChemMedChem* **2017**, *12*, 1235−1253.

(56) Shakya, T.; Stogios, P. J.; Waglechner, N.; Evdokimova, E.; Ejim, L.; Blanchard, J. E.; McArthur, A. G.; Savchenko, A.; Wright, G. D. A Small Molecule Discrimination Map of the Antibiotic Resistance Kinome. *Chem. Biol.* **2011**, *18*, 1591−1601.

(57) Huang, N.; Shoichet, B.; Irwin, J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.