HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2015 October 05.

Published in final edited form as:

J Chem Inf Model. 2015 August 24; 55(8): 1757–1770. doi:10.1021/acs.jcim.5b00232.

PoLi: A Virtual Screening Pipeline Based On Template Pocket And Ligand Similarity

Ambrish Roy, **Bharath Srinivasan**, and **Jeffrey Skolnick***

Center for the Study of Systems Biology Georgia Institute of Technology 250 14th Street NW, Atlanta, GA 30318

Abstract

Often in pharmaceutical research, the goal is to identify small molecules that can interact with and appropriately modify the biological behavior of a new protein target. Unfortunately, most proteins lack both known structures and small molecule binders, prerequisites of many virtual screening, VS, approaches. For such proteins, ligand homology modeling, LHM, that copies ligands from homologous and perhaps evolutionarily distant template proteins, has been shown to be a powerful VS approach to identify possible binding ligands. However, if we want to target a specific pocket for which there is no homologous holo template protein structure, then LHM will not work. To address this issue, in a new pocket based approach, *PoLi*, we generalize LHM by exploiting the fact that the number of distinct small molecule ligand binding pockets in proteins is small. *PoLi* identifies similar ligand binding pockets in a holo-template protein library, selectively copies relevant parts of template ligands and uses them for VS. In practice, *PoLi* is a hybrid structure and ligand based VS algorithm that integrates 2D fingerprint-based and 3D shape-based similarity metrics for improved virtual screening performance. On standard DUD and DUD-E benchmark databases, using modeled receptor structures, *PoLi* achieves an average enrichment factor of 13.4 and 9.6 respectively, in the top 1% of the screened library. In contrast, traditional docking based VS using AutoDock Vina and homology-based VS using FINDSITE^{filt} have an average enrichment of 1.6 (3.0) and 9.0 (7.9) on the DUD (DUD-E) sets respectively. Experimental validation of *PoLi* predictions on dihydrofolate reductase, DHFR, using differential scanning fluorimetry, DSF, identifies multiple ligands with diverse molecular scaffolds, thus demonstrating the advantage of *PoLi* over current state-of-the-art VS methods.

Introduction

Identifying lead molecules that bind to a given target protein is a fundamental challenge in pharmaceutical research. This issue has been addressed using both experimental high throughput screening (HTS) and computational *in-silico* (commonly referred as virtual screening, VS) approaches[1]. Although HTS is currently the best method for lead identification, the dependence of the results on experimental factors, “chemical space” coverage, applicability for all targets, along with the cost and time required to perform such

*All correspondence should be addressed to skolnick@gatech.edu.

Availability *PoLi* is freely available as a webserver at http://cssb.biology.gatech.edu/PoLi.

Competing interests The authors declare no competing financial interest.

screens limit their applicability[2]. For this reason, new computational approaches that can efficiently screen large databases are needed, as they not only complement HTS but also have much higher throughput and greatly reduced cost and increased speed[3].

Based on the availability of target protein structures, either structure or ligand-based VS calculations are performed to identify potential lead molecules. The most commonly used structure-based VS approach is molecular docking, which does not require *a priori* knowledge of known binders[4] and can target a specific binding pocket of interest. Molecular docking involves screening database molecules based on their calculated interaction energy with the receptor binding site[5]. As such, its performance relies heavily on the receptor structure quality and flexibility[6]. For example, ~90% of docking accuracy is lost if models of trypsin and HIV-1 protease with a root-mean-square deviation from native, RMSD, >1.5Å are used[7]. It also depends on the presence of water molecules, the conformations of database molecules, and the sensitivity of the scoring function used for evaluating protein-ligand interactions[8]. Another structure based variant docks small molecule fragments to screen for promising leads[9]. However, distinguishing binding and non-binding fragments is a challenge in these methods, as fragments bind with very low weak binding affinity which cannot be captured using the inaccurate scoring functions that we currently have[10]. Moreover, like other small molecule docking approaches, fragment-based approaches also require a high-resolution structure, which is not always available. To address this problem homology models that are very closely related to the template proteins in the PDB have been used; moreover, the models frequently require a lot of side-chain refinement[11].

In the absence of a target receptor structure, ligand-based VS approaches are generally used. Ligand-based VS is robust, but requires at least one known bioactive molecule, that is used as a seed to fish out database molecules with similar chemotypes. Most common ligand-based VS approaches evaluate a 2D fingerprint[12], pharmacophore[13] or 3D shape-based similarity[14] between known bioactive and database molecules. Thus, most structure and ligand-based VS methods require either an experimentally solved receptor structure or an experimentally determined bioactive molecule. As such, they cannot be readily applied to many proteins of therapeutic interest.

To address these significant limitations, we recently described two new virtual screening approaches[15, 16]. The first, FINDSITE^{filt} [15], can use either experimental or predicted low-resolution target protein structures to screen database molecules based on 2D fingerprint similarity with template ligands in the PDB holo template library. FINDSITE^{comb} includes FINDSITE^{filt} for proteins having holo templates, but for those proteins lacking holo templates, also uses an artificially generated template library of predicted tertiary structures whose binding ligands are found in the CHEMBL[17] and DrugBank[18] ligand binding databases. Template ligands are copied from globally related protein structures based on structural similarity to the target, without considering where the ligand actually binds in the template protein. These methods have the inherent advantages of speed, lack of requirement of high-resolution protein structures, and do not need known binders. Although both approaches achieve good enrichment in identifying out active molecules, FINDSITE^{filt}, in particular, depends on the availability of proteins with a similar fold in the holo template

library for effective virtual screening. More importantly, both FINDSITE^{filt} and FINDSITE^{comb} were not developed with the goal of targeting a specific binding pocket.

To begin to generalize our approach, we developed a shape-based virtual screening algorithm, LIGSIFT[16] that screens database molecules based on their 3D shape and chemical feature similarity to a target seed ligand. LIGSIFT was benchmarked using the 3D similarity of database molecules to a known binding ligand to the target protein, as provided by the DUD database[19], and its performance without known binding ligands was not established. Thus, a new pocket centric approach that can target a specific binding pocket of interest, overcome the requirement of global fold similarity between template and target structures, and which combines both 2D and 3D based ligand similarity metrics for virtual screening using ligands identified from holo templates is needed.

Based on these ideas and the fact that the space of protein-ligand binding pockets is small and close to complete[20], we developed a new virtual screening pipeline, *PoLi*, that first predicts the ligand binding pocket in the target protein, selectively copies parts of template ligands based on binding-pocket alignment, then performs virtual screening of database molecules based on combined 2D and 3D ligand similarity metrics to the selected template small molecules. Large scale *in-silico* benchmarking followed by *in-vitro* high-throughput experimental validation of predictions on *E. coli* DHFR, establishes *PoLi* as an effective virtual screening approach.

Results

Overview of PoLi pipeline

PoLi is based on the basic idea that the number of distinct binding pockets is small[20], and for many query proteins that lack any known active molecule, one can detect binding pockets in the query protein structure, identify similar pockets in the holo-template library, copy ligands from similar pockets and subsequently use them for ligand-based virtual screening. Figure 1 shows the schematic representation of *PoLi*. The input to the *PoLi* pipeline is a 3D structure of the query protein. If an experimentally determined structure of query protein is already available, it can be used directly; otherwise starting from a query protein sequence, the TASSER-VMT[21] structure modeling pipeline is used to generate 3D models of the query protein. The next step in the hierarchical pipeline is the detection of ligand binding sites and copying of ligands from similar binding pocket of holo-template proteins. In *PoLi*, ligand binding pockets in the query protein's structure are predicted using two different approaches: first by global structural superposition of holo-proteins in the PDB library on the query protein structure using the TM-align[22] structure alignment algorithm, and second by detecting pockets using the ConCavity[23] algorithm. These predicted pockets are then structurally aligned against known ligand binding pockets in the PDB holo-template library, using the sequence-order independent binding-site comparison algorithm APoc[24], and template ligands from similar binding pockets are copied in the query ligand-binding pocket. These copied template ligands are later pruned to remove parts of the template ligand that interact with the unaligned region of template binding pocket, and then used in ligand-based VS. Virtual screening in *PoLi* is performed using a combination of 2D fingerprint based and 3D-shape based similarity metrics, where the 2D path-based

fingerprint is generated using OpenBabel[25] and 3D similarity is calculated using a variant of the LIGSIFT[16] algorithm. LIGSIFT is a small molecule structural alignment algorithm that uses an atom-centered smooth Gaussian function to describe the ligand structure and perform rapid overlay to measure 3D shape and chemical similarity. The ligand 3D similarity between molecules in LIGSIFT is evaluated using a size-independent scoring function (a scaled TC). The statistical significance of the similarity score (*P-value*) is estimated based on millions of comparisons of randomly selected ligands[16]. A detailed description of the pipeline modules is provided in the Materials and Methods section.

Detection of template seed ligands for virtual screening

We first validate our approach to detect chemically similar ligands using a pocket-based search. These detected ligands will be used as seed ligands for ligand-based virtual screening. The objective of this exercise is to show that the ligands copied from template proteins have a statistically significant chemical similarity to the native ligand when they come from structurally similar pockets as assessed using the ligand binding pocket structural comparison algorithm APoc[24]. In practice, we selected a non-redundant set of 30,000 ligand-pairs with statistically significant chemical similarity (LIGSIFT 3D chemical similarity *P-value* < 0.001) and 35,000 ligand pairs that lack significant chemical similarity score from the PDB holo template library, such that the corresponding receptor pairs share < 30% sequence identity. Figure 2 shows the performance of APoc's pocket similarity[24] to detect templates that have chemically similar ligands bound to them, in comparison to TM-align[22] (global structural similarity) and HHalign[26] (threading). Predictions are labeled as correct if the *P-value* of the 3D chemical similarity score between copied template ligand and the query ligand is < 0.001. As shown in Figure 2, the pocket similarity based approach (APoc) outperforms both TM-align and HHalign in detecting true positives. For instance at 95% specificity, APoc identifies 34% of chemically similar ligand pairs, TM-align global structural template matching recovers 18.5% true positives, while HHalign only identifies 14.5% true positives. This establishes that pocket similarity is the best of the three approaches to identify templates that have ligands with overlapping chemical features.

Benchmarking performance on DUD and DUD-E databases

Benchmarking of virtual screening algorithms was done on 40 DUD database proteins[19] and 65 proteins included in the DUD-E database[27]. Both are routinely used for testing scoring functions and virtual screening methods. Our objective here is to analyze the overall performance of the *PoLi* pipeline, which includes structure modeling of the receptor, binding site prediction and virtual screening of database molecules (see Materials and Methods). We also ran the same pipeline using experimentally solved protein structures to assess the effect of model quality on virtual screening performance. For structure modeling of target proteins and binding site predictions using both modeled and experimental structure, closely related homologous proteins were excluded from template libraries using a sequence identity threshold of 30%.

Model quality of target proteins—Since model quality and accuracy of binding site predictions are expected determinants of structure-based methods for virtual screening, including *PoLi*, we first examine the quality of predicted protein structures. Figures 3A & B

present the global and local structure quality of predicted TASSER[21] models. The global structure quality of models is measured as TM-score[28], with values ranging between [0,1] with a higher score indicating better structural match between the model and native structure. Statistically, a TM-score < 0.3 means random structural similarity and TM-score >0.4 indicates that the protein pairs have a similar fold. The average TM-scores of DUD and DUD-E set proteins are 0.76±0.18 and 0.73±0.12 respectively (Fig 3A), clearly highlighting that the predicted structure of most proteins share high structural similarity with the experimentally determined structure. Two proteins in both sets, namely *hmgr* (hydroxymethylglutaryl-CoA reductase) and *sahh* (S-adenosyl-homocysteine hydrolase) in the DUD set, and *nos1* (Nitric oxide synthase) and *pa2ga* (Phospholipase A2 group IIA) in the DUD-E set, have incorrectly predicted structures; i.e., the TM-score between the model and experimental structure is < 0.4. For these proteins, the structural confidence C-score of model[29] is also < −3; i.e. they can be easily recognized as having poorly predicted structures even in the absence of experimentally determined structures (Table S2). Fig 3B shows the structure quality of the predicted models near the known ligand binding site of the co-crystallized ligand. The mean Cα RMSD of binding pocket residues (residues within <4.5 Å from co-crystallized ligand in experimental structure) in the DUD proteins is 4.3±5.7Å and in the DUD-E proteins, it is 3.3±3.2Å. In most cases, the structure near the known ligand binding pocket is also reasonably well predicted (Table S2), with some local structural variations (typical of any homology based structure modeling algorithm). This is not surprising, as functionally important regions are generally more conserved than other parts of the protein and are more likely to be correctly predicted. Nevertheless, for some proteins, the structural variations of the binding pocket residues can be large (Cα RMSD >5Å), because of two reasons: (a) the global structure itself is incorrectly predicted and so is the binding pocket (e.g. in hydroxymethylglutaryl-CoA reductase), or (b), while the global fold is basically correct, the structure of the ligand binding site is only partially correct. For example, it could be an inter-domain binding pocket with one incorrectly predicted domain (e.g. in glycogen phosphorylase beta). Such structural variations affect both binding pocket predictions and have a seriously adverse effect on the performance of molecular docking methods that use these models.

Analysis of binding site predictions—Figure 3C shows the performance of the *PoLi* pipeline in recapitulating known ligand binding sites as provided in the DUD and DUD-E databases[19, 27] using both modeled and experimental structures. Using modeled structures, ligand-binding pockets are correctly identified (within 5Å from the geometric center of the experimentally solved ligand-protein) in 32 of the 40 DUD set proteins, and in 52 of the 65 DUD-E set proteins. When experimental structures are used, binding pockets can be correctly predicted for 36 proteins in DUD set and 60 proteins in DUD-E set, within the same distance cutoff. Among the modeled protein structures with incorrectly predicted binding pockets (pocket distance >5Å), 5 of the 8 proteins in DUD set and 7 of the 13 proteins in DUD-E set have binding pocket residues with a Cα RMSD >5Å. For the remaining predicted and experimental structures even though binding pocket cavities could be detected, they lacked a significant match (*P*-value < 0.001 and PS-score > 0.35) to known ligand binding pockets in the PDB holo template library. This is one of the main limitations of LHM based binding site predictions. Thus, these VS predictions are of poor quality. Also

in some targets (e.g. in HIV reverse transcriptase), the known ligand-binding site is interfacial (formed by contacts of protein chains in a complex) and cannot be always predicted using monomeric structures alone (especially in those models having structural variations near the pocket), a limitation of this approach.

Virtual screening performance on DUD and DUD-E targets—The above analyses have shown that: (a) a pocket-based approach is better than both global similarity and homology based approaches in detecting templates whose ligands have similar chemical features, and (b) for most proteins, computationally generated models have a correctly predicted fold, whose ligand binding pockets can also be correctly identified in ~80% of the cases. In this section, we examine the next module of the *PoLi* pipeline: the ability to identify active molecules in the DUD[19] and DUD-E[27] databases. Performance is evaluated using standard evaluation metrics: (a) the Enrichment Factor (EF) of the screened compound library, (b) the Hit Rate (HR) of active molecules and (c) the Receiver Operating Characteristic (ROC) curve. Descriptions of these metrics are given in Materials and Methods.

Table 1 shows the virtual screening performance of *PoLi* using both computationally generated models and experimentally determined receptor structures. The average enrichment in the top 1% of the screened library is 13.4 and 9.6 for DUD and DUD-E set modeled receptor structures, and the average hit rates are 38.0 and 14.3 respectively. When experimental structures are used, the enrichment rates in the top 1% increase to 15.2 and 9.6, and the hit rate increases to 43.3 and 14.6 respectively for the DUD and DUD-E sets. A paired Wilcoxon signed rank *t*-test between EF1% achieved using model and experimental structures has a *p*-value of 0.44 on the DUD set and 0.30 on the DUD-E set, suggesting that the difference in VS performance using model and experimental structure is not statistically significant. Moreover, using a known binder of each target protein (taken from the experimental structure in PDB), the best average EF1% obtained using LIGSIFT shape-based screening is 17.4 and 18.7 for the DUD and DUD-E sets respectively; this is notable since *PoLi* predictions were generated by using ligands copied from templates with < 30% sequence identity.

Since model quality and accuracy of binding pocket predictions directly affect the performance of the *PoLi* pipeline, we further analyzed the VS results only for proteins with reasonable quality model (TM-score > 0.5) and for those proteins in which one of the predicted pockets overlap with the known ligand binding site in the experimental structure (pocket distance <5 Å). Since most proteins are reasonably well predicted, using correctly modeled structures, the EF1% on DUD and DUD-E set, marginally improved to 14.1 and 9.9 respectively (Table 1), an increase of approximately 3–5% compared to EF1% obtained for all the proteins. A more significant improvement is observed when proteins in which the known ligand binding site was recapitulated as one of the binding site predictions. The EF1% for DUD and DUD-E are 15.9 and 11.0 respectively, an improvement of approximately 14–18%.

It is interesting to observe that using both modeled and experimental structures the performance of *PoLi* is consistently lower on the DUD-E set compared to the DUD set,

while performance remains similar when known binders are used as input for LIGSIFT-based VS. This decrease in performance cannot be attributed to either structure quality, as the average TM-score for both sets ~0.7, or to the accuracy of binding pocket predictions, as just 20% of modeled proteins and ~10% of experimental structures in both the sets have predicted pockets at a distance >5Å from the geometric center of the experimentally solved ligand location in the protein's structure.

We sought to analyze this further by examining the highest 3D and 2D molecular similarity between database molecules and collected template ligands. Table 2 clearly shows that the main reason for the decrease in performance on the DUD-E set (using both experimental and modeled structures) is because of increased overlap between the active and decoy molecular similarity distributions. More specifically, there is an overall decrease of 3D similarity scores in the DUD-E compared to the DUD database. A detailed statistical analysis performed by taking random samples from DUD and DUD-E database and analyzing the difference between 3D similarity scores of actives and decoy molecules reveals that the mean of the difference distribution is 0.08 in the DUD set and 0.02 in the DUD-E set. Also a Welch's *t*-test performed on the difference distributions has a *P*-value < 2.2E-16, suggesting that difference between the highest similarity scores of actives and decoys in the DUD set was significantly greater than in the DUD-E set.

Comparison with control methods for virtual screening—Without known binders, molecular docking is the most widely used virtual screening approach and has been benchmarked on numerous occasions using experimental structures[30, 31]. Another virtual screening approach, that is becoming increasingly popular copies ligands from homologous/structurally analogous template proteins and uses them as seeds for ligand-based virtual screening[15, 32]. Here, template ligands are copied, and either a single or combination of different 2D molecular similarity metrics is used for ranking the database molecules.

As our experimental control, we employed the widely used molecular docking tool AutoDock Vina[33], our in-house developed VS algorithm FINDSITE^{filt} (as it also uses a PDB holo template library) and shape-based VS using LIGSIFT. Docking runs of AutoDock Vina were performed with default options, and the entire receptor structure was enclosed within a box during the docking simulations (as if the binding pocket were unknown). Furthermore, to avoid any bias arising due to differences in holo template library, both FINDSITE^{filt} and LIGSIFT used the same set of templates as *PoLi* for virtual screening. FINDSITE^{filt} uses a 2D fingerprint similarity metric (Eq. 6) between these selected templates and database molecules, while LIGSIFT uses these template ligands as seeds (without any pruning) for shape-based structural alignment with database molecules (Eq. 4). Thus, FINDSITE^{filt} is the VS performance achieved using a 2D approach, while LIGSIFT is representative of a 3D VS algorithm.

Table 3 reports the AUC, EF and HR obtained on the DUD and DUD-E sets using modeled protein structures. The average enrichment factors of *PoLi*, LIGSIFT, FINDSITE^{filt} and AutoDock Vina in the top 1% of the screened library (EF1%) are 13.4, 11.8, 9.0 and 1.6 respectively on the DUD set. A similar trend is also observed on DUD-E set where *PoLi*, LIGSIFT, FINDSITE^{filt} and AutoDock Vina achieve EF1%s of 9.6, 5.9, 7.9 and 3.0

respectively. Fig S1 shows the distribution of AUC and EF1% for the DUD and DUD-E set proteins using a boxplot. A paired Wilcoxon signed rank *t*-test between EF1% of *PoLi* and control methods (LIGSIFT, FINDSITE^{filt} and AutoDock Vina), after Boneferroni correction for multiple comparison, have *p*-values of 8.45E-02, 4.18E-02 and 1.51E-06 respectively on the DUD set and 0.0004, 0.0099 and 0.0011 respectively on the DUD-E set of proteins. It is clear from these results that establishing which molecular similarity metrics (3D shape-based or 2D fingerprint based) is better is difficult, as their performance can vary with the protein target. Nevertheless, fusion of 2D and 3D similarity metrics based on their Z-score (Eq. 7) shows the best performance in virtual screening on the tested databases. The observed improvement of *PoLi* is also partially due to the pruning of template ligands. Biased structural overlap of ligands near the hot-spot regions also contributed to the enrichment of actives in the DUD set, where EF1% increased from 12.3 for unbiased structural overlap to 13.4 for biased overlap. For DUD-E set the performance was similar, where EF1% was 9.7 for unbiased structural overlap and 9.6 for biased overlap. Molecular docking using AutoDock Vina has the worst performance in identifying active molecules. One might expect that without explicitly providing the exact location of target binding site, molecular docking will certainly result in poor enrichment of active molecule. However, a similar analysis done by Feinstein and Brylinski[32] have shown that even when the predicted binding site in modeled receptor structures of the DUD-E set were specified, the resulting EF1% was 2.45 and 2.86 on high and medium quality models. These results suggest that traditional docking-based approaches cannot correctly evaluate protein-ligand interactions on predicted protein structures, as they frequently have incorrect side-chain orientations.

Predictions using globally unrelated template proteins—An important advantage of *PoLi* over existing template-based methods[15, 32] for virtual screening is that it can copy ligands from proteins with different folds but similar pockets and use them for ligand-based virtual screening. To examine this in greater detail, we performed an experiment in which binding site predictions and ligand copying were done using templates with unrelated fold (TM-score < 0.4) and templates with similar fold. Table 4 shows the result of this analysis on the DUD and DUD-E databases. It is encouraging to observe that using ligands copied from globally unrelated template proteins, *PoLi* can achieve an EF1% of 7.1 on the DUD set and 2.7 on DUD-E set proteins. These EF1% values are significantly higher on the DUD set and are similar for DUD-E targets when compared to the EF1% obtained using molecular docking (Table 3), which is currently the best approach for screening database molecules in the absence of any homologous/structurally analogous holo-template protein. Similarly, when we restrict *PoLi* to only use template ligands from related folds (TM-score >0.4), the EF1% on DUD and DUD-E targets increases to 12.0 and 8.7 respectively, which is still lower than that achieved using default *PoLi* pipeline (Table 1) that uses all templates ligands irrespective of the fold they were collected from. It needs to be mentioned that when we restricted *PoLi* to use only template proteins with similar global fold, then 8 proteins in DUD set and 11 proteins in DUD-E set failed to generate any predictions because of lack of similar template pockets. For the subset of proteins where predictions could be made using globally related template proteins, the EF1% is 15.0 and 10.5 on DUD and DUD-E sets respectively. On the same set, a combination of both globally related and unrelated template ligands yield EF1% of 15.9 and 10.0 respectively. These results highlight that even though

template ligands copied from globally related proteins on average yield better enrichment during virtual screening; ligands copied from unrelated folds improve prediction coverage. For example EF1% for targets that could only be predicted after copying ligands from globally unrelated template structures (shown as Failed in Table 4) are 4.0 and 7.3 on DUD and DUD-E set proteins. Also, unrelated fold template ligands complement the ligands templates copied from globally related template proteins to improve the overall virtual screening performance, as observed for DUD database proteins (shown as Combined in Table 4).

Pocket specific virtual screening performance—Another important advantage of *PoLi* compared to other LHM methods[15, 32], is its ability to generate pocket specific predictions, similar to docking approaches. To analyze if pocket specific predictions can yield better virtual screening performance, we analyzed the EF1% and AUC of ranked database molecules for the top 5 predicted pockets treated individually (Table 5). As shown in the table, in both the DUD and DUD-E databases, the best virtual screening performance (both EF1% and AUC) is achieved using the top predicted pocket, which has the maximum number of superposed template ligands (pocket 1). Using modeled receptor structures, pocket 1 results in an average AUC and EF1% of 0.77 and 13.3 on the DUD set, and 0.74 and 9.4 on the DUD-E set. Interestingly, virtual screening on other predicted pockets (pocket 2–5) also resulted in non-random ranking of database molecules (AUC > 0.5 and EF1% > 0). Moreover, the combined ranking procedure used in *PoLi*, which combines predictions from all the pockets, results in slightly improved predictions compared to individual pocket based predictions (compare Table 1 & 5). This suggests that some of the experimentally known active molecules in the DUD and DUD-E databases could bind in pockets different from pocket 1. For example, both experimentally verified canonical and alternate binding sites in PPAR[34] were predicted by *PoLi* and resulted in non-random predictions (AUC > 0.5 and EF1% >0) for both sites.

Experimental validation of PoLi VS—To demonstrate the utility of *PoLi* as a better VS option in identifying small molecule binders, experimental validation was carried out using a high-throughput DSF approach. The method relies on the increase in fluorescence quantum-yield of the extrinsic fluorophore reporter dye Sypro orange upon its interaction with an unfolded protein. In the presence of the ligand that binds to and stabilizes the protein of interest, the transition midpoint of unfolding shifts to higher temperatures, the magnitude of which is proportional to the strength of binding.

Escherichia coli DHFR, an enzyme that is the sole source of cellular tetrahydrofolate and thus pivotal for nucleic acid synthesis, was chosen for its immense medical importance[35]. The top 90 predictions from *PoLi* (approximately the top 3% of the ligand library) were tested. Out of 76 interpretable curves, (i.e. those showing a single sigmoidal transition and reasonably good Q values; see Methods), 14 curves showed a substantial shift in their thermal unfolding transition midpoint indicative of ligand binding (Fig.4 and Fig.5). This indicates a success rate of 18.4%. Table 6 shows the thermal shift assay parameters for all hits. 7 out of the 14 hits obtained were within the top 10 ranks assigned by the *PoLi* VS algorithm with a distinct positive skew to the distribution of top ranking hits when plotted

against the rank. Moreover, 13 of the 14 hits have consistently low μM affinities in spite of the high T_m of 51.9 °C for the protein-alone. This is a clear indication of the strength of the methodology in identifying experimentally verified binders as top ranking predictions.

Figure 4 shows the thermal melting curves, their first derivatives and the non-linear fits used to estimate thermal melt parameters for the various classes of molecules that showed unambiguous binding to prokaryotic DHFR. Figure 5 provides the chemical structures for these hits.

The algorithm was capable of picking up derivatives of 1,3,5-triazine-2,4-diamine; this represents the most populated group of identified ligands (Fig. 5A). Among molecules belonging to this class, NSC133071 shows the highest shift with a ΔT_m of 14.3 °C followed by NSC168184 & CHEMBL597262 with ~10.6 °C each, CHEMBL333873 with 9.6 °C, NSC117268 with 8.5 °C and NSC104129 with 5.4 °C, respectively (Table 6, Fig 4A & 4E and Fig 5A). An approximate estimate of the dissociation constant for NSC133071 shows that it binds tightly to *E. coli* DHFR, with a 6.4 ± 1.5 μM K_D (Table 6). The tighter binding of this molecule compared to others from this class can be ascribed to possible favorable contacts made by the [3-chloro-4-(3-phenoxypropoxy)phenyl] substituent at the 1st position of the triazine ring. It should be noted here that cycloguanil, a molecule belonging to the 1,3,5-triazine-2,4-diamine class, is a known inhibitor of *Plasmodium falciparum* DHFR [36]. However, to the best of our knowledge, no report exists on either binding or inhibition of *E. coli* DHFR by molecules predicted by *PoLi* VS and experimentally validated in the current study. Thus, all hits are novel binders. Moreover, in spite of the presence of 1,3,5-triazine-2,4-diamine ring, it would be difficult to predict the binding of NSC117268 to *E. coli* DHFR solely relying on 2-D ligand comparison methodologies or SAR intuition (Table 6). The presence of two bulky ortho ring substituents at the 1st and 6th position on the core ring precludes intuitive assumptions about binding. We posit that the 3-D method of comparison facilitated the prediction of NSC117268 as a potential binder.

The second class of molecules predicted to bind to *E. coli* DHFR, and subsequently validated experimentally, are derivatives of quinazoline-1,3-diamine (Fig. 4B, 4F and 5B). In previous studies from our lab[37–39], we have demonstrated the binding and potent inhibition of *E. coli* DHFR by two of these molecules (NSC339578 and NSC309401); both contain a pyrroloquinazoline core ring. The prediction of these molecules by *PoLi* as potential binders validates the VS approach and demonstrates its predictive power. Furthermore, a novel molecule NSC305782 showed binding to the enzyme with a ΔT_m of 14.4 °C, indicative of strong binding.

The third class of predicted molecules contains either a diaminopteridine ring (NSC740) or a diaminopyrimidine ring (NSC7364 and NSC71669) (Fig 4C, 4G and 5C). NSC740, commonly known as methotrexate, is a well-known DHFR inhibitor acting on both prokaryotic and eukaryotic homologs[35, 40]. Likewise, NSC7364 is commonly known as metoprine and is also a known inhibitor of DHFR from various sources[41]. Prediction of the above two molecules serves as an internal quality control of the VS algorithm's predictive ability and reinforces our confidence in the novel ligands that are predicted. The sole novel hit from this class, NSC71669, with two trifluoromethyl phenyl substituents on the

diaminopyrimidine ring gave a ΔT_m of 15 °C that translates into an approximate dissociation constant of 05.2 ± 1.3 μM. Once again, it should be noted here that NSC71669 would have been difficult to predict solely relying on 2-D comparison methodology (Table 6) or SAR intuition.

Lastly, the fourth class contains two hits (NSC89759 and NSC11150) with structures containing 2,4-dihydroxyphenyl rings that are very different from known DHFR inhibitors (Fig 4D, 4H and Fig 5D). This class of compounds would require further experimental proof before establishing their veracity as genuine DHFR binders/inhibitors. If these molecules are true hits, they represent novel structural scaffolds amenable to further exploration as potential DHFR inhibitors.

In conclusion, *PoLi* predicted 14 ligands as binders of *E. coli* DHFR, with 10 of them being novel. Further, it offers the advantage of predicting diverse ligands as potential binders in that it uses a 3-D metric that aids in selecting ligands that may get overlooked if only a 2D metric of ligand comparison is employed.

Discussion

Drug discovery pipelines have many bottlenecks, but new computational methods capable of identifying multiple novel lead molecules that likely bind to the protein of interest could improve the situation. In that regard, computational approaches that employ molecular similarity based searches and small molecule docking are the two most commonly used methods for virtual ligand screening. While molecular similarity based VS requires *a priori* knowledge of at least one known binder, for molecular docking, receptor structure quality is crucial for success. Such limitations have proven to be quite problematic. Methods that can use computationally generated receptor structures will allow us to approach drug discovery from a Systems Biology perspective and investigate the interaction of lead molecules at the proteome level. In that regard, we have developed a number of methods that can use modeled receptor structures for lead identification[15, 42]. Our initial efforts in this direction utilized ligands from structurally related template proteins for ligand-based VS[15]. While the capability of this method has been both computationally and experimentally demonstrated for its ability to correctly predict new lead molecules for diverse targets[15, 37], it has some inherent limitations: (a) template ligands are used without any pruning to remove parts that interact with template binding site region bearing no similarity with target pocket, (b) template ligand selection is limited to proteins sharing global structural similarity to target, and (c) the predictions are not pocket specific and cannot be used for targeting a specific binding pocket of interest.

To address these limitations, we have developed *PoLi*, which copies ligands from related pockets (irrespective of the global fold of the template protein), prunes the ligand to avoid false positive matches and then uses them in virtual screening. Moreover, since specific pockets can be targeted, one might be able to identity ligands with novel models of action. Other special features of *PoLi* include: (a) biased structural overlap between the database molecule and template ligand to promote overlap in *hot-spot* regions of the target's binding

pocket, and (b) ranking of database molecules using a data fusion technique that combines 2D and 3D molecular similarity scores for improved virtual screening performance.

On the widely used DUD and DUD-E benchmark databases, *PoLi* shows improved performance in detecting active molecules compared to all other methods used in this study. Notably, even when template proteins with similar fold (TM-score > 0.4) are excluded, *PoLi* achieves an EF1% of 7.1 on DUD database proteins and 2.7 on DUD-E database proteins, which is significantly higher for the DUD set and similar for DUD-E set when compared to the EF1% achieved using AutoDock Vina molecular docking. Considering that many proteins lack a globally related template protein in the PDB holo template library, this gives *PoLi* a significant advantage over other LHM based virtual screening algorithms[15, 32].

Experimental demonstration of an 18.4% success rate to identify lead molecules that bind the pharmaceutically relevant target, *E. coli* DHFR, demonstrates the power of the methodology. With 14 total hits, 10 of which are novel, it becomes amply clear that the VS is capable of finding novel analogues from chemical classes that constitute known DHFR inhibitors. Further, the demonstration that the methodology is capable of predicting binders based on a 3-D metric of comparison, as exemplified by NSC117268 and NSC71669, offers a distinct advantage over traditional 2-D comparison and SAR intuition. For example, using 2D fingerprint similarity as the only scoring metric and with same set of templates as input, only 5 of these 14 hits would have ranked among the top 90 predictions that were experimentally validated using differential scanning fluorimetry.

In summary, *PoLi* is a new hybrid approach for virtual screening that has multiple advantages over contemporary approaches. Nevertheless, the somewhat low enrichment of active molecules (EF1%) in the DUD-E database results from the rather small difference between active and decoy molecules. A more elaborate screening procedure that evaluates the interactions made by database molecules in the target binding pocket can provide a potential solution. This type of approach will be examined in future studies.

Materials and Methods

Structure modeling and binding site identification

For each target protein, structural models are generated using the TASSER-VMT[21] automated structure modeling pipeline, wherein template proteins in the non-redundant PDB library are selected using the SP3 threading algorithm[43], followed by multiple TASSER refinement using a variable number of templates and SPICKER clustering[44]. For benchmarking, we removed homologous template proteins from both the threading library and holo template binding site library (described in next section) using a threshold of 30% pairwise sequence identity.

Given a target structure, that can be either modeled or experimental, ligand-binding pockets are predicted using two different approaches. In the first, the superposition matrix from the TM-align[22] global structural alignment is used to overlay template ligands onto the target structure and predict the pockets based on residues that make contact (distance < 4.5Å) with the superimposed ligand. Next, binding pocket similarity of this predicted pocket (in the

target) and original template ligand-binding site is evaluated using the APoc pocket alignment algorithm[24], to filter out cases where even though the receptors share fold similarity (TM-score >0.4), their ligand binding pockets are not similar (APoc *P-value* > 0.001 or PS-score < 0.35). The second approach to predict pockets uses the cavity detection algorithm ConCavity[23] to find pockets. Then, these pockets scan the holo-template binding site library using APoc. Then, ligands of matched pockets in the PDB (with *P-value* < 0.001 and PS-score >0.35) are copied onto the target structure. Finally, all superimposed template ligands are clustered based on their spatial distance, measured from the center of mass of the ligand, using an average-linkage clustering algorithm and a threshold distance of 4.5Å.

Holo templates and small molecule screening library

The holo template library required by *PoLi* for binding pocket prediction and virtual screening was compiled from the May 14, 2014 release of BioLiP database[45]. Downloaded protein-ligand complexes were filtered to remove nucleic acids and small molecules with less than 6 atoms. This filtering process resulted in 40,158 receptors with 44,098 non-redundant ligands and binding sites.

The small molecule screening library is compiled from two different sources. A large fraction (2628 molecules) of this library was compiled from NCI/DTP Open Chemical Repository molecules. In addition, 400 molecules were added using Malaria Box donated by Medicines for Malaria Venture (MMV). A maximum of 200 low energy conformation of these molecules were generated using RDKit conformer generation tool[46] were used for shape-based screening (described below).

PoLi virtual screening pipeline

Figure 1 shows the schematic representation of the *PoLi* pipeline. Starting from the tertiary structure of a protein, the first step is to identify potential small molecule binding sites in the target protein structure. The modus operandi of small molecule binding site prediction in *PoLi* is based on the structural alignment of putative target pockets with a known template ligand binding site. This also enables copying of template ligands in the predicted ligand-binding pocket using the superposition matrix generated during the pocket alignment. Since *PoLi* relies on this binding site comparison to selectively copy template ligands, an advantage of this approach is its ability to copy ligands from protein structures that have different global folds but have similar ligand binding pockets. Up to the top 200 template ligands, ranked based on the harmonic mean of binding pocket similarity (APoc PS-score) and the identity of binding site residues are selected and clustered based on their spatial distance. Then, ligand-based virtual screening uses these selected template ligands.

Ligand pruning and identification of hot-spot regions—Naïvely copying template ligands and using them in virtual screening usually leads to spurious results, as parts of the template ligand that interact with unaligned regions of the template binding site can also be copied. Moreover, since both target and template binding pockets have their own sets of ligand binding residues, even structurally aligned residues in the binding pocket alignment are not always chemically similar and can potentially make disparate interactions. In *PoLi*, these issues are addressed by only copying parts of the template ligand that interact with

template residues that are chemically similar to the aligned target residue. This is performed by first defining template binding site residues, which are at a distance < 4.5 Å of heavy atoms from the ligand. Also a map between the heavy atom index of the template ligand and the residue index of the template receptor is built. Next, an APoc alignment between the template and query pocket is used to define the aligned and unaligned template binding site residues. This is followed by deletion of atoms, which do not make any contact with aligned template residues, with an exception not to delete all atoms that are part of an aromatic/non-aromatic ring if at least one atom of the same ring makes contact with any aligned template residue.

A *hot spot* is defined as the location on the protein that has a high ligand binding propensity. These regions are usually experimentally detected by screening large libraries of fragment-sized organic compounds for binding to target proteins using NMR or X-ray crystallography and identifying regions that have large fragment clusters[47]. Based on a similar concept, we tried to identify parts of template ligand that can make interactions in the *hot-spot* region by clustering pharmacophores of superposed template ligands in order to bias the LIGSIFT structural alignment near these *hot spot* regions. However, it is difficult to detect pharmacophore clusters that can make similar interactions, as the copied template ligands are unaligned to each other (Fig S2B).

We addressed the problem of identifying the *hot spot* region by examining the number of potential interactions that a target binding site residue can make with all copied template ligands (based on its occurrence in binding site alignment with the template residues that interact with ligand). Lets say for a given target protein, we selected P template proteins, and for a given template protein p ($p \in P$) the bound ligand has L atoms. Let T be the set of binding site residues that interact with L and are also conserved (both structurally aligned and chemically similar) in in the APoc binding site alignment. Since, template residue t ($t \in T$) is structurally aligned with target residue q, we assume that template ligand atom a ($a \in L$) can potentially make similar interactions with q. Now, to bias the small molecule structural alignment near *hot spot* regions or the regions that have high propensity to make interactions, a weight h is assigned to each template ligand atom a that can potentially interact with q (Fig S2C), and is defined as:

$$h_a = \sum_{q=1}^{Q} \lambda_{aq} W_q. \quad (1)$$

In equation 1, λ_{aq} is a step function which equals to 1 when atom a is ≤ 4.5 Å from q and 0 otherwise. Q is the set of query binding site residues where $q \in Q$ and W_q is the weight assigned to query binding site residue, and is defined as:

$$W_q = \frac{C_q}{\sum_{q=1}^{Q} C_q}, \text{where}$$
$$C_q = \sum_{p=1}^{P} \sum_{t=1}^{T} \sum_{a=1}^{L} \delta_{qt} I_{at}. \quad (2)$$

In equation 2, C_q is the number of potential interactions that can be made by residue *q*, δ_{qt} is a step function which is equal to *1* when target residue *q* is structurally aligned and chemically similar to template binding site residue *t* and is 0 otherwise. I_{at} is also a step function which is equal to *1* when template ligand atom *a* is at a distance ≤ 4.5Å from residue *t*.

Scoring of database molecules using template ligands—*PoLi* uses a combination of 2D and 3D chemical similarity metrics to score the ligand database molecules. 3D chemical and shape similarity is calculated using a variant of the LIGSIFT algorithm[16], which uses different molecular overlay techniques to find the best volume overlap between template ligand T and database molecule D. Structural superpositions are scored as a shape-density overlap volume (V_{TD}), calculated as the sum of the overlaps of individual atom's Gaussian functions (with similar chemical nature), defined as:

$$V_{TD}=\sum_{i\in T}\sum_{j\in D} h_i\rho_i\rho_j\, exp\left(\frac{\alpha_i\alpha_j d_{ij}^2}{\alpha_i+\alpha_j}\right)\left(\frac{\pi}{\alpha_i+\alpha_j}\right)^{3/2}, \text{where}$$
$$\rho_i(\boldsymbol{r})=\varphi_i\, exp\left\{-\alpha_i(r-R_i)^2\right\}, \text{where } \alpha_i=\pi\left(3\varphi_i/4\pi\sigma_i^3\right)^{2/3} \quad (3)$$

i and *j* are the heavy atom indices, ρ_i and ρ_j are the atomic Gaussian distributions of each atom and d_{ij} is the distance between atom *i* and *j*, α_i is the decay factor, $\phi_i = 2\sqrt{2}$ is the amplitude, R_i is the atomic coordinate for the *i*th atom, σ_i is its van der Waals radius and h_i is the *hot spot* weighting term to reward the overlap near the hot-spot regions in the target. Once the maximum overlap (V_{TD}) is attained, similarity between two molecules is calculated using a ligand size independent scaled Tanimoto Coefficient (sTC), defined as:

$$\text{sTC}=\frac{\text{TC}_{3D}+s_0}{1+s_0}, \text{where}\quad \text{TC}_{3D}=\frac{V_{TD}}{V_T+V_D-V_{TD}}. \quad (4)$$

Here, TC_{3D} is the Tanimoto coefficient (TC) of the 3D shape/chemical similarity, V_T and V_D are the chemical density volume of template molecule T and database molecule D calculated using the Gaussian model, V_{TD} is the molecular volume overlap between molecules T and D, and s_0 is the scaling factor to ensure that the similarity scores of the same statistical significance are size-independent. A combination of shape and chemical similarity in the ratio 1:1 is used for measuring 3D similarity in *PoLi*.

2D chemical similarity between molecules is generally evaluated using the TC of bit fingerprints, defined as:

$$\text{TC}_{2D}=\frac{c}{a+b-c}, \quad (5)$$

where *a* is the count of bits on the 1st string, *b* is the count of bits on the 2nd string and *c* is the count of bits in both strings. In *PoLi*, we use an average Tanimoto Coefficient (*ave*TC) of 1024 bit Daylight-fingerprints generated using OpenBabel [25] API, which is defined as:

$$ave\text{TC}=\frac{\text{TC}+\text{TC'}}{2}, \quad (6)$$

where TC' is Tanimoto coefficient calculated for bits that are set off rather than on in the fingerprints.

Ranking of database molecules—It is a challenging problem to rank database molecules using multiple seed ligands and two different scoring functions without any supervised initial training on the dataset. Therefore, in *PoLi* we adopted an unsupervised data fusion technique, where a fused similarity score *Fsim* of the *y*th database molecule is defined as:

$$FSim_y = \max_{l\in(1,\ldots,N)} \left[d_l \times \left(Z - \text{score}_{ly}^{2D} + Z - \text{score}_{ly}^{3D} \right) \right], \text{where}$$
$$\text{Score}_{ly} = w \frac{\sum_{i=1}^{Nc} Sim(T_i, D_y)}{Nc} + (1-w) \max_{l\in(1,\ldots,Nc)} \left[Sim\left(T_l, D_y\right) \right] \quad (7)$$

In Eq. 7, Z-score$_{ly}$ (2D/3D) is the *Z*-score of similarity between template ligand *l* and database molecule *y*, *N* represents the set of all selected template ligands, d_l is the density of the cluster to which template ligand *l* belongs, *Nc* is the number of template ligands in that cluster, *w* is a weight parameter (defined as *w*=0.3), and *Sim* is the 2D (Eq. 6) or 3D (Eq. 4) similarity score between template ligand (T) and database molecule (D).

Benchmarking sets and evaluation

We have used two types of benchmarking to evaluate *PoLi*. In the first, *in-silico VS* predictions were done on DUD and DUD-E database targets. The DUD database contains 40 target proteins with active and decoys molecules in the ratio of 1:36; while DUD-E database contains a list of 102 targets with an average of 224 active molecules and 50 decoys for each active molecule. For validation, we have used 40 proteins listed in the DUD database and 65 targets of DUD-E database. 37 proteins of DUD-E set that were already included in DUD set were not included to avoid redundancy. Moreover, both experimental and modeled receptor structures of these proteins have been used to objectively evaluate the effect of model quality on virtual screening performance.

The performance of *PoLi* in these *in-silico* virtual screening runs is evaluated using standard evaluation metrics: (a) The Receiver Operating Characteristic (ROC) curve and (b) The Enrichment Factor of the screened database and (c) The Hit Rate (HR). The ROC curve depicts the true positive rate as a function of false positive rate, and the area under the curve (AUC) is used to quantify the shape of the ROC curve. AUC values range between [0–1], where an AUC below 0.5 is equivalent to random performance. Much more important metrics for practical purposes are measures like the Enrichment Factor (EF) and Hit Rate (HR) that are used to evaluate the performance in the top *x*% of the screened library, where the EF is defined as:

$$EF^{x\%} = \frac{\text{No. of True Positives}^{x\%} / N_{selected}^{x\%}}{N_{actives} / N_{total}}, \quad (8)$$

where *x* represents fraction of screened library and is set to 1%, 5% and 10% to analyze the performance for a broad range of screened molecules in the database. We have also used HR as an evaluation metric, which defined as:

$$HR^{x\%} = \frac{EF^{x\%}_{actual}}{EF^{x\%}_{ideal}} \times 100, \quad (9)$$

The second set of experiments simulates the real world scenario, where we use the *PoLi* pipeline to generate ligand binding predictions for *E. coli* DHFR, while excluding all template proteins with >30% sequence identity to the target protein. Top ranked predictions in our small molecule library are then experimentally validated using high-throughput differential scanning fluorimetry (described below).

Experimental validation using differential scanning fluorimetry

Reagents—All reagents and chemicals, unless mentioned otherwise, were procured from Sigma-Aldrich (St. Louis, MO) with the following exceptions: HEPES, pH 7.3 buffer was obtained from Fischer Bioreagents and dimethyl sulfoxide (DMSO) from MP Biomedicals LLC. Sypro orange dye was obtained from Invitrogen (Carlsbad, CA). 96-well PCR-plates and plate seals were from Eppendorf (Eppendorf, NY, USA). *E. coli* dihydrofolate reductase, DHFR, was provided by Prof. Eugene Shakhnovich, Harvard University. The library of small molecules and drugs containing oncology drug set III (97 compounds), mechanistic set II (816 compounds) diversity set III (synthetic) (1597 compounds) and natural product set (118 compounds) were provided by the open chemical repository of Developmental Therapeutics Program (DTP) of the National Cancer Institute (NCI), National Institutes of Health (NIH) (http://dtp.cancer.gov). Furthermore, a set of 400 diverse drug-like and probe-like compounds was provided as 10 mM stock solutions in dimethyl sulfoxide by Medicines for Malaria Venture (MMV) (http://www.mmv.org/malariabox). All provided compounds had been demonstrated to possess antimalarial activity against the blood-stage of *P. falciparum* and were selected to represent structural diversity, ease of oral absorption and minimum toxicity.

Acquisition and quantification of thermal shift assays—High-throughput thermal shift assays were carried out following established guidelines[48, 49]. Briefly, samples were aliquoted in 96-well PCR plates and protein melting curves were obtained by heating the samples from 25 °C to 74 °C using a 1 °C/min heating ramp in a RealPlex quantitative PCR instrument (Eppendorf, NY, USA), with Sypro orange dye (Invitrogen) as the extrinsic fluorescent reporter. A uniform final concentration of 5X was used in all experiments. The dye was excited at 465 nm and emission recorded at 580 nm using the instrument's filters. One data point was acquired for each degree increment. Unfolding was carried out in a total reaction volume of 20 μl, with 100 mM HEPES pH 7.3, 150 mM NaCl and 5 μM *E. coli* DHFR. Appropriate dye and protein controls were included in each plate as an internal reference. All experiments were done with experimental replicates, with the mean value considered for further analysis. Furthermore, the curves obtained were processed to subtract the background signal from dye alone or dye-small molecule controls.

Each melting curve was assigned a quality score (Q), the ratio of the melting-associated increase in fluorescence (ΔF_{melt}) to the total fluorescence range (ΔF_{total}). Q = 1 is a high-quality curve, while Q = 0 indicates no thermal transition[49].

Data analysis—The validity of the PoLi's top 90 predictions on E. coli dihydrofolate reductase was assessed by the thermal melt assay methodology. Protein unfolding curves showing a single sigmoidal thermal transition were selected and normalized for further analysis. Initially, the curves were fit to Boltzmann's equation (Eq. 10) to obtain the melting temperature, T_m, from the observed fluorescence intensity, I by:

$$I = I_{min} + \frac{I_{max} - I_{min}}{1 + e\left(\frac{T_m - T}{a}\right)}, \quad (10)$$

where I_{min} and I_{max} are the minimum and maximum intensities; *a* denotes the slope of the curve at the unfolding transition midpoint temperature, T_m. However, due to unfolding-associated aggregation of the protein that resulted in decreasing SO fluorescence at higher temperatures, the fits were unconvincing giving wide margins of error (Fig. 4A–D). To overcome this problem and to estimate the melting temperature more accurately, the first derivative of each melting curve was derived and fit to a Gaussian whose mean gave an accurate estimate of the T_m (Fig 4E–H). The fluorescence intensity was used to compute the fraction unfolded (f_u) and approximate thermodynamic parameters were estimated by van't Hoff [50] and Gibbs-Helmholtz analyses[51]. Further, rough estimates of ligand-binding affinity at T_m were computed by employing Equation 10[52], with slight modifications.

$$K_L(T_m) = \frac{e\left\{\frac{-\Delta H}{R}\left(\frac{1}{T_m} - \frac{1}{T_0}\right)\right\}}{[L]}, \quad (11)$$

where $K_L(T_m)$is the ligand association constant and [L] is the free ligand concentration at T_m ($[L_{Tm}] \sim [L]_{total}$, when $[L]_{total} >>$ the total concentration of protein. K_D is the inverse of $K_L(T_m)$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This project was funded by GM-37408 of the Division of General Medical Sciences of the NIH. The authors wish to thank Prof. Eugene Shakhnovich, Harvard University, for providing purified *E. coli* DHFR protein. We would also like to thank the Developmental Therapeutics Program of the National Cancer Institute and the Medicines for Malaria Venture (MMV) for providing the small molecules used in this study.

Abbreviations

VS	Virtual screening
HTS	High Throughput Screening

DSF	Differential Scanning Fluorimetry
DHFR	Dihydrofolate reductase

References

1. Hung CL, Chen CC. Computational Approaches for Drug Discovery. Drug development research. 2014; 75:412–418. [PubMed: 25195585]
2. Shoichet BK. Screening in a Spirit Haunted World. Drug discovery today. 2006; 11:607–615. [PubMed: 16793529]
3. The Academic Pursuit of Screening. Nature chemical biology. 2007; 3:433. [PubMed: 17637766]
4. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-Based Virtual Screening for Drug Discovery: A Problem-Centric Review. The AAPS journal. 2012; 14:133–141. [PubMed: 22281989]
5. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. Nature reviews. Drug discovery. 2004; 3:935–949. [PubMed: 15520816]
6. McGovern SL, Shoichet BK. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. Journal of medicinal chemistry. 2003; 46:2895–2907. [PubMed: 12825931]
7. Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. Journal of medicinal chemistry. 2004; 47:45–55. [PubMed: 14695819]
8. Sousa SF, Fernandes PA, Ramos MJ. Protein-Ligand Docking: Current Status and Future Challenges. Proteins. 2006; 65:15–26. [PubMed: 16862531]
9. Hall DR, Ngan CH, Zerbe BS, Kozakov D, Vajda S. Hot Spot Analysis for Driving the Development of Hits into Leads in Fragment-Based Drug Discovery. Journal of chemical information and modeling. 2012; 52:199–209. [PubMed: 22145575]
10. Verdonk ML, Giangreco I, Hall RJ, Korb O, Mortenson PN, Murray CW. Docking Performance of Fragments and Druglike Compounds. Journal of medicinal chemistry. 2011; 54:5422–5431. [PubMed: 21692478]
11. Kolb P, Kipouros CB, Huang D, Caflisch A. Structure-Based Tailoring of Compound Libraries for High-Throughput Screening: Discovery of Novel Ephb4 Kinase Inhibitors. Proteins. 2008; 73:11–18. [PubMed: 18384152]
12. Willett P. Similarity-Based Virtual Screening Using 2d Fingerprints. Drug discovery today. 2006; 11:1046–1053. [PubMed: 17129822]
13. Taminau J, Thijs G, De Winter H. Pharao: Pharmacophore Alignment and Optimization. Journal of molecular graphics & modelling. 2008; 27:161–169. [PubMed: 18485770]
14. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA, Nevins N, Jain AN, Kelley B. Molecular Shape and Medicinal Chemistry: A Perspective. Journal of medicinal chemistry. 2010; 53:3862–3886. [PubMed: 20158188]
15. Zhou H, Skolnick J. Findsite(Comb): A Threading/Structure-Based, Proteomic-Scale Virtual Ligand Screening Approach. Journal of chemical information and modeling. 2013; 53:230–240. [PubMed: 23240691]
16. Roy A, Skolnick J. Ligsift: An Open-Source Tool for Ligand Structural Alignment and Virtual Screening. Bioinformatics. 2015; 31:539–544. [PubMed: 25336501]
17. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. Chembl: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic acids research. 2012; 40:D1100–1107. [PubMed: 21948594]

18. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. Drugbank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. Nucleic acids research. 2006; 34:D668–672. [PubMed: 16381955]
19. Huang N, Shoichet BK, Irwin JJ. Benchmarking Sets for Molecular Docking. Journal of medicinal chemistry. 2006; 49:6789–6801. [PubMed: 17154509]
20. Skolnick J, Gao M. Interplay of Physics and Evolution in the Likely Origin of Protein Biochemical Function. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:9344–9349. [PubMed: 23690621]
21. Zhou H, Skolnick J. Template-Based Protein Structure Modeling Using Tasser(Vmt.). Proteins. 2012; 80:352–361. [PubMed: 22105797]
22. Zhang Y, Skolnick J. Tm-Align: A Protein Structure Alignment Algorithm Based on the Tm-Score. Nucleic acids research. 2005; 33:2302–2309. [PubMed: 15849316]
23. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3d Structure. PLoS computational biology. 2009; 5:e1000585. [PubMed: 19997483]
24. Gao M, Skolnick J. Apoc: Large-Scale Identification of Similar Protein Pockets. Bioinformatics. 2013; 29:597–604. [PubMed: 23335017]
25. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An Open Chemical Toolbox. Journal of cheminformatics. 2011; 3:33. [PubMed: 21982300]
26. Soding J. Protein Homology Detection by Hmm-Hmm Comparison. Bioinformatics. 2005; 21:951–960. [PubMed: 15531603]
27. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (Dud-E): Better Ligands and Decoys for Better Benchmarking. Journal of medicinal chemistry. 2012; 55:6582–6594. [PubMed: 22716043]
28. Zhang Y, Skolnick J. Scoring Function for Automated Assessment of Protein Structure Template Quality. Proteins. 2004; 57:702–710. [PubMed: 15476259]
29. Zhang Y, Devries ME, Skolnick J. Structure Modeling of All Identified G Protein-Coupled Receptors in the Human Genome. PLoS computational biology. 2006; 2:e13. [PubMed: 16485037]
30. Perola E, Walters WP, Charifson PS. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. Proteins. 2004; 56:235–249. [PubMed: 15211508]
31. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A Critical Assessment of Docking Programs and Scoring Functions. Journal of medicinal chemistry. 2006; 49:5912–5931. [PubMed: 17004707]
32. Feinstein WP, Brylinski M. Efindsite: Enhanced Fingerprint-Based Virtual Screening against Predicted Ligand Binding Sites in Protein Models. 2014; 33:135–150.
33. Trott O, Olson AJ. Autodock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. Journal of computational chemistry. 2010; 31:455–461. [PubMed: 19499576]
34. Hughes TS, Giri PK, de Vera IM, Marciano DP, Kuruvilla DS, Shin Y, Blayo AL, Kamenecka TM, Burris TP, Griffin PR, Kojetin DJ. An Alternate Binding Site for Ppargamma Ligands. Nature communications. 2014; 5:3571.
35. Schweitzer BI, Dicker AP, Bertino JR. Dihydrofolate Reductase as a Therapeutic Target. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 1990; 4:2441–2452. [PubMed: 2185970]
36. Yuthavong Y, Tarnchompoo B, Vilaivan T, Chitnumsub P, Kamchonwongpaisan S, Charman SA, McLennan DN, White KL, Vivas L, Bongard E, Thongphanchang C, Taweechai S, Vanichtanankul J, Rattanajak R, Arwon U, Fantauzzi P, Yuvaniyama J, Charman WN, Matthews D. Malarial Dihydrofolate Reductase as a Paradigm for Drug Development against a Resistance-Compromised Target. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:16823–16828. [PubMed: 23035243]

37. Srinivasan B, Zhou H, Kubanek J, Skolnick J. Experimental Validation of Findsite(Comb) Virtual Ligand Screening Results for Eight Proteins Yields Novel Nanomolar and Micromolar Binders. Journal of cheminformatics. 2014; 6:16. [PubMed: 24936211]
38. Srinivasan, B.; Skolnick, J.; Zhou, H. Molecules with Potent Dhfr Binding Affinity and Antibacterial Activity. 2014. 20140329840
39. Srinivasan B, Skolnick J. Insights into the Slow-Onset Tight-Binding Inhibition of Escherichia Coli Dihydrofolate Reductase: Detailed Mechanistic Characterization of Pyrrolo [3,2-F] Quinazoline-1,3-Diamine and Its Derivatives as Novel Tight-Binding Inhibitors. The FEBS journal. 2015; 282:1922–1938. [PubMed: 25703118]
40. Stone SR, Montgomery JA, Morrison JF. Inhibition of Dihydrofolate Reductase from Bacterial and Vertebrate Sources by Folate, Aminopterin, Methotrexate and Their 5-Deaza Analogues. Biochemical pharmacology. 1984; 33:175–179. [PubMed: 6367748]
41. McCormack JJ. Dihydrofolate Reductase Inhibitors as Potential Drugs. Medicinal research reviews. 1981; 1:303–331. [PubMed: 7050566]
42. Zhou H, Skolnick J. Findsite(X): A Structure-Based, Small Molecule Virtual Screening Approach with Application to All Identified Human Gpcrs. Molecular pharmaceutics. 2012; 9:1775–1784. [PubMed: 22574683]
43. Zhou H, Zhou Y. Fold Recognition by Combining Sequence Profiles Derived from Evolution and from Depth-Dependent Structural Alignment of Fragments. Proteins. 2005; 58:321–328. [PubMed: 15523666]
44. Zhang Y, Skolnick J. Spicker: A Clustering Approach to Identify near-Native Protein Folds. Journal of computational chemistry. 2004; 25:865–871. [PubMed: 15011258]
45. Yang J, Roy A, Zhang Y. Biolip: A Semi-Manually Curated Database for Biologically Relevant Ligand-Protein Interactions. Nucleic acids research. 2013; 41:D1096–1103. [PubMed: 23087378]
46. Landrum, G. Rdkit: Open-Source Cheminformatics.
47. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. Fragment-Based Identification of Druggable 'Hot Spots' of Proteins Using Fourier Domain Correlation Techniques. Bioinformatics. 2009; 25:621–627. [PubMed: 19176554]
48. Niesen FH, Berglund H, Vedadi M. The Use of Differential Scanning Fluorimetry to Detect Ligand Interactions That Promote Protein Stability. Nature protocols. 2007; 2:2212–2221. [PubMed: 17853878]
49. Crowther GJ, He P, Rodenbough PP, Thomas AP, Kovzun KV, Leibly DJ, Bhandari J, Castaneda LJ, Hol WG, Gelb MH, Napuli AJ, Van Voorhis WC. Use of Thermal Melt Curves to Assess the Quality of Enzyme Preparations. Analytical biochemistry. 2010; 399:268–275. [PubMed: 20018159]
50. John DM, Weeks KM. Van't Hoff Enthalpies without Baselines. Protein science : a publication of the Protein Society. 2000; 9:1416–1419. [PubMed: 10933511]
51. LiCata VJ, Liu CC. Analysis of Free Energy Versus Temperature Curves in Protein Folding and Macromolecular Interactions. Methods in enzymology. 2011; 488:219–238. [PubMed: 21195230]
52. Lo MC, 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. Analytical biochemistry. 2004; 332:153–159. [PubMed: 15301960]

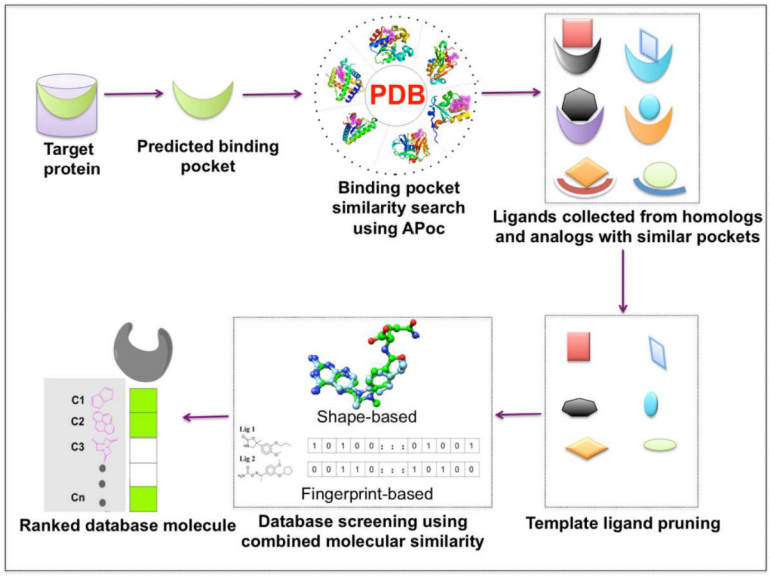

Figure 1.
Schematic flowchart of the *PoLi* virtual screening pipeline.

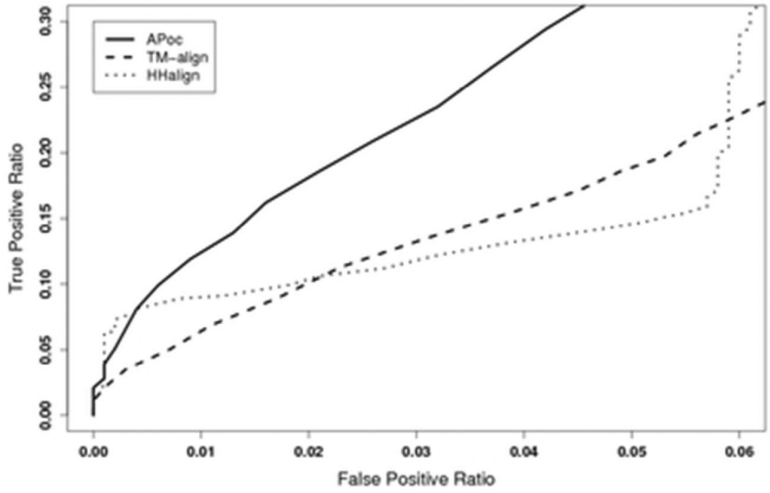

Figure 2.

Comparison of a pocket based method (APoc) with global structure alignment and homology based approaches to detect similar ligands. The benchmark shows the ability of different approaches to recognize 30,000 pairs of similar ligands from 35,000 pairs of chemically dissimilar ligands.

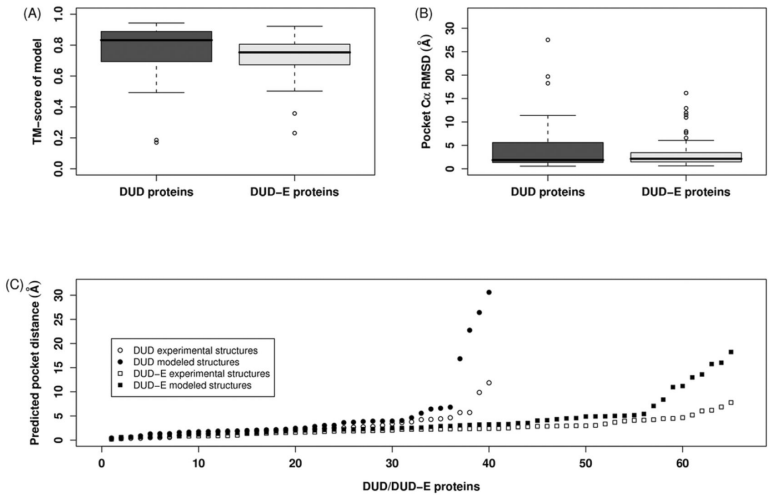

Figure 3.

Structure quality and binding site prediction accuracy for DUD and DUD-E proteins. Box and whiskers plot of (A) TM-score and (B) ligand binding pocket C_α RMSD of TASSER models to the experimentally determined structures. (C) Distance between the geometric center of the ligand in the co-crystallized complex and the center of the best predicted ligand-binding pocket in the 40 DUD and 65 DUD-E protein targets.

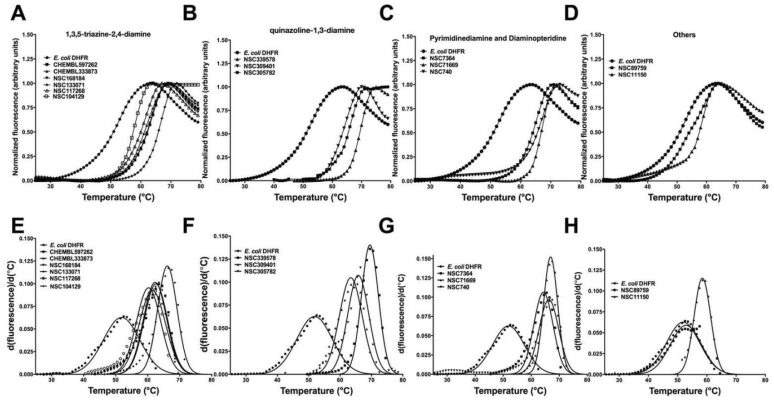

Figure 4.

Thermal unfolding curves of *E. coli* DHFR **A)** Primary unfolding curves for hits belonging to the 1,3,5 triazine-2, 4-diamine group **B)** Primary unfolding curves for hits belonging to the quinazoline-1, 3-diamine group **C)** Primary unfolding curves for hits belonging to the pyrimidinediamine and aminopteridine group **D)** Primary unfolding curves for hits belonging to chemical classes distinct from any reported DHFR inhibitors **E)** Gaussian fit of first-derivative for curves in (A) **F)** Gaussian fit of first-derivative for curves in (B) **G)** Gaussian fit of first-derivative for curves in (C) **H)** Gaussian fit of first-derivative for curves in (D). On the plots A-D, the y-axis represents the normalized fluorescence and the x-axis represents the temperature in degrees Celsius. The experimental data points were fit to the respective equations using the nonlinear curve-fitting algorithm of GraphPad Prism v 6.0e.

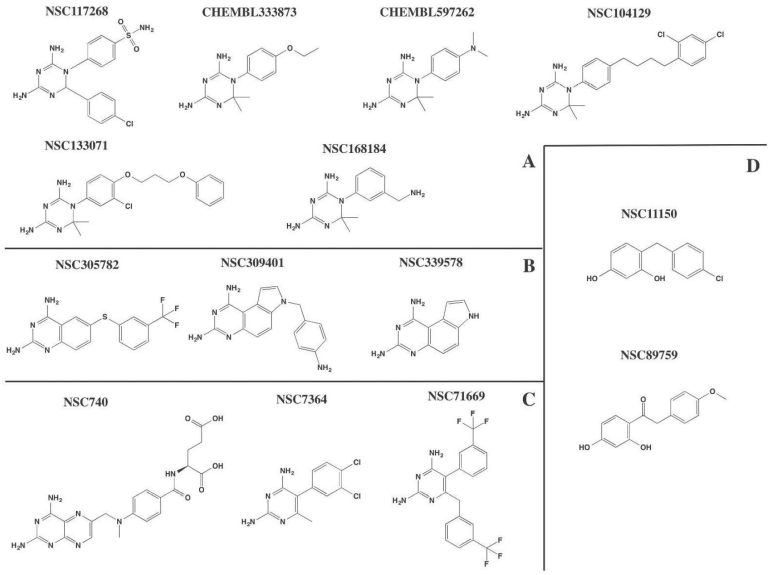

Figure 5.

Structures of small molecules showing binding to *E. coli* DHFR as assessed by the thermal shift assay methodology **A)** 1,3,5-triazine-2,4-diamine derivatives **B)** quinazoline-1,3-diamine derivatives **C)** Pyrimidinediamine and diaminopetridine derivatives **D)** 2,4 dihydroxyphenyl derivatives. The SDF files for the small molecules were downloaded from Pubchem (http://pubchem.ncbi.nlm.nih.gov) and the figure was generated using ChemBioDraw 14.0.

Table 1

Virtual screening performance of *PoLi* on 40 DUD targets and 65 DUD-E targets using experimental and modeled structures.

Target receptor	AUC *av* | *sd*	$EF^{1\%}$ *av* | *sd*	$EF^{5\%}$ *av* | *sd*	$EF^{10\%}$ *av* | *sd*	$HR^{1\%}$ *av* | *sd*	$HR^{5\%}$ *av* | *sd*	$HR^{10\%}$ *av* | *sd*
DUD-E (LIGSIFT)	0.73±0.20	18.7±18.1	6.6±4.4	4.2±2.3	29.2±25.7	20.0±14.1	25.3±14.6
DUD-E (exp.)	0.72±0.16	9.6±13.5	4.8±4.4	3.4±2.3	14.6±19.1	14.0±13.9	20.1±15.1
DUD-E (model TM> 0.5)	0.74±0.16	9.9±13.5	4.9±4.3	3.6±2.3	14.9±18.8	14.5±13.4	21.4±15.3
DUD-E(model pdist<5 Å)	0.76±0.16	11.0±14.3	5.4±4.4	3.9±2.3	16.6±19.9	15.9±13.9	23.4±15.6
DUD-E (model)	0.73±0.16	**9.6±12.7**	4.7±4.2	3.6±2.3	**14.3±17.0**	13.9±13.2	21.1±15.3
DUD (LIGSIFT)	0.77±0.20	17.4±11.1	7.8±5.4	4.7±2.7	49.4±31.5	39.2±27.1	47.2±27.5
DUD (exp.)	0.78±0.19	15.2±11.4	7.2±5.2	4.7±2.7	43.3±32.4	31.9±24.4	41.3±26.0
DUD (model TM> 0.5)	0.78±0.18	14.1±10.1	7.2±4.7	4.6±2.7	40.1±28.8	31.8±22.1	40.7±25.5
DUD (model pdist< 5Å)	0.80±0.19	15.9±9.9	8.1±4.7	5.1±2.7	45.2±28.3	34.9±21.8	44.1±25.4
DUD (model)	0.78±0.18	**13.4±10.3**	7.0±4.8	4.6±2.9	**38.0±29.3**	30.7±22.3	39.5±26.1

av: average; *sd*: standard deviation; exp: experimentally determined structure; TM: TM-score of model to experimental structure; pdist: distance between predicted pocket and center of mass of ligand in crystal structure

Table 2

Analysis of molecular similarity scores between database molecules and template ligands to understand the decrease in performance of *PoLi* on DUD-E database.

Database	Receptor structure	3D similarity		2D similarity	
		Actives *av* | *sd*	Decoys *av* | *sd*	Actives *av* | *sd*	Decoys *av* | *sd*
DUD-E	Experimental	0.52±0.06	0.50±0.04	0.61±0.09	0.57±0.06
	Model	0.52±0.06	0.50±0.04	0.61±0.09	0.56±0.06
DUD	Experimental	0.58±0.09	0.53±0.05	0.62±0.09	0.58±0.06
	Model	0.57±0.06	0.53±0.04	0.62±0.09	0.57±0.06

av: average; *sd*: standard deviation

Table 3

Performance of *PoLi*, LIGSIFT, FINDSITE^{filt} and AutoDock Vina on DUD and 65 DUD-E targets using modeled structures. *PoLi*^{unbiased} performance is obtained without performing biased structural overlap in hot-spot regions.

Method	AUC *av* | *sd*	EF$^{1\%}$ *av* | *sd*	EF$^{5\%}$ *av* | *sd*	EF$^{10\%}$ *av* | *sd*	HR$^{1\%}$ *av* | *sd*	HR$^{5\%}$ *av* | *sd*	HR$^{10\%}$ *av* | *sd*
DUD-E database							
AutoDockVina	0.60±0.13	3.0±2.8	2.1±1.5	1.8±1.0	4.9±4.7	6.3±4.8	11.1±6.9
FINDSITE^{filt}	0.69±0.16	7.9±11.3	3.9±3.8	2.9±2.2	12.1±17.0	11.2±11.6	16.8±13.1
LIGSIFT	0.67±0.14	5.9±8.7	3.5±3.4	2.7±1.9	8.9±11.1	10.3±10.1	15.8±12.3
PoLi ^{unbiased}	0.72±0.17	9.7±13.5	4.8±4.2	3.5±2.3	14.4±17.4	14.0±13.0	21.0±15.3
PoLi	0.73±0.16	**9.6±12.7**	4.7±4.2	3.6±2.3	**14.3±17.0**	13.9±13.2	21.1±15.3
DUD database							
AutoDockVina	0.50±0.16	1.6±2.2	1.5±1.3	1.3±1.0	4.6±6.1	6.7±5.9	11.6±9.3
FINDSITE^{filt}	0.70±0.20	9.0±10.3	4.4±4.5	3.1±2.5	25.8±29.4	20.8±22.5	28.8±25.4
LIGSIFT	0.71±0.20	11.8±11.5	5.4±4.6	3.7±2.6	33.3±32.9	23.0±19.9	31.3±21.9
PoLi ^{unbiased}	0.77±0.19	12.3±10.4	6.6±4.8	4.5±2.9	35.0±29.8	29.0±22.2	39.3±26.9
PoLi	0.78±0.18	**13.4±10.3**	7.0±4.8	4.6±2.9	**38.0±29.3**	30.7±22.3	39.5±26.1

av: average; *sd*: standard deviation

Table 4

Performance of *PoLi* on DUD-E and DUD database using templates with similar fold and those with random structure similarity (TM-score < 0.4).

Templates	AUC *av* | *sd*	$EF^{1\%}$ *av* | *sd*	$EF^{5\%}$ *av* | *sd*	$EF^{10\%}$ *av* | *sd*	$HR^{1\%}$ *av* | *sd*	$HR^{5\%}$ *av* | *sd*	$HR^{10\%}$ *av* | *sd*
DUD-E database							
Same fold	0.71±0.17	8.7±13.3	4.3±4.2	3.2±2.5	12.9±17.4	12.7±12.5	19.0±15.6
Unrelated fold	0.62±0.15	**2.7±4.1**	2.1±2.2	1.9±1.6	4.3±6.6	6.2±6.5	11.4±10.3
Same fold*	0.75±0.16	10.5±14.0	5.2±4.2	3.8±2.3	15.6±18.0	15.3±12.2	22.9±14.4
Combined*	0.74±0.16	10.0±13.1	4.8±4.0	3.7±2.3	14.7±16.5	14.1±11.4	21.7±14.2
Failed(combined)	0.69±0.25	7.3±10.4	4.1±5.3	3.0±2.8	12.5±19.4	12.7±19.9	18.0±20.5
DUD database							
Same fold	0.74±0.21	12.0±11.0	6.3±5.3	4.2±3.3	33.9±31.0	27.5±23.5	35.9±29.5
Unrelated fold	0.68±0.20	**7.1±8.7**	4.3±4.3	3.1±2.5	20.0±25.0	19.2±20.6	27.9±24.1
Same fold#	0.80±0.20	15.0±10.4	7.9±4.8	5.2±2.9	42.3±29.4	34.4±21.6	44.9±26.7
Combined#	0.80±0.16	15.9±10.0	7.7±4.7	5.0±2.6	44.9±28.5	33.5±22.2	43.3±25.3
Failed(combined)	0.70±0.27	4.0±4.6	3.8±3.9	3.1±2.4	11.1±12.7	18.2±19.0	29.3±23.7

av: average; *sd*: standard deviation;

*average over 54 DUD-E targets with predictions generated using similar fold template;

#average over 32 DUD targets with predictions generated using similar fold template; Combined: Predictions generated using both similar and unrelated fold templates; Failed: Proteins targets where no predictions could be generated due to lack of similar pockets in similar fold templates.

Table 5

Pocket specific predictions **by** *PoLi* on DUD-E and DUD database.

Pocket (# protein)	AUC *av* | *sd*	$EF^{1\%}$ *av* | *sd*	$EF^{5\%}$ *av* | *sd*	$EF^{10\%}$ *av* | *sd*	$HR^{1\%}$ *av* | *sd*	$HR^{5\%}$ *av* | *sd*	$HR^{10\%}$ *av* | *sd*
DUD-E database							
Pocket 1 (65)	0.74±0.13	9.4±13.1	4.7±4.1	3.5±2.3	14.2±17.0	14.0±4.8	20.9±15.3
Pocket 2 (45)	0.64±0.13	2.5±3.1	2.1±2.1	1.9±1.4	4.1±5.5	6.4±7.9	11.6±10.2
Pocket 3 (27)	0.60±0.16	2.5±3.5	2.2±1.9	1.9±1.6	3.8±5.2	6.2±5.6	11.1±9.4
Pocket 4 (12)	0.66±0.13	2.4±2.9	2.5±1.8	2.4±1.5	3.9±5.0	7.5±6.4	14.1±10.1
Pocket 5 (6)	0.60±0.18	3.5±5.3	2.6±4.3	2.0±2.5	5.0±7.1	7.7±13.2	12.0±15.5
DUD database							
Pocket 1 (40)	0.77±0.15	13.3±9.1	6.8±4.4	4.5±2.7	37.7±26.1	30.0±21.8	38.8±25.5
Pocket 2 (36)	0.64±0.18	2.3±4.0	2.4±3.0	2.2±2.2	6.5±11.6	11.8±15.2	21.3±22.5
Pocket 3 (23)	0.63±0.21	4.1±5.7	2.8±3.3	2.3±2.3	11.6±16.0	13.2±16.0	22.0±22.9
Pocket 4 (14)	0.65±0.22	3.3±9.7	2.8±3.6	2.5±2.1	9.1±26.9	13.3±17.9	24.2±21.8
Pocket 5 (8)	0.74±0.13	4.9±5.9	3.9±3.1	3.3±2.1	14.3±17.4	18.7±16.0	31.5±21.4

Table 6

Summary of virtual ligand screening, thermal shift assay and binding parameters for the hits obtained on *E. coli* DHFR.

Identity	Rank	Rank^{2D}	Q$^{\#}$	T_m (° C)	ΔT_m (° C)	K_D (μM)^c
Protein	NA	NA	1.00	51.9	NA	NA
NSC339578*	6	777	0.35	69.5	17.6	02.4 ± 0.6
NSC71669	75	863	0.32	66.9	15.0	05.2 ± 1.3
NSC305782	46	1485	0.20	66.3	14.4	06.2 ± 1.2
NSC740*	18	674	0.34	66.3	14.4	06.2 ± 1.6
NSC133071	25	119	0.41	66.2	14.3	06.4 ± 1.5
NSC7364*	5	1303	0.43	64.4	12.5	10.8 ± 2.1
NSC309401*	7	129	0.31	63.6	11.7	13.7 ± 1.8
CHEMBL597262	1	41	0.42	62.6	10.7	18.4 ± 2.7
NSC168184	3	109	0.23	62.4	10.5	19.5 ± 3.5
CHEMBL333873	2	90	0.31	61.5	9.6	25.6 ± 3.8
NSC117268	60	254	0.43	60.4	8.5	35.7 ± 6.3
NSC11150	77	69	0.50	58.4	6.5	65.6 ± 11.1
NSC104129	10	80	0.32	57.3	5.4	91.9 ± 14.0
NSC89759	51	66	0.30	55.1	3.2	182.1± 21.6

*Indicates reported inhibitors of DHFR independently picked up by *PoLi* and validated experimentally. Rank^{2D} is the rank of identified inhibitors using 2D fingerprint similarity (TC) using same set of templates as used by *PoLi*,

#quality score (Q) is the ratio of the melting-associated increase in fluorescence (ΔF_{melt}) and total range in fluorescence (ΔF_{total}). A Q value of 1 represents a high-quality curve, while a value of 0 shows an absence of melting as described earlier[49]. K_D^C is the dissociation constant computed from the magnitude thermal shifts obtained relative to the protein alone.