PNAS
www.pnas.org

Supplementary Information for

Leveraging non-structural data to predict structures and affinities of protein–ligand complexes

Joseph M. Paggi, Julia A. Belk*, Scott A. Hollingsworth*, Nicolas Villanueva, Alexander S. Powers, Mary J. Clark, Augustine G. Chemparathy, Jonathan E. Tynan, Thomas K. Lau, Roger K. Sunahara, Ron O. Dror

*** These authors contributed equally to this work.**

Ron O. Dror

Email: ron.dror@stanford.edu

This PDF file includes:

Supplementary Information Text

Assembly of data for use in learning the ComBind scoring function

Curation of experimental protein–ligand complex structures

In order to learn the ComBind scoring function, we curated a set of protein–ligand complex structures representing each of the major drug targets catalogued by Santos *et al.*, 2017 (**Supplementary Table 2**). This set of target proteins was chosen through a combination of manual curation and adaptation of the PDBbind refined set (1). For each target, we included up to 21 structures, each with a distinct ligand bound, selecting the structures with alphabetically lowest PDB code when more than 21 were available. Structures with duplicate ligands, mutant proteins, or no small molecule in the orthosteric site were excluded.

Preparation of protein–ligand complex structures and ligands

For all of the results presented in this study, we performed "cross-docking." Specifically, for each target, we chose the structure with the alphabetically first PDB code as the input 3D structure of the protein and then docked other ligands to this reference structure. This simulates a real-world application where only one structure of the target protein is available, and the user wants to predict poses for ligands not present in that structure.

To prepare protein structures for use in docking, we first prepared structures using the Schrodinger suite. All waters were removed, the tautomeric state of the ligand present in the experimentally determined structure was assigned using Epik at pH 7.0 +/– 2.0, hydrogen bonds were optimized, and energy minimization was performed with non-hydrogen atoms constrained to an RMSD of less than 0.3 Å from the initial structure. The ligand was then removed.

For ligands to be docked, the tautomeric state was assigned using Epik tool at target pH 7.0 +/– 2.0. The single most favorable state was considered for docking. Torsion angles were randomized before docking.

Skilled chemists can often improve the overall success rate of docking through careful manual preparation of the protein structure—for example, by diligent placement of waters or consideration of side chain rotamers. Such a procedure is subjective and was thus not employed in our performance benchmarks. In our experience, however, careful manual preparation of target structures improves the results of ComBind and ComBindVS even more than those of per-ligand docking methods, because such preparation increases the accuracy of the helper ligand poses and thus the value of the information gleaned from them.

Generation of docked poses

Here we use the commercial docking software package Glide (2, 3) to generate candidate poses and assign a per-ligand score to each. We selected Glide because it is widely used in the pharmaceutical industry and because it ranks among the most accurate docking packages in comparative studies (4, 5). We emphasize, however, that the ComBind approach can utilize any per-ligand scoring function and pose sampling strategy, including those implemented in any standard docking package.

Ligands were docked using default Glide SP settings except that "Enhanced Sampling" was set to 4, quadrupling the number of ligand conformers considered. For each ligand, we produced up to the 100 most highly ranked poses (for some ligands fewer than 100 poses passed Glide's

internal filters). We also considered using Glide XP but found that Glide XP produced a correct candidate pose substantially less often than Glide SP (**Supplementary Table 1**). Glide XP and SP performed similarly in terms of how frequently the top-ranked pose is correct. Additionally, we considered using Induced Fit Docking (IFD). While IFD produced at least one correct candidate pose more often than Glide SP, the performance in terms of how often the top-ranked pose is correct was worse.

A variety of "flexible receptor docking" methods have been developed that allow deformation of the target protein when sampling ligand poses (5-7). These methods have proven highly valuable in cases where the user knows in advance that protein flexibility is important to binding of the query ligand. When used as fully automated pose prediction methods without such prior information, however, flexible receptor docking methods frequently underperform rigid docking methods such as Glide, as observed in our benchmarks of the popular Induced Fit Docking method (6) (**Supplementary Table 1)** and reported previously for other flexible receptor docking methods (7). Such methods are more likely to sample a correct pose but also more likely to sample incorrect poses that outscore correct poses, although recent work has shown substantial improvement (8).

Determining the quality of docked poses

The accuracy of each pose was quantified by the non-hydrogen-atom RMSD from the experimentally determined pose. To compute the RMSD, each complex was aligned to the structure used for docking based on non-hydrogen-atoms within 15 Å of the ligand, and the RMSD was then computed between the docked pose and the same ligand's pose in the aligned complex. We denote poses at most 2.0 Å RMSD from their aligned experimentally determined pose as being "near-native" or "correct."

Quantifying the similarity of binding poses for distinct ligands

Protein–ligand interaction similarity

Three interaction types were considered: hydrogen bonds, salt bridges, and hydrophobic contacts. We designed quantitative measures to assess the presence of these interactions between the ligand and a given protein residue (**Supplementary Table 5**). The hydrogen bond and salt bridge interaction measures were designed to give a value of 1 for interactions meeting established criteria (9). A soft boundary was added to give borderline cases values between 0 and 1, in order to prevent discontinuities. The hydrophobic contact measure approximates the hydrophobic surface contact area by considering the number of protein–ligand atom pairs in contact with each other. Again, a soft boundary (in this case, between an atom pair being or not being in contact) was used to prevent very similar poses from leading to very different values. We denote the interaction value for interaction type k, for pose ℓ_i of ligand i, with protein residue r as $X_r^{(k)}(\ell_i)$.

Interaction similarities for a pair of poses (for two different ligands bound to the same target protein) were computed separately for each interaction type. The interactions made between the ligand and each residue of the target protein residue were tabulated and then the similarity between the resulting lists for each pose was measured by the Tanimoto coefficient (10). The Tanimoto coefficient was modified by the addition of pseudo counts, which serve to make the metric well defined if neither ligand forms a particular type of interaction and to reward poses

that share larger numbers of interactions in absolute terms. We define the interaction similarity, for interaction type k between a pair of poses ℓ_i, ℓ_j (for ligands i and j, respectively), as

$$s^{(k)}(\ell_i, \ell_j) = \frac{1 + \sum_{r \in R} \sqrt{X_r^{(k)}(\ell_i) X_r^{(k)}(\ell_j)}}{2 + \sum_{r \in R} \left[X_r^{(k)}(\ell_i) + X_r^{(k)}(\ell_j) \right] - \sum_{r \in R} \sqrt{X_r^{(k)}(\ell_i) X_r^{(k)}(\ell_j)}},$$

where R is the set of all protein residues.

When computing hydrogen bond similarity, a case where a given protein residue acts as a hydrogen bond donor for one ligand and a hydrogen bond acceptor for another ligand is not considered a shared interaction.

Substructure similarity

To compute the substructure similarity for a pair of candidate poses, the maximum common substructure of the two ligands is identified using Canvas (Schrodinger LLC) and then mapped onto each candidate pose. Finally, the RMSD between these two sets of atoms is computed and used as the measure of substructure similarity. We defined custom atom and bond types for computation of the common scaffold (**Supplementary Table 6**). Substructure similarity is not considered for pairs of ligands with a maximum common substructure of less than half the size of the smaller ligand. Hydrogen atoms were not included in the substructure nor when determining the total number of atoms in each ligand.

Importance of similarity types

Removing any of the similarity types from the ComBind potential reduced ComBind's performance (**Supplementary Fig. 5**). In particular, both protein–ligand interaction similarity and substructure similarity contribute substantially to ComBind's accuracy. Protein–ligand interaction similarity is the more important of the two, particularly when using a diverse set of helper ligands.

Computation of similarity statistics

Using the set of protein–ligand complex structures described above, we characterized the extent to which distinct ligands binding a common target adopt similar poses, as quantified by the interaction and substructure similarity metrics described above. (We note that the three ion channel targets were not included in these statistics because they were added after the rest of our study had been completed.)

When computing these statistics, we docked the ligands using Glide and then identified poses that are near-native among the candidate poses ranked in the top 100 by Glide. We used these docked poses, as opposed to the experimentally determined pose, in order to ensure that the statistics will be applicable to the scoring of candidate poses generated by Glide. We computed the empirical distribution of each similarity type across all pairs of near-native poses using a Gaussian kernel density estimate with standard deviation of 0.03 for interaction similarities and 0.18 for substructure similarities. To reduce bias near the boundaries, we applied reflected boundary conditions (11).

We capped substructure similarities at 6 Å (that is, substructure similarities greater than 6 Å were set to 6 Å), as the sparsity of near-native pose pairs for higher values led to overly rough

distributions. We denote the similarity distribution over near-native poses for interaction type k as $f_k(x; \text{Native})$.

We computed equivalent similarity distributions using all pairs of candidate poses produced by Glide, regardless of whether they are near-native. We denote the resulting distributions as $f_k(x; \text{Reference})$.

To combine the distributions for the four similarity types into a single joint distribution, we assume that the interaction types are conditionally independent and express the joint distribution as a product of the distributions for each interaction type. That is:

$$f\big(s(\ell_i, \ell_j); \text{Native}\big) = \prod_k f_k\big(s^{(k)}(\ell_i, \ell_j); \text{Native}\big), \text{ and}$$

$$f\big(s(\ell_i, \ell_j); \text{Reference}\big) = \prod_k f_k\big(s^{(k)}(\ell_i, \ell_j); \text{Reference}\big).$$

where $s(\ell_i, \ell_j)$ is the vector of $s^{(k)}(\ell_i, \ell_j)$'s for each similarity type k.

Description of the ComBind method

The ComBind score

We describe a hypothesized set of binding poses of a set of n ligands as $L = \ell_1, \ell_2, \ldots, \ell_n$, where ℓ_i specifies the hypothesized pose for ligand i.

Per-ligand scoring functions, which consider each ligand independently, would determine an optimal set of poses $\hat{L}$ by choosing the binding pose with minimum docking score for each ligand or, equivalently, by minimizing

$$E^{\text{dock}}(L) = \sum_{i=1}^{n} E^{\text{dock}}(\ell_i)$$

where $E^{\text{dock}}(\ell_i)$ is the output of a per-ligand scoring function (such as that reported by Glide) for pose ℓ_i of ligand i.

In our method, we add pairwise terms that tend to favor sets of similar poses:

$$E^{\text{ComBind}}(L) = CE^{\text{dock}}(L) + \frac{1}{(n-1)} \sum_{(i,j), i \neq j} -\log \frac{f\big(s(\ell_i, \ell_j); \text{Native}\big)}{f\big(s(\ell_i, \ell_j); \text{Reference}\big)}.$$

Intuitively, these pairwise terms reward pose pairs with similarity values more often observed in near-native (correct) pose pairs than in reference pose pairs (i.e., pose pairs chosen at random from among all candidates). The idea of comparing the distribution of features in correct solutions to the distribution in all possible solutions has been used in statistical potentials for biomolecular structure prediction (12-14) and in the naïve Bayes machine learning model (15). We divide the pairwise scores by the number of ligands minus 1, in order to hold the relative contribution of singleton and pairwise terms constant for different numbers of helper ligands.

Consistent with their reported units of kcal/mol, we find that Glide scores have the mathematical properties of an energy; namely, the negative log likelihood ratio of a pose being near-native is linear in its Glide score (**Supplementary Fig. 3**). By construction, the pairwise terms we introduce in this study also have this property. This congruence implies that these singleton and

pairwise terms can be additively combined (as this is the equivalent of multiplying likelihood ratios).

The per-ligand docking scores are scaled by the constant factor C in order to be consistent with the pairwise terms. For example, if the docking scores were on average 10 times the negative log likelihood ratio of a pose being near-native, they would need to be scaled by 1/10. This constant factor can be identified by performing logistic regression with the docking scores as features and whether each pose is near-native as the response. For Glide scores, the appropriate constant is close to 1 (0.9) (**Supplementary Fig. 3**), and we chose to set it to 1 for simplicity.

Optimization procedure

We use coordinate descent to compute a set of poses that minimizes the ComBind score. At first, L is randomly initialized. L is then iteratively improved by iterating through the ligands, in a random order, and updating the selected ligand's pose to the argument minimum of $E^{\mathrm{ComBind}}(L)$ assuming that the other poses in L are correct. This procedure is repeated until no more updates can be made. Each update can be computed efficiently because it depends only on the partial contribution of the selected ligand's pose to the ComBind score:

$$\hat{\ell}_q = \underset{\ell_q}{\operatorname{argmin}} \left[C E^{\mathrm{dock}}(\ell_q) + \frac{1}{(n-1)} \sum_{i \neq q} -\log \frac{f\left(s(\ell_q, \ell_i); \mathrm{Native}\right)}{f\left(s(\ell_q, \ell_i); \mathrm{Reference}\right)} \right].$$

In order to account for the non-convex nature of the ComBind score, we repeat this algorithm from 500 initial configurations, explicitly including the initial configuration corresponding to the generic scoring function predictions at least once and return the best scoring configuration. Empirically this procedure converges to the same result over multiple runs.

Benchmarking

We evaluated the performance of ComBind on the 30 target proteins listed in **Supplementary Table 2.** We only considered ligands that have less than 50% scaffold overlap with the ligand that was originally present in the experimental structure used for docking. We found that ligands with higher scaffold overlap were substantially easier to dock, likely due to the binding pocket being well shaped to accommodate the similar ligand (**Supplementary Table 1**). Additionally, we only consider ligands for which there is at least one correct candidate pose, since only in these cases is it possible for either ComBind or Glide to make a correct prediction. Importantly, this subsetting was only done for the query ligands, not the helper ligands downloaded from ChEMBL described below.

For each of the 245 unique ligands meeting these criteria, we identified other ligands known to bind the respective target protein from the ChEMBL database and then used ComBind to jointly predict their binding poses. Importantly, when evaluating the performance of our method on a particular target protein, we excluded the data for that target protein from the similarity statistics.

Selection of helper ligands

For all targets, we downloaded K_i or IC_{50} data (whichever was more numerous) from ChEMBL (16). We removed ligands that did not meet the following criteria: a ChEMBL confidence score of 9 (the highest value), molecular weight < 800 Da, and K_i or IC_{50} < 1 μM. Ligand structures were generated from the SMILES strings provided by ChEMBL.

We benchmarked two criteria for selecting which ChEMBL ligands to use as helper ligands for each query ligand: (1) the highest affinity binders that do not share a chemical scaffold, and (2) the ligands that share the largest chemical substructure with the query ligand. To define the size of the common substructure, we used the same maximum common substructure definition as that used to compute substructure similarity. For selection method (1), we added helper ligands in order of affinity, not adding a ligand if it has greater than 80% substructure overlap with any ligand already in the selected set of helpers.

The benchmarking results presented in the figures were obtained using the following ligand selection criteria and number of helper ligands: **Fig. 4A** and **Supplementary Fig. 5A**: 20 helper ligands selected using criterion (1); **Fig. 4B**: the indicated number of ligands selected using criterion (1); **Supplementary Fig. 4A** and **Supplementary Fig. 5B**: 20 helper ligands selected using criterion (2); and **Supplementary Fig. 4B**: the indicated number of ligands selected using criterion (2). For a handful of targets, fewer than 20 helper ligands were available meeting our criteria. In these cases, we used the minimum of the indicated number of ligands and the number of available ligands. In **Fig. 4A** and **Supplementary Fig. 4A**, the individual targets listed are those for which poses were predicted for two or more query ligands. Data for targets for which only a single query ligand's pose was predicted are also included in aggregate performance metrics (for each target class and overall).

Performance evaluation

We developed an overall performance metric to represent the expected performance in drug development campaigns. For each protein family, we computed the average performance, then weighted each by the fraction of FDA-approved drugs targeting the protein family, as reported in Santos *et al.*, 2017.

Description of the ComBindVS method

ComBindVS takes as input a structure of the target protein, a set of ligands known to bind the target (helper ligands), and a library of candidate molecules to screen. First, poses are predicted for the set of known binders using ComBind. Second, for each candidate molecule, per-ligand docking software (Glide for the results presented here) is used to produce many candidate poses and a pose is selected from the candidate poses that minimizes the ComBind score with respect to the poses for the helper ligands. In these evaluations, the poses for the helper ligands are fixed, since we do not know if the candidate molecule can bind the target protein. Third, the ComBind score of each candidate molecule in its predicted pose is used as a prediction of its affinity relative to other molecules, and for virtual screening, the candidate molecules are ranked by this score.

In ComBindVS, we replaced the substructure similarity term with a shape similarity term when determining poses and scores for candidate molecules (i.e., the second and third steps above). Our substructure similarity term is defined only for pairs of ligands with a common substructure of at least half the size of the smaller ligand, which sometimes leads to large differences in score between two very similar candidate molecules because ones shares a sufficiently sized substructure with a helper ligand while the other does not. The shape similarity term captures much of the same information as the substructure similarity term but avoids these sudden "jumps" in score between similar molecules. We computed the shape similarity of two poses using the Schrodinger shape screening tool with 'pharm' atom-typing using the 'inplace' mode

(17). These shape similarity values were incorporated into the ComBind scoring function identically as were the interaction similarity and substructure similarity metrics.

For each candidate molecule, we considered up to 30 candidate poses (decreased from 100 for computational efficiency; for the same reason, we left "Enhanced Sampling" at the default value instead of 4).

Ligand chemical similarity and its integration with per-ligand docking or ComBindVS

Ligand chemical similarity methods can largely be separated into "2D" methods that consider the presence of similar chemical groups in the provided ligands and "3D" methods that consider the relative 3D positioning of functional groups able to form particular intermolecular interactions. We chose a widely used method in each of these two families to compare against ComBindVS. We computed 2D ligand chemical similarity using the Tanimoto similarity metric acting on diameter-four extended connectivity fingerprints (ECFP4) with Morgan atom typing, as implemented in RDKit version 2020.03.1. We computed 3D ligand chemical similarity using the Schrodinger shape screening tool using 'pharm' atom-typing (17). We averaged these similarity scores across the helper ligands.

To combine these scores with each other, as well as with the ComBind or per-ligand docking scores, we used use a framework similar to that described above for incorporation of pose similarity scores. In particular, we computed the empirical distribution of each similarity metric for the active and decoy compounds across the DUD-E dataset. Given a ligand to score, we compute the appropriate similarity metric, look up the frequency of that similarity value in the active and decoy distributions, and then compute the negative log ratio of these two values. These scores are then simply combined by addition. Data from the target protein being evaluated were excluded from the statistics. We found this method of integrating the Glide or ComBind scores with the ligand-based scores to perform comparably to or better than the z-score method (18) and the product-of-ranks method (19) (**Supplementary Fig. 9C,D**).

The "ligand-based" method for which we present results in **Figure 5** incorporates the 2D and 3D ligand chemical similarity metrics described above. The combination of these two methods performed comparably to or better than either method alone (**Supplementary Fig. 9B**).

ComBindVS performance evaluation

We considered data from all 102 targets in the DUD-E benchmark set (20). For each target protein, we used the provided structure for docking. These protein structures and the provided ligand data were processed using the same protocol described above for our pose prediction benchmark. We removed systems where the crystallographic ligand is chelated to a metal because automated use of Glide is not recommended in these cases (21).

For each target protein, we randomly selected 1, 5, and 10 ligands to use as helper ligands in ComBindVS or the ligand-based method. In addition, the crystallographic ligand was included as a helper ligand, and only its near-native poses were considered by ComBind. In cases where docking failed to produce any near-native poses, the native pose was used. For each protein, we repeated this procedure 5 times with different helper ligands and averaged the results, to lower the variance of the results. For each target protein, we excluded any data involving that protein from the statistics used in the ComBind scoring function.

To assess the impact of the chemical similarity of the candidate molecules to the helper ligands, we split the candidate molecules into two bins based on the maximum 2D Tanimoto similarity to any of the helper ligands: 0 to 0.2 and 0.2 to 0.3. We did not evaluate performance for candidate molecules with a maximum Tanimoto coefficient greater than 0.3; due to the methodology used for generation of decoys in DUD-E, very few (~1%) of the decoys fall in this regime, making the virtual screening problem uncharacteristically easy. So that the test set would not change for different numbers of helper ligands, we performed this filter considering the full set of 10 helper ligands for each target. We only included target proteins for which there are at least 10 actives and 100 decoys meeting each similarity threshold. For the threshold of 0.2, this left 39 targets, and for the range of 0.2 to 0.3 this left 76 targets.

In addition to enrichment factors, we quantified performance using the adjusted area under log receiver operating characteristic curve (aulogROC) metric (22). Both metrics are designed to reward strong performance in the low false-positive-rate regime, which is the most relevant because in virtual screening projects typically only the highest-scoring molecules are synthesized and assayed. The aulogROC is computed by plotting the true-positive-rate as a function of the logarithm of the false-positive rate, then integrating the area under the curve. We integrated over 0.001 to 1, and then subtracted 0.14462 from the result, such that a random ranking of candidate molecules has an expected aulogROC of 0. Qualitatively similar results were observed with enrichment factors and the aulogROC metric (**Supplemental Fig. 9A**). We note that the lower performance of per-ligand docking in the 0 to 0.2 Tanimoto similarity bin compared to the 0.2 to 0.3 bin is likely due to per-ligand docking performing best on candidate molecules that are chemically similar to the co-crystalized ligand (23).

ComBindVS computing requirements

Virtual screening of very large compound libraries can be computationally intensive. Fortunately, ComBindVS increases computing requirements only modestly relative to per-ligand docking—by less than a factor of two, even when using dozens of known binders as helper ligands.

Prediction of binding poses of antipsychotics at the D_2R

Execution of the ComBind method

We predicted binding poses for the typical antipsychotics spiperone, mespiperone, benperidol, and pimozide at the human D_2 dopamine receptor (D_2R). We prepared the ligands using the Schrodinger ligprep tool, considering the unprotonated tautomer and both inversions of the protonated tautomer. The same docking protocol was used as described above, except that the top 300 poses were considered by ComBind, in order to account for the use of the 3 tautomeric states of the ligand.

Quantitative binding affinity prediction

We assembled a set of piperidine butyrophenones from ChEMBL by downloading all ligands from ChEMBL release 27 for the human or rat D_2R annotated with a K_i, then selecting all ligands matching the SMARTS pattern "[cH]1[cH]c[cH][cH]c1C(=O)[C;R0][C;R0][C;R0][N;r6]". The helper ligands (spiperone, pimozide, benperidol, and n-methyl spiperone) were excluded from this set. The ComBindVS protocol was used to produce quantitative affinities exactly as described above. The p-value reported in this section was obtained by bootstrapping.

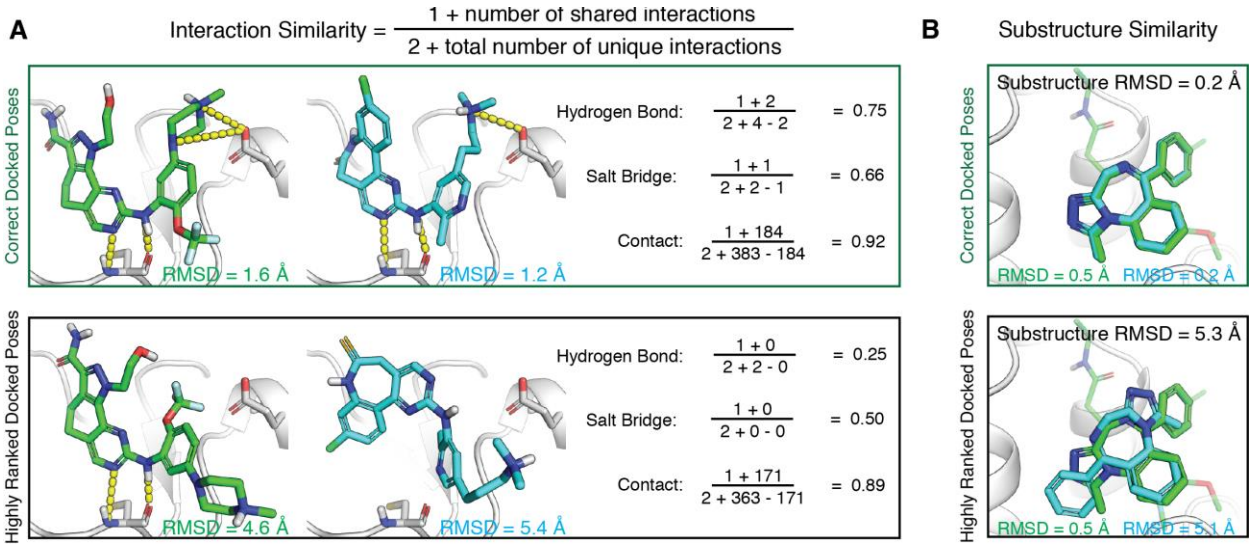
D_2 Dopamine receptor mutagenesis

Wild type (wt) human D_2R in pcDNA3.1 was kindly provided by the laboratory of Jonathan Javitch (Columbia University, New York, NY). Mutations were introduced through of a modified QuikChange (Stratagene, La Jolla, CA) mutagenesis protocol using the following primers V91F: 5'-GGTCATGCCCTGGTTTGTCTACCTGG-3', S193A: 5'CGTGGTCTACGCCTCCATCGTCTCC-3', S193V: 5'-CGTGGTCTACGTCTCCATCGTCTCC-3', S193L: 5'-CGTGGTCTACCTCTCCATCGTCTCC-3', W100L: 5'-GGTAGGTGAGTTGAAATTCAGCAGG-3', C118M: 5'-GGACGTCATGATGATGACGGCGAGC-3', W386F: 5'-CGTGTTCATCATCTGCTTTCTGCCCTTCTTC-3', F389L: 5'-GCTGGCTGCCCTTATTCATCACACACATCC-3'.

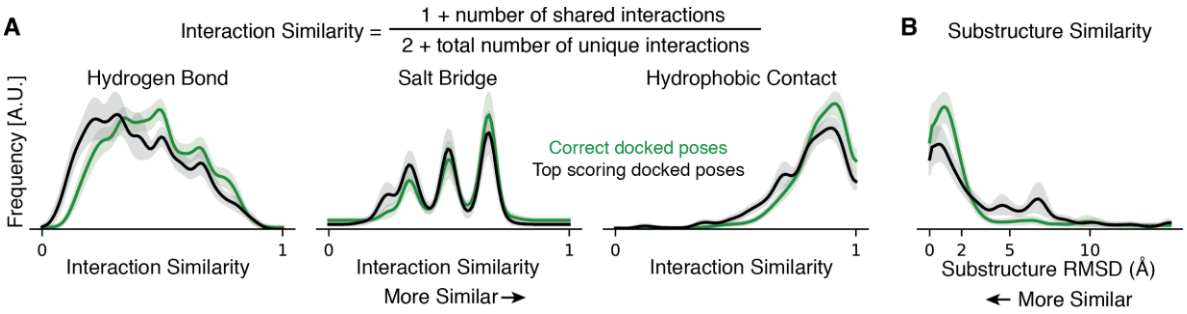
Membrane preparation and radioligand binding

Membranes were isolated from HEK293T cells transiently transfected with D_2R(wt) or D_2R-mutants. Briefly, cells were harvested 48 hr post-transfection (with Lipofectamine 2000), rinsed with PBS, lifted with harvesting buffer (0.68 mM EDTA, 150 mM NaCl, 20 mM HEPES, pH 7.4), and centrifuged at 200 x g for 3 min. The cells were resuspended in ice cold homogenizing buffer (10 mM HEPES, pH 7.4, 100 mM NaCl, 0.5 mM EGTA), homogenized using a Tissue Tearer (BioSpec, Bartlesville, OK) for 30 sec, and centrifuged at 20,000 x g for 20 min. Membranes were resuspended in Binding Buffer (20 mM HEPES, pH 7.4, 100 mM NaCl) using a Dounce glass homogenizer, flash frozen in liquid N_2 and stored at –80°C.

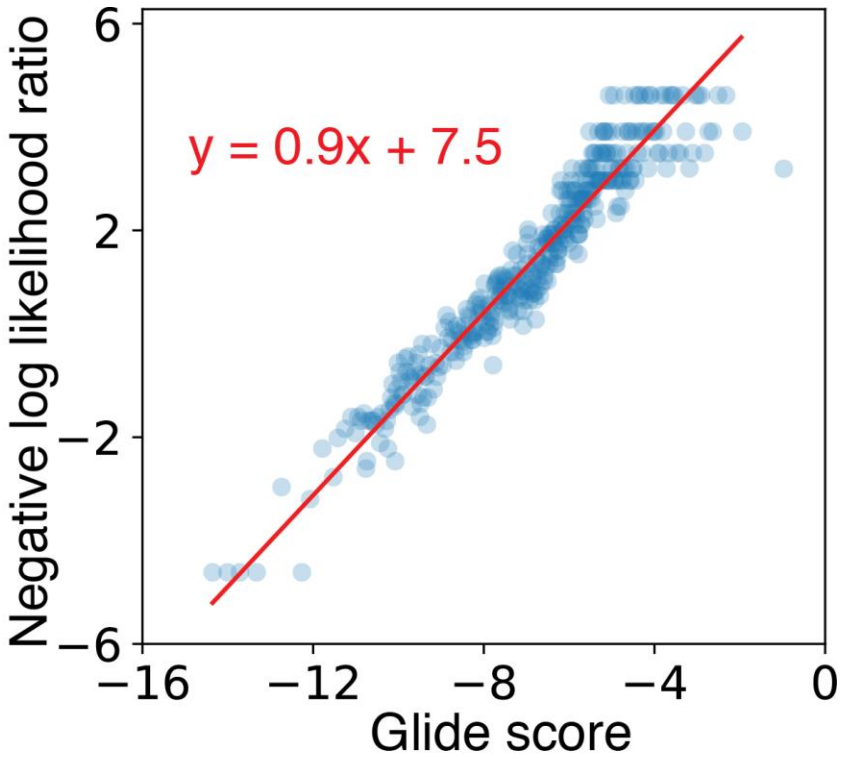
For saturation binding assays, cell membranes (0.6–20 μg per well, depending on the mutant) were incubated for 1.5 hr at 30°C with [^{3}H]-spiperone (Perkin Elmer, Waltham, MA) (0.02–12 nM, depending on the Kd of the D_2R mutant) in Binding Buffer containing 0.001% BSA, 1 mM ascorbic acid, and 100 nM GDP with or without 20 μM (+)butaclamol (to determine non-specific binding). For competition binding assays, cell membranes (0.6–20 μg, depending on the D_2R mutant) were incubated for 1.5 h at 30°C with [^{3}H]-spiperone (0.05–0.6 nM, depending on the Kd of the D_2R mutant) in Binding Buffer containing 0.001% BSA, 1 mM ascorbic acid, 100 nM GDP and 0–0.1 nM test compound (purchased from Millipore-Sigma, St Louis, MO), or 20 μM (+)butaclamol (to determine non-specific binding). Sample membranes were harvested by vacuum filtration on 96-well GF/C filter plates, washed with ice cold binding buffer to remove unbound radioligand, and allowed to dry before adding Microscint 0 (Perkin Elmer, Waltham, MA) for counting in a Top Count Scintillation Counter (Perkin Elmer/Packard, Waltham, MA). Data were fit to a one site binding curve to determine Kd for [^{3}H]-spiperone saturations, or to a one-site competition binding curve to calculate Ki of test compounds using Prism (GraphPad, San Diego CA).



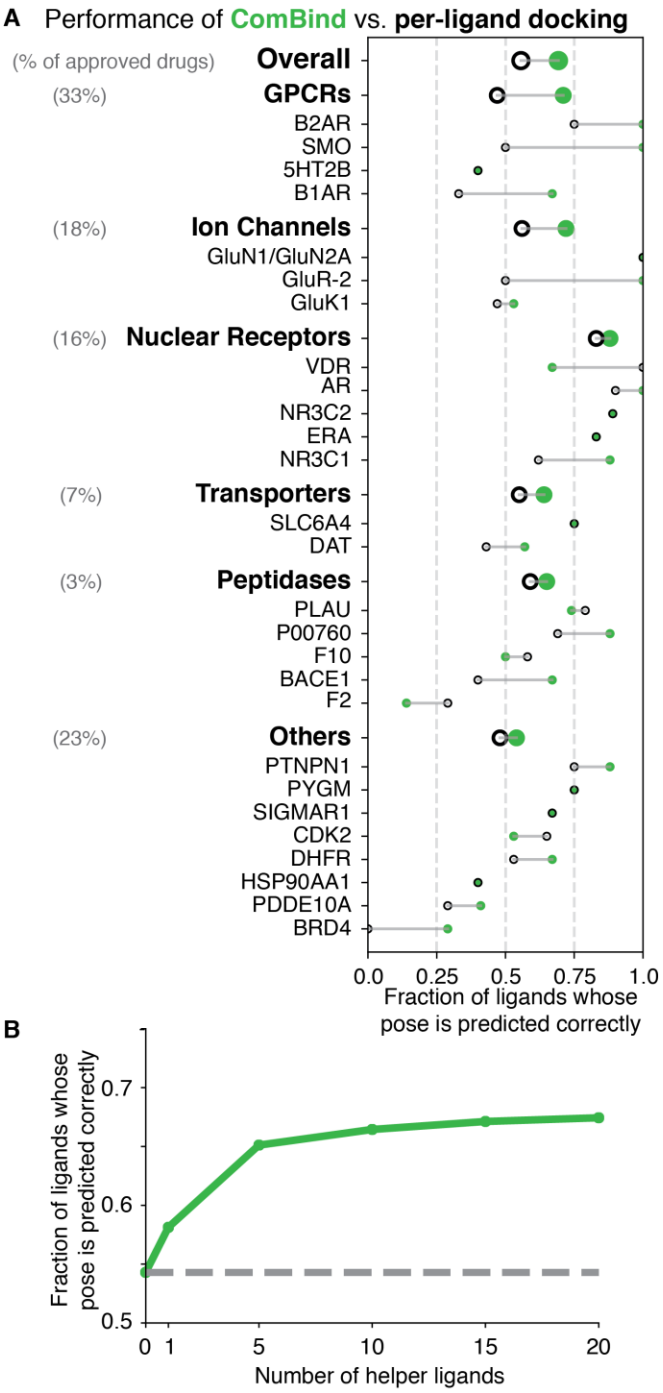
Supplementary Figure 1: Examples of interaction similarity and substructure similarity computation. (A) Comparison of interactions formed by two ligands bound to *PLK1*, for a pair of correct poses (top) and randomly chosen poses (bottom). (B) Overlays of two ligands that share a common substructure bound to *BRD4* for correct docked poses (top) and randomly chosen highly ranked docked poses (bottom).



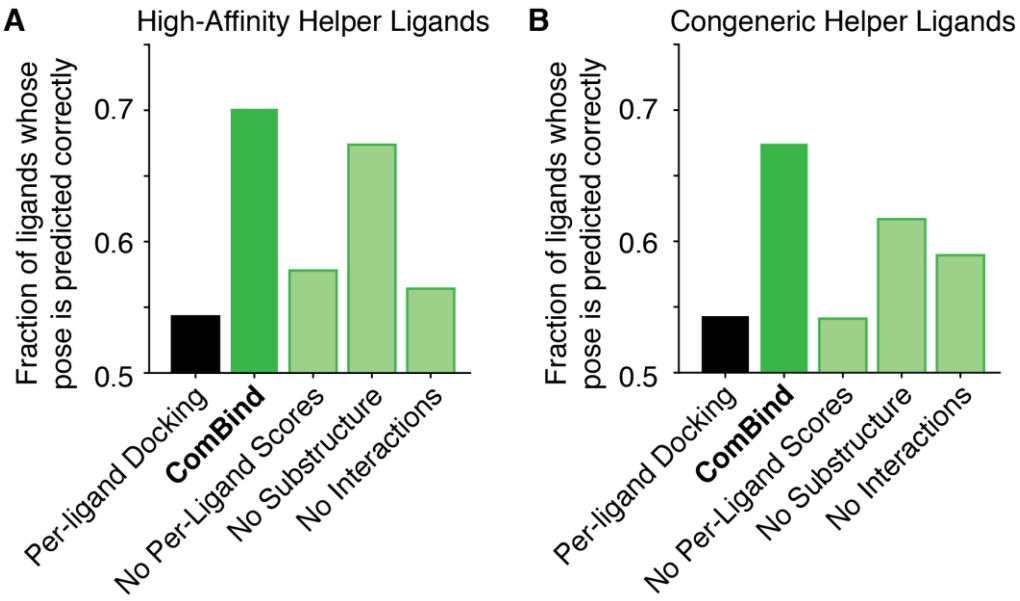
Supplementary Figure 2: A state-of-the-art per-ligand scoring function (Glide) underestimates the similarity of binding poses of different ligands binding to the same target protein. (A) and (B) are identical to Fig. 2C and D, respectively, except that the black curves in this figure are computed using only the pose ranked first by Glide for each ligand.



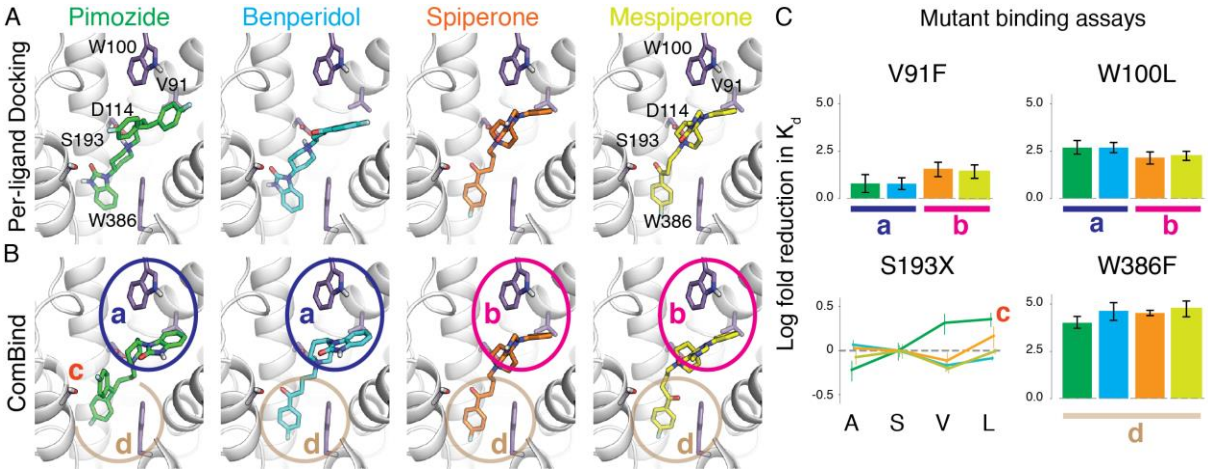
Supplementary Figure 3: The output of Glide's per-ligand scoring functions is in units of energy similar to those of ComBind's pairwise potential. A quantile plot showing the relationship between Glide scores and the negative log likelihood ratio of a pose being correct. For each of the docked poses of each ligand in our benchmark set, we computed the Glide score and determined whether the pose was correct. We split all of the resulting data into quantiles based on Glide scores, with each quantile containing 100 poses. Each point in the plot represents the mean Glide score and negative log likelihood ratio for a given quantile. The red line shows the best-fit linear relationship between these two quantities as determined by logistic regression.



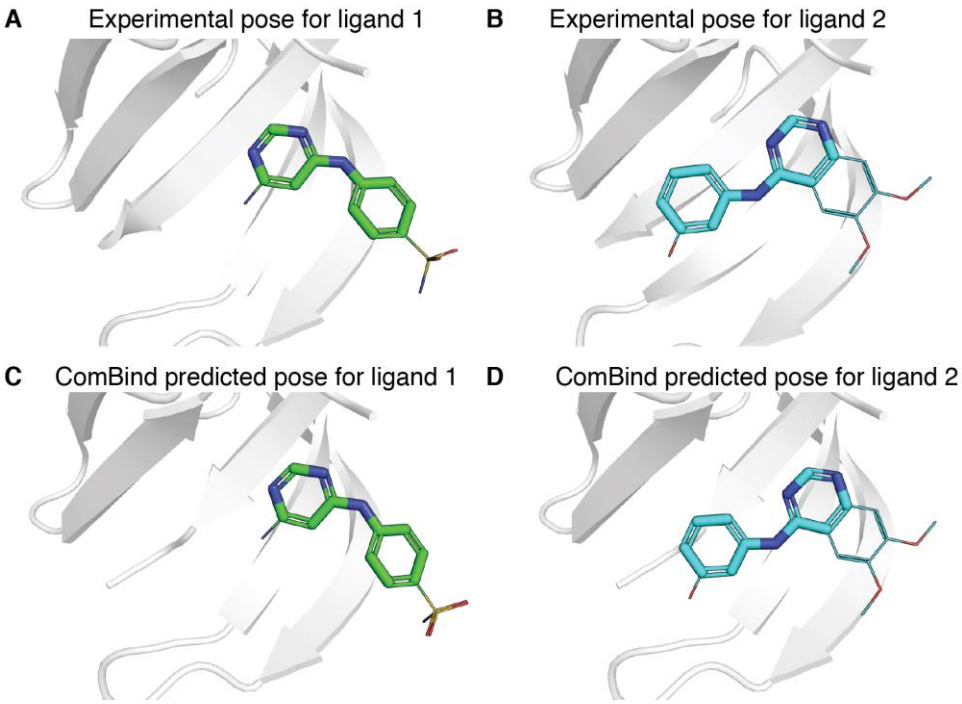
Supplementary Figure 4: ComBind performance using a congeneric series of ligands. This figure corresponds to **Fig. 4,** but with helper ligands selected from ChEMBL ligands according to the "congeneric" criterion (i.e., ligands that share the greatest common substructure with the query) instead of the "high affinity" criterion.



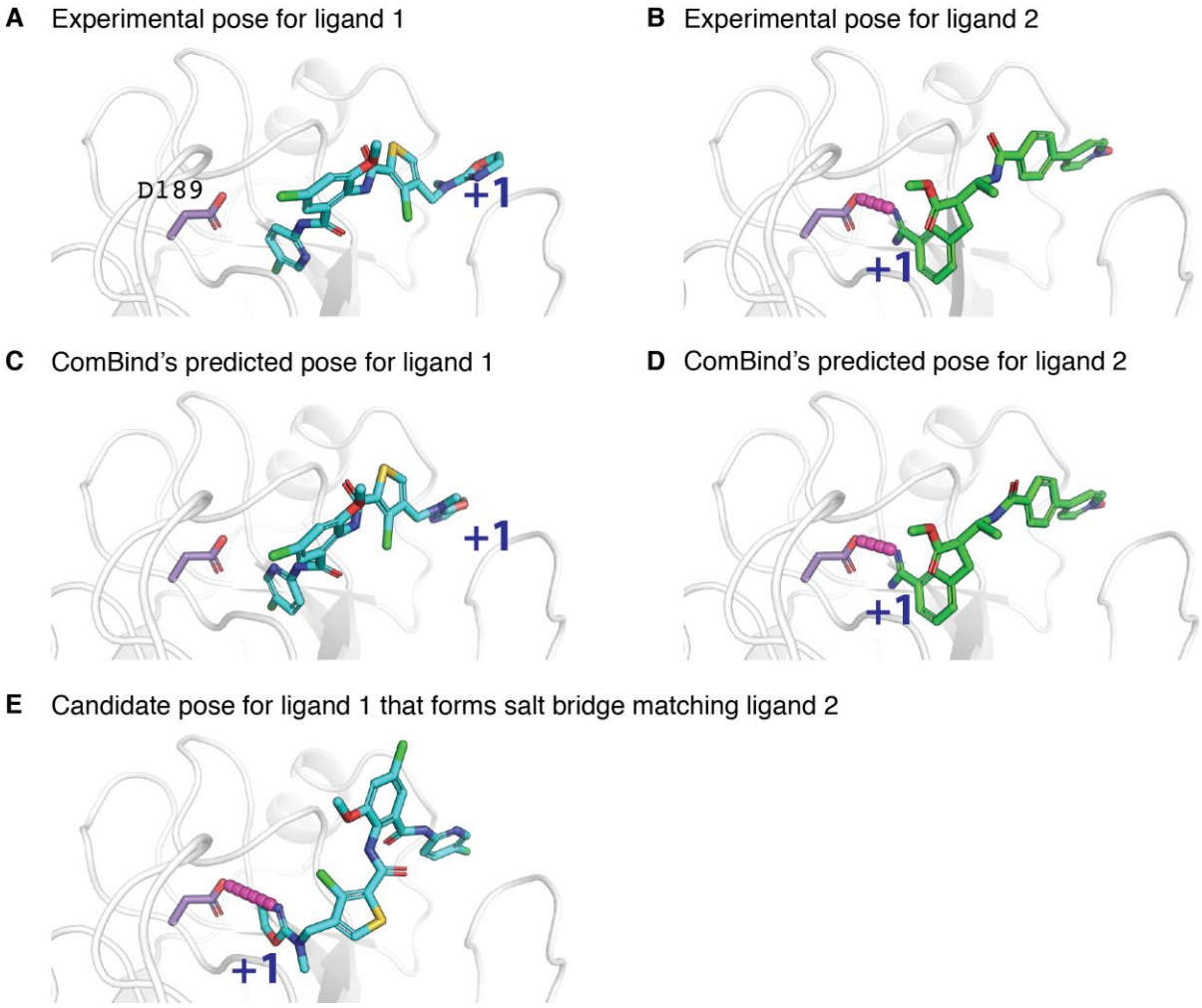
Supplementary Figure 5: Importance of components of the ComBind scoring function. Performance using various components of the ComBind scoring function when using helper ligands chosen by either the high-affinity (A) or congeneric (B) ChEMBL ligand selection criterion. ComBind uses per-ligand docking scores, similarity scores based on interactions, and similarity scores based on relative positions of shared substructures. "Per-ligand docking" (Glide) omits all similarity scores. The remaining bars ("No Per-Ligand Scores," "No Substructure," and "No Interactions") show the effects of omitting per-ligand scores, substructure position similarity scores, and protein–ligand interaction similarity scores, respectively, from the ComBind potential.



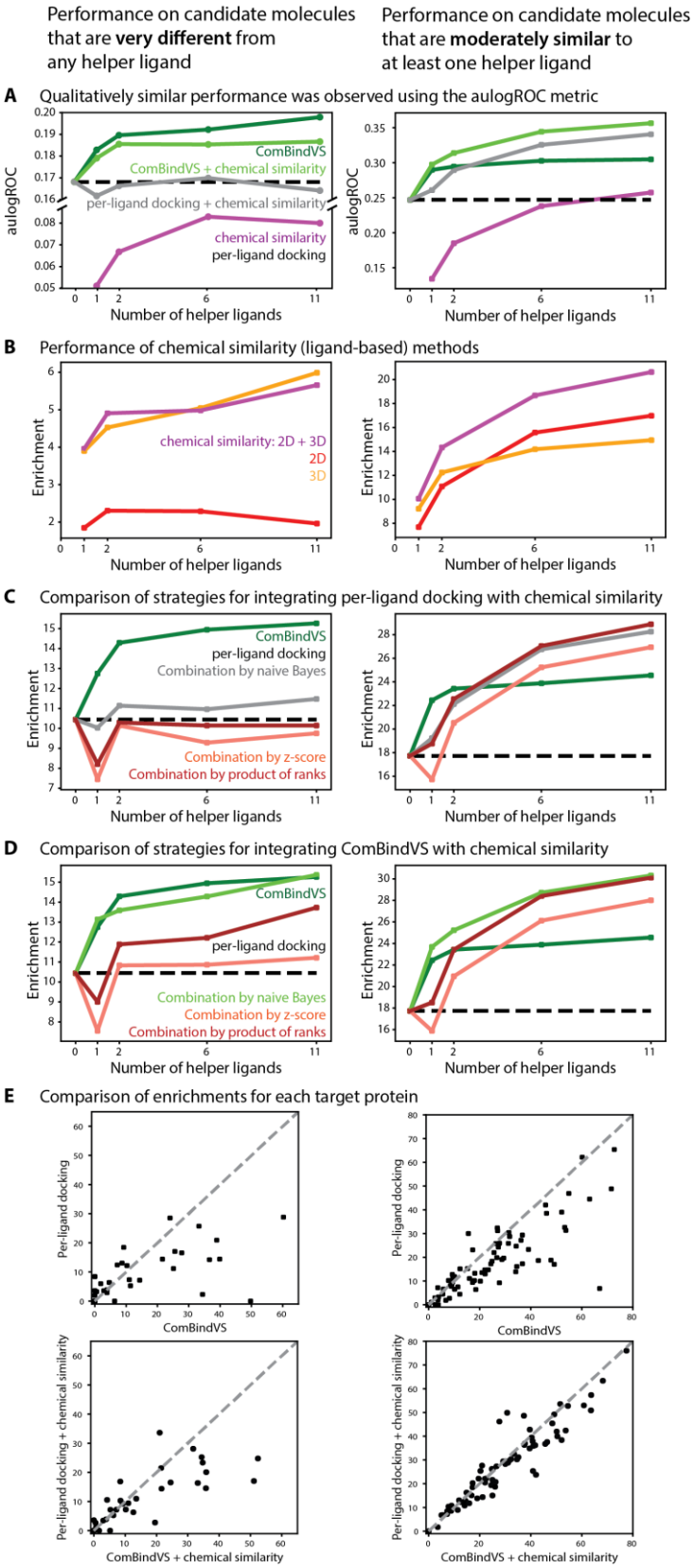
Supplementary Figure 6: Prediction and validation of the binding poses of antipsychotics at the D_2 dopamine receptor—additional data. (A) Binding poses of pimozide, benperidol, spiperone, and mespiperone as predicted by Glide. (B) Binding poses of the same ligands, as predicted by ComBind. (C) Results of mutagenesis studies designed to test ComBind's binding pose predictions. Ligands are color-coded as in panel A. Error bars show standard error of the mean. S193 was mutated to A, S, V and L; these results are discussed in the main text. Unlike Glide, ComBind predicts that all four ligands will position a fluorobenzene ring at the bottom of the binding pocket, packing favorably against Trp386 (W386). Indeed, mutating W386 to a smaller residue (Phe) reduced affinity to a similar extent for all of the ligands, with a slightly smaller effect for pimozide, which packs less tightly against W386 according to ComBind's prediction. At the top of the ligand binding pocket, near Val91 (V91) and Trp100 (W100), ComBind predicts that the pimozide and benperidol will place identical functional groups that differ somewhat from those of spiperone and mespiperone. Indeed, mutation of these residues affects pimozide and benperidol slightly differently from spiperone and mespiperone.



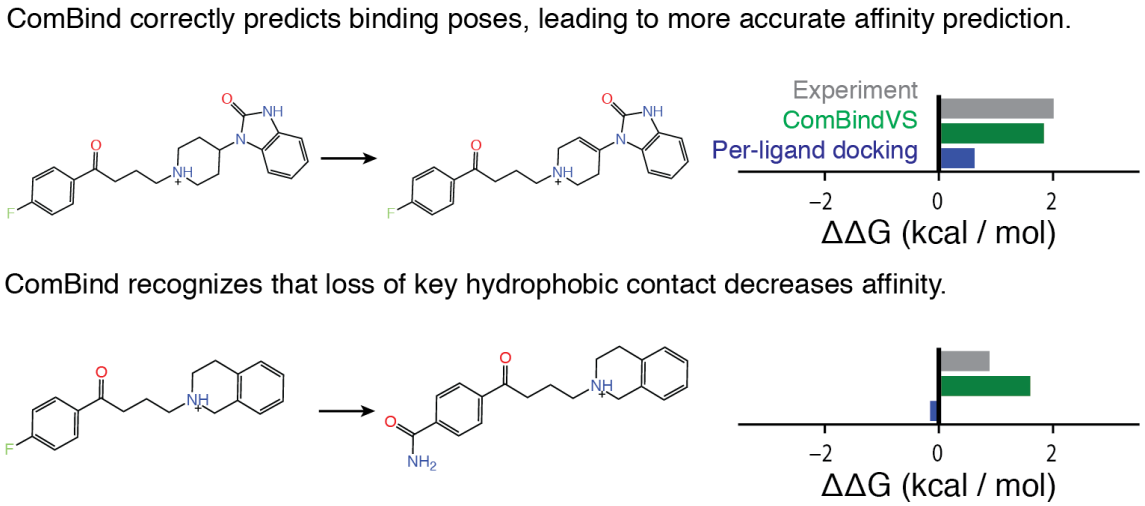
Supplementary Figure 7: Example of a case where ComBind correctly predicts that a shared chemical scaffold is placed differently for different ligands. We show two ligands that bind the kinase CDK2. These ligands share a common scaffold but adopt significantly different binding poses. In A and B, we show their experimentally determined poses (PDB: 1JSV and PDB: 1DI8, respectively). In C and D, we show the poses predicted by ComBind for the two ligands. The shared scaffold is shown in the thicker sticks and parts of the ligands that differ are shown in the thinner lines.



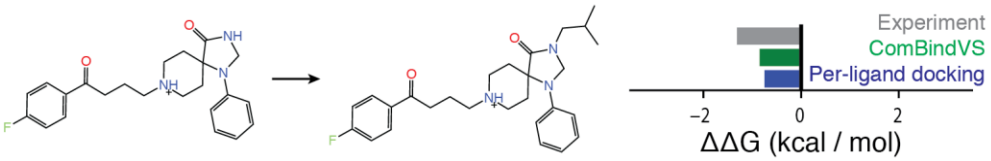
Supplementary Figure 8: Example of a case where ComBind correctly predicts that ligands form distinct interactions with the protein. We ran ComBind for 20 ligands that bind F10. While most of the ligands have a positively charged group, only some of them position it to form a salt bridge with D189 (e.g., ligand 1, shown in panel A) while others orient it in the complete opposite direction (e.g., ligand 2, shown in panel B). ComBind correctly predicts both binding poses (C, D). (E) One of the candidate poses for ligand 1 forms the same salt bridge as ligand 2. ComBind correctly avoided choosing this pose, even though choosing it would have led to more similar interactions between ligands.


Performance on candidate molecules that are **very different** from any helper ligand
Performance on candidate molecules that are **moderately similar** to at least one helper ligand
A
Qualitatively similar performance was observed using the aulogROC metric
aulogROC
ComBindVS
ComBindVS + chemical similarity
per-ligand docking + chemical similarity
chemical similarity
per-ligand docking
Number of helper ligands
B
Performance of chemical similarity (ligand-based) methods
Enrichment
chemical similarity: 2D + 3D
2D
3D
Number of helper ligands
C
Comparison of strategies for integrating per-ligand docking with chemical similarity
ComBindVS
per-ligand docking
Combination by naive Bayes
Combination by z-score
Combination by product of ranks
D
Comparison of strategies for integrating ComBindVS with chemical similarity
E
Comparison of enrichments for each target protein
Per-ligand docking
ComBindVS
Per-ligand docking + chemical similarity
ComBindVS + chemical similarity

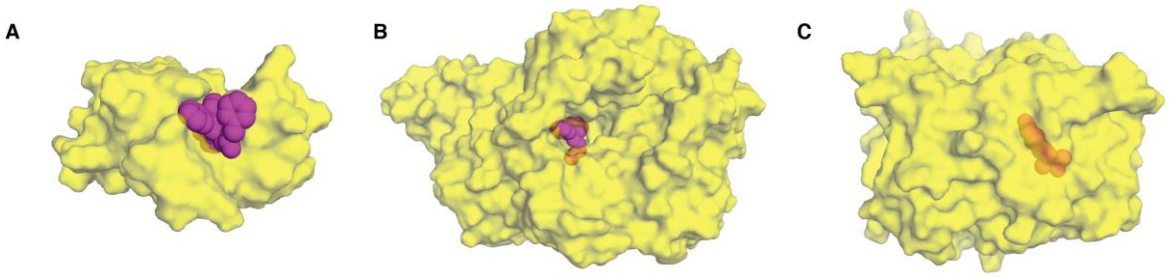
Supplementary Figure 9: Virtual screening performance using aulogROC metric, performance for individual targets, performance of ligand-based methods, and comparison of integration strategies. The left column shows performance for candidate molecules very different from any helper ligand ($Tc \leq 0.2$), and the right column shows performance for candidate molecules moderately similar to at least one helper ligand ($0.2 < Tc < 0.3$). (A) Average performance on DUD-E benchmark set as quantified by the adjusted area under log receiver operating characteristic curve (aulogROC) metric, defined in Methods. (B) Comparison of performance of a 2D ligand-based method, a 3D ligand-based method, and a combination of the two. (C) Comparison of strategies for integrating results of ligand-based screening and per-ligand docking. The "combination by naïve Bayes" strategy is referred to as "per-ligand docking + chemical similarity" in other figures. See Methods for a complete description of each integration strategy. (D) Comparison of methods for integrating results of ligand-based screening and ComBindVS. The "combination by naïve Bayes" strategy is referred to as "ComBindVS + chemical similarity" in other figures. (E) Comparison of enrichment factor 1% for each target protein in the DUD-E benchmark for the indicated scoring methods.



Supplementary Fig. 10. Examples of how ComBindVS can improve relative binding affinity predictions for analogs. (A) Benperidol to droperidol. ComBind predicts benperidol's binding pose more accurately than per-ligand docking, which predicts a "flipped" pose (see main text). As a result, ComBind more accurately predicts the change in affinity due to chemical modification. (B) ChEMBL3818994 to ChEMBL3819427. When a substitution is made preventing a hydrophobic contact shared by the helper ligands, ComBind correctly predicts that this will decrease rather than increase binding affinity. Experimental ΔΔG values are the difference in $RT\log(K_i)$ between each pair of ligands. Predicted ΔΔG values are the difference in ComBindVS or per-ligand docking scores between each pair of ligands.



Supplementary Fig. 11. ComBindVS correctly predicts that addition of a hydrophobic group to the secondary amine of spiperone increases ligand affinity, even though this addition decreases interaction similarity to all the helper ligands. In the example shown here, ComBind predicts that addition of an isobutyl group to the secondary amine of spiperone (resulting in molecule ChEMBL334349) increases binding affinity. We observed similar results for other hydrophobic substituents. ΔΔG values are computed as in **Supplementary Fig. 10**.



Supplementary Figure 12. ComBind improves pose prediction accuracy for binding pockets with diverse properties. (A) A poorly formed, shallow binding pocket in bromodomain-containing protein 4 (BRD4; PDB entry 2YEL). (B) A deep binding pocket in the dopamine transporter (DAT; PDB entry 4M48). (C) A fully enclosed binding pocket in the androgen receptor (AR; PDB entry 1E3G). Each protein is shown as a transparent yellow surface. Each ligand is shown as magenta spheres; these spheres appear orange when occluded by the protein (i.e., for the entire ligand in C and parts of it in B and A). For each of these target proteins, ComBind predicts ligand binding poses more accurately than per-ligand docking (Fig. 4).

Supplementary Tables

Supplementary Table 1: Performance of Glide SP and Glide XP on our benchmark set. The data presented in this table does not include ligands that share a substantially sized chemical scaffold with the ligand present in the experimental structure used for docking. Including such ligands increases the success rate for both Glide SP and Glide XP (to 49%, 53%, 47%, respectively).

# Ligands	Is the top-ranked pose correct?			Is any candidate pose correct?		
	SP	*XP*	*IFD*	*SP*	*XP*	*IFD*
327	44%	45%	40%	81%	63%	81%

Supplementary Table 2: Structural data used for benchmarking ComBind. From left to right, columns represent: Protein family, protein name, Uniprot ID, ChEMBL target ID, number of ligands, number of ligands that do not share a scaffold with the ligand present in the experimental structure used for docking, and number of ligands that do not share a scaffold with the ligand present in the experimental structure used for docking and have at least one correct candidate pose. The right-most column corresponds to the number of ligands included in our benchmarks for each target protein.

PROTEIN FAMILY	PROTEIN	UNIPROT	CHEMBL	# TOTAL LIGANDS	# DIVERSE LIGANDS	# DIVERSE LIGANDS WITH AT LEAST ONE CORRECT CANDIDATE POSE
GPCR	5-HT_{2B}	P41595	CHEMBL1833	5	5	5
	β_1AR	P07700	CHEMBL213	11	6	6
	β_2AR	P07550	CHEMBL210	7	4	4
	mGluR5	P41594	CHEMBL2564	4	3	1
	Smo	Q99835	CHEMBL5971	4	3	2
ION CHANNEL	GluN1/2A	Q05586 Q12879	CHEMBL1907604	8	6	4
	GluR-2	P19491	CHEMBL3503	15	7	6
	GluK1	P22756	CHEMBL2919	18	18	15
TRANSPORTER	DAT	Q7K4Y6	CHEMBL238	8	8	7
	SERT	P31645	CHEMBL228	4	4	4
	GLUT1	P11166	CHEMBL2535	2	1	1
NUCLEAR RECEPTOR	ER	P03372	CHEMBL206	20	14	12
	GR	P04150	CHEMBL2034	16	10	8
	MR	P08235	CHEMBL1994	12	10	9
	AR	P10275	CHEMBL1871	19	12	10
	VDR	P11473	CHEMBL1977	20	3	3
PROTEASE	F2	P00734	CHEMBL204	20	19	15
	F10	P00742	CHEMBL244	20	19	12
	PLAU	P00749	CHEMBL3286	20	20	19
	P00760	P00760	CHEMBL3769	20	19	16
	BACE1	P56817	CHEMBL4822	20	19	7

PHOSPHORYLASE	PYGM	P00489	CHEMBL4696	20	5	4
PHOSPHATASE	PTPN1	P18031	CHEMBL335	20	19	8
TRANSCRIPTION FACTOR	BRD4	O60885	CHEMBL1163125	16	13	7
CHAPERONE	HSP90-α	P07900	CHEMBL3880	20	16	10
PHOSPHO-DIESTERASE	PDE10A	Q9Y233	CHEMBL4409	20	19	17
RECEPTOR	σ_1	Q99720	CHEMBL287	4	4	3
ELATASE	ELANE	P08246	CHEMBL248	8	1	1
REDUCTASE	DHFR	P00374	CHEMBL202	20	20	15
KINASE	Cdk2	P24941	CHEMBL301	20	20	17

Supplementary Table 3: ComBind is robust to cases where some of the ligands considered have no correct (near-native) candidate pose. Here we show the results of running ComBind for 20 ligands that bind PTPN1. We considered ligands whose binding poses have been determined experimentally, so that we could assess whether the predicted poses are correct. For over half of the ligands, there were no correct candidate poses (likely because these ligands induce a conformational change in the binding pocket). Despite this, ComBind produces more accurate pose predictions than state-of-the art per-ligand docking software. The ligands used in the predictions correspond to those present in the following PDB structures: 1C88, 1C86, 1GFY, 1ECV, 1C83, 1C84, 1L8G, 1KAV, 1BZJ, 1NWL, 1G7F, 1QXK, 1PYN, 1G7G, 1NZ7, 1NNY, 1NO6, 1ONZ, 1NL9, 1ONY.

	Ligand																			
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
Is any candidate pose correct?	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	N	N	N	N	N	N	N	N	N	N	N
Is Glide's predicted pose correct?	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Is ComBind's predicted pose correct?	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	N	N	N	N	N	N	N	N	N	N	N	N

Supplementary Table 4: Ligands used in predictions for the β1 adrenoceptor. From left to right, columns represent: index of ligand (a–k are as shown in **Fig. 3**; xtal denotes the cocrystallized ligand in the protein structure used for docking**)**, name of ligand, mode of action, and PDB ID of the experimental structure.

Index	**Ligand**	**Mode of action**	**Structure**
xtal	cyanopindolol	antagonist	2VT4
a	dobutamine	partial agonist	2Y00
b	carmoterol	partial agonist	2Y02
c	isoprenaline	full agonist	2Y03
d	salbuterol	partial agonist	2Y04
e	carazolol	inverse agonist	2YCW
f	iodocyanopindolol	antagonist	2YCZ
g	4-(piperazin-1-yl)-1H-indole	antagonist	3ZPQ
h	4-methyl-2-(piperazin-1-yl) quinoline	antagonist	3ZPR
i	bucindolol	antagonist	4AMI
j	carvedilol	inverse agonist	4AMJ
k	methylcyanopindolol	inverse agonist	5A8E

Supplementary Table 5: Definitions for the measures used to quantity the presence of each of the three interaction types considered in this study.

Hydrogen Bond

A = a hydrogen bond acceptor; D = a hydrogen bond donor; H = the associated hydrogen

$$\text{distance term} = \begin{cases} 1 & \text{if } \text{distance}(H,A) \leq 2.5\ \text{Å} \\ \frac{3.0\ \text{Å} - \text{distance}(H,A)}{0.5\ \text{Å}} & \text{if } 2.5\ \text{Å} < \text{distance}(H,A) \leq 3.0\ \text{Å} \end{cases}$$

$$\text{angle term} = \begin{cases} 1 & \text{if } \text{angle}(D,H,A) \geq 120°. \\ \frac{\text{angle}(D,H,A) - 90°}{30°} & \text{if } 90° \leq \text{angle}(D,H,A) < 120° \end{cases}$$

hydrogen bond value = distance term * angle term

Salt Bridge

N = an atom with a negative formal charge; P = an atom with a positive formal charge

$$\text{salt bridge value} = \begin{cases} 1 & \text{if } \text{distance}(N,P) \leq 4.0\ \text{Å} \\ \frac{5.0\ \text{Å} - \text{distance}(N,P)}{1.0\ \text{Å}} & \text{if } 4.0\ \text{Å} < \text{distance}(N,P) \leq 5.0\ \text{Å} \end{cases}$$

Hydrophobic Contact

$A_1, A_2, \ldots, A_n$ = all carbon or halogen atoms in the ligand; $B_1, B_2, \ldots, B_m$ = all carbon atoms in the given protein residue

distance term for A_i and B_j =

$$\begin{cases} 1 & \text{if } \text{distance}(A_i, B_j) \leq 1.25\ r_{ij} \\ \frac{1.75\ r_{ij} - \text{distance}(A_i, B_j)}{0.5\ r_{ij}} & \text{if } 1.25\ r_{ij} < \text{distance}(A_i, B_j) \leq 1.75\ r_{ij} \end{cases}$$

where r_{ij} is the sum of the van der Waals radii of A_i and B_j

hydrophobic contact value = $\sum_{i=1}^{n} \sum_{j=1}^{m}$ distance term for A_i and B_j

Supplementary Table 6: Atom types used in maximum common substructure definition. SMARTS pattern and intuitive description of each atom type used when searching for common substructures. Each atom in a molecule is assigned the most specific atom type (lowest in the table) that applies to it.

SMARTS	Description
(*)	Any Atom
(#1)	Hydrogen
(#6)	Carbon
(#6; r5; CX4) (#6; r6)	Saturated carbon in 5-member ring Carbon in 6-member ring
c1ccccc1	Carbon-only aromatic ring
(CR0)	Carbon not in a ring
(#7)	Nitrogen
(#7; r5)	Nitrogen in 5-member ring
(#8)	Oxygen
O=*	Ketone Oxygen
(#8; r5)	Oxygen in 5-member ring
(#15)	Phosphorus
(#16)	Sulphur
(#16; r5)	Sulphur in 5-member ring
(#9, #17, #35, #53)	Halogens

References

1. R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, The PDBbind database: methodologies and updates. *J Med Chem* **48**, 4111-4119 (2005).

2. R. A. Friesner *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47**, 1739-1749 (2004).

3. R. A. Friesner *et al.*, Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **49**, 6177-6196 (2006).

4. Z. Wang *et al.*, Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys Chem Chem Phys* **18**, 12964-12975 (2016).

5. N. S. Pagadala, K. Syed, J. Tuszynski, Software for molecular docking: a review. *Biophys Rev* **9**, 91-102 (2017).

6. W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner, R. Farid, Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* **49**, 534-553 (2006).

7. P. A. Ravindranath, S. Forli, D. S. Goodsell, A. J. Olson, M. F. Sanner, AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLoS Comput Biol* **11**, e1004586 (2015).

8. E. B. Miller *et al.*, Reliable and Accurate Solution to the Induced Fit Docking Problem for Protein-Ligand Binding. *J Chem Theory Comput* **17**, 2630-2639 (2021).

9. C. Bissantz, B. Kuhn, M. Stahl, A medicinal chemist's guide to molecular interactions. *J Med Chem* **53**, 5061-5084 (2010).

10. D. Bajusz, A. Rácz, K. Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* **7**, 20-20 (2015).

11. M. C. Jones, Simple boundary correction for kernel density estimation. *Statistics and Computing* **3**, 135-146 (1993).

12. M. J. Sippl, Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des* **7**, 473-501 (1993).

13. M. J. Sippl, Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**, 859-883 (1990).

14. R. Das, D. Baker, Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A* **104**, 14664-14669 (2007).

15. D. J. Hand, K. Yu, Idiot's Bayes: Not So Stupid after All? *International Statistical Review / Revue Internationale de Statistique* **69**, 385-398 (2001).

16. A. Gaulton *et al.*, The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945-D954 (2017).

17. G. M. Sastry, S. L. Dixon, W. Sherman, Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *J Chem Inf Model* **51**, 2455-2466 (2011).

18. G. M. Sastry, V. S. Inakollu, W. Sherman, Boosting virtual screening enrichments with data fusion: coalescing hits from two-dimensional fingerprints, shape, and docking. *J Chem Inf Model* **53**, 1531-1542 (2013).

19. A. E. Cleves, A. N. Jain, Structure- and Ligand-Based Virtual Screening on DUD-E(+): Performance Dependence on Approximations to the Binding Pocket. *J Chem Inf Model* **60**, 4296-4310 (2020).

20. M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **55**, 6582-6594 (2012).

21. M. P. Repasky *et al.*, Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J Comput Aided Mol Des* **26**, 787-799 (2012).

22. M. M. Mysinger, B. K. Shoichet, Rapid context-dependent ligand desolvation in molecular docking. *J Chem Inf Model* **50**, 1561-1573 (2010).

23. L. Chaput, J. Martinez-Sanz, N. Saettel, L. Mouawad, Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *J Cheminform* **8**, 56-56 (2016).