Combined physics- and machine-learning-based method to identify druggable binding sites using SILCS-Hotspots

Erik B. Nordquist,[1,#] Mingtian Zhao,[1,#] Anmol Kumar,[1] Alexander D. MacKerell, Jr.[1*]

[1]Computer Aided Drug Design Center, Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Baltimore, Maryland 21201, United States.

[#]These authors contributed equally to the work.
*Corresponding author: A.D.M. Jr., alex@outerbanks.umaryland.edu

Keywords: Site identification by ligand competitive saturation, protein-ligand interaction, orthosteric, allosteric, computer-aided drug design, CADD, binding site prediction

Author Contributions:
A.D.M. Jr. conceived of and designed the study. All authors contributed to material preparation, data collection and analysis. The first draft of the manuscript was written by E.B.N. and all authors participated in revision of the manuscript.

Abstract

Identifying druggable binding sites on proteins is an important and challenging problem, particularly for cryptic, allosteric binding sites that may not be obvious from X-ray, cryo-EM, or predicted structures. The Site-Identification by Ligand Competitive Saturation (SILCS) method accounts for the flexibility of the target protein using all-atom molecular simulations that include various small molecule solutes in aqueous solution. During the simulations the combination of protein flexibility and comprehensive sampling of the water and solute spatial distributions can identify buried binding pockets absent in experimentally-determined structures. Previously, we reported a method for leveraging the information in the SILCS sampling to identify binding sites (termed Hotspots) of small mono- or bi-cyclic compounds, a subset of which coincide with known binding sites of drug-like molecules. Here we build in that physics-based approach and present a machine learning model for ranking the Hotspots according to the likelihood they can accommodate drug-like molecules (e.g. molecular weight > 200 daltons). In the independent validation set, which includes various enzymes and receptors, our model recalls 65% and 88% of experimentally-validated ligand binding sites in the top 10 and 20 ranked Hotspots, respectively. Furthermore, we show that the model's output Decision Function is a useful metric to predict binding sites and their potential druggability in new targets. Given the utility the SILCS method for ligand discovery and optimization the tools presented represent an important advancement in the identification of orthosteric and allosteric binding sites and the discovery of drug-like molecules targeting those sites.

Introduction

There has been no time like the present for structure-based drug design (SBDD) given the number of protein structures solved at or near atomic resolution currently available in the Protein Data Bank [1], with >200,000 experimental structures and >1,000,000 computed structure models [2], and the >200,000,000 computed structures in the AlphaFold Database [3]. These structural models cover a plethora of potential drug targets [4]. Furthermore, just as GPUs have revolutionized deep-learning models for protein structure prediction [3,5,6], they have also brought all-atom molecular dynamics (MD) simulations of large proteins at meaningful timescales into routine reach [7,8]. This combination, along with advances in our understanding of the molecular nature of disease and the associated growth of personalized medicine, has the potential to produce many new therapeutic agents.

After target identification, the critical first step in the SBDD process is either to identify binding sites of known ligands or identifying candidate sites for virtual screening. Despite the boom in computational power, many widely-used tools for identifying binding sites do not fully account of the conformational flexibility of proteins. The standard methods of protein-ligand binding site prediction rely on extremely efficient methods which generally rely on static structure-based analysis, conventional molecular docking, and/or machine-learning [9]. When a representative structure is available and the binding pocket is relatively well-defined, methods including FTMap [10–12] and FPocket [13] are effective. Some methods employ template based modeling to predict binding sites when only a sequence is known [14–17]. There are many similar-performing

machine-/deep-learning models [9,18] that incorporate sequence-homology, structural features, molecular docking, and consensus to predict ligand binding sites [19–23]. To remain highly computationally efficient, methods reliant on static structures necessarily neglect protein backbone flexibility, thus cannot capture protein allostery or cryptic binding sites [24–28]. In addition, the traditional molecular docking approaches used in available methods [29–33], while efficiently sampling known ligand-protein interactions [12,23], rely on continuum electrostatic models and/or statistical potentials to estimate the energetics of binding. Such methods are limited in their ability to accurately account for the complex balance of enthalpic and entropic costs and desolvation contributions that contribute to ligand binding.

A powerful way to overcome these limitations is through the use of all-atom cosolute MD simulations [34,35]. Cosolute methods are conceptually similar to experimental fragment-based drug design [36,37] wherein proteins are co-crystallized with various small solutes to determine their binding sites [38]. In general, these methods involve solvating the target biomolecule with various small molecules to analyze the distribution of the molecules over the course of the simulation. This approach is widely-employed [39–44] including by MDmix [34,45], pMD-Membrane [46,47], Mix-MD [48–50], SWISH and SWISH-X [51,52], Cosolvent Analysis Toolkit (CAT) [53], and SILCS [35,54,55]. The coarse grain MD cosolute method Colabind was recently released [56], which allows substantially faster sampling than all-atom MD, but with corresponding accuracy sacrifices. The success of the all-atom cosolute MD methods is due to advances in efficient, GPU-enabled molecular dynamics software packages [57–60], combined with consistent improvements in the accuracy of all-atom force fields [61–65], such that accurate sampling of the interactions of solutes with flexible proteins in the presence of explicit atomistic water is readily achievable.

Specifically, the present study is based on the SILCS methodology. SILCS samples the protein conformational ensemble in the presence of multiple solutes and water while alternating between an oscillating chemical potential Grand Canonical Monte Carlo (GCMC) sampling scheme and conventional MD [66,67] that dramatically accelerates the rates of penetration of solutes and water into hydrophobic pockets and other buried cavities. After extensive sampling, the occupancies of the solute molecules and water are converted to functional group-type specific free energy maps, or FragMaps. An example of the FragMaps surrounding the protein TEM-1 β-lactamase is depicted in Figure 1A, and Figure 1B shows molecular renderings of the 8 solutes used in the standard SILCS simulations. These FragMaps form the basis for all subsequent analysis in SILCS, such as performing molecular docking of small molecules in the field of the maps [68,69]. In a previous paper, a method was presented for identifying a comprehensive set of fragment binding sites, or Hotspots, on proteins [70], and subsequently applied to RNA [71]. Although some Hotspots correspond with the known binding sites of small molecules (Figure 1C), it was unclear which Hotspots were really ‘druggable’ using only the previous method. Here we define druggable as being suitable for binding drug-like molecules, such as those with molecular weight (MW) > 200 Da.

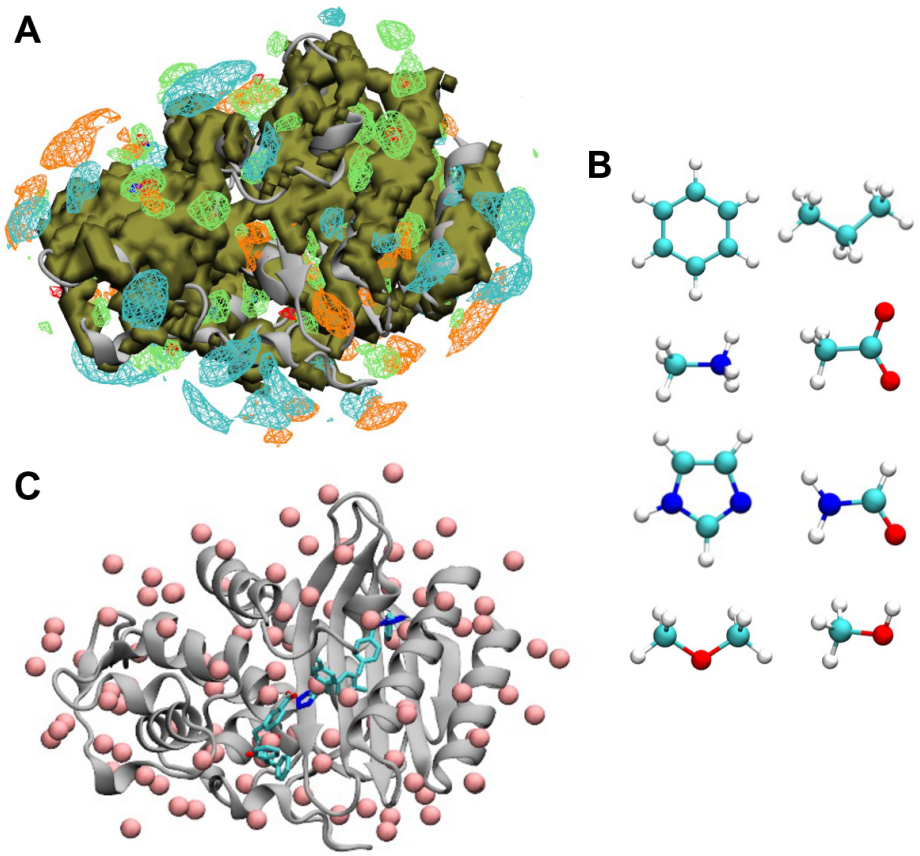


Figure 1: Example SILCS FragMap and Hotspots and depiction of the SILCS solutes. A) TEM-1 β-lactamase is rendered in NewCartoon style (PDB: 1JWP), with the various FragMaps contoured at -1.2 kcal/mol. The green map corresponds to generic apolar carbons (propane and benzene carbon), the red corresponds to hydrogen-bond acceptors, the blue corresponds to hydrogen-bond donors, the cyan corresponds to positive charges (methylammonium nitrogen), the orange corresponds to negative charges (acetate oxygen), gold corresponds to alcohols (methanol oxygen), and the solid tan surface is the Exclusion map. **B)** Depiction of the 8 solutes used in the SILCS GCMC/MD simulations, namely: benzene, propane, methylammonium, acetate, imidazole, formamide, dimethyl ether, and methanol. The molecules are rendered in CPK style, where cyan atoms are carbons, red atoms are oxygen, blue atoms are nitrogen, and white atoms are hydrogen. **C)** Depiction of TEM-1 in NewCartoon style, with the Hotspots rendered as pink spheres, and with the crystallographic ligands from PDBs 1ERO and 1PZO. The ligands are colored as in panel B).

In this study we present a new set of tools to identify Hotspots that contribute to binding sites for drug-like molecules. The method first calculates a range of properties characterizing each Hotspot, which are then used as features in a machine learning (ML) algorithm that predicts the likelihood of each Hotspot participating in a drug-like binding site. For model training Hotspot identified as being in a druggable site were 1) within 12 Å of at least one adjacent Hotspot, 2) within 5 Å of the non-hydrogen atoms of a crystal location of a drug-like ligand, and 3) partially buried. The first criteria assumes that a drug-like molecule is comprised of a minimum of two linked fragments. The second criteria is experimental validation of Hotspots being located in a site

which binds a drug-like molecule through X-ray crystallography. The third criteria is based on the assumption that binding sites are pockets in which the ligands are partially buried [72–74] as determined by an empirical relative buried surface area cutoff described below. For the training set, the developed ML model identifies 76% and 80%, of druggable sites in the top 10 and 20 Hotspots, respectively. In the validation set it recovers 65% and 88% of druggable sites in the top 10 and 20 total Hotspots, respectively.

Methods

SILCS workflow

The overall workflow was to run standard SILCS GCMC/MD simulations of the target proteins solvated in water with a variety of solute molecules (Figure 1B) at 0.25 M for a total of 1 μs as previously described [35,55]. Analysis of the occupancies, and therefore free energy affinities, of each solute gives an atom-type specific 3D affinity map (FragMap) over the entire 3D space of the protein, as well as an Exclusion map containing all the voxels with zero solute or water occupancy (Figure 1A). The PDB identifiers of the protein structures used for the SILCS simulations are provided in Table S1. Note that wherever possible, an apo structure was used for the SILCS simulations; else, a structure with minimal ligand size was used. Any ligands were removed from the structure prior to the simulations. For transmembrane proteins, the membrane orientation was determined using the PPM (Positioning of Proteins in Membranes) webserver [75,76], after which a bilayer composed of 1-palmitoyl-2-oleoyl-sn-*glycero*-3-phosphocholine (POPC) and cholesterol (9:1 ratio) was constructed using the CHARMM-GUI webserver [77,78]. The CHARMM-GUI webserver was also used to generate small missing loops (<12 amino acids) and to adjust the protonation state of titratable residues [77,78]. The protonation state of titratable residues at pH 7.0 was determined using PropKa3 [79]. The FragMaps were obtained using SILCS software version 2019 (SilcsBio LLC) and Gromacs version 2019, except for ANGPTL4, TEM-1, and $GABA_BR$, for which SILCS software version 2023 [80] and Gromacs version 2022 were used [57,58].

After calculating the FragMaps, we performed the SILCS-Hotspots calculation as described in our previous work [70]. The Hotspots calculation consists of comprehensively docking a library 90 mono- and bicyclic fragments [81] with MW < 190 Da into the FragMaps and Exclusion map. Then two rounds of clustering are performed to identify binding sites that include one or more of the fragments (Figure 1C). Each original Hotspot is then defined by the number of fragments in that site and the LGFE scores of those fragments from which features such as the minimum (e.g. most favorable) LGFE or mean LGFE over all the fragments in that Hotspot are calculated and used for ranking. The SILCS-Hotspots calculations were run using version 2019, except for all proteins in the validation set, where version 2023 was used [80]. The SILCS-Hotspots docking performed for this study utilized a new GPU implementation of SILCS-MC docking (Zhao and MacKerell, *manuscript in preparation*).

Additional characterization of Hotspots as potential druggable binding sites was performed by screening a database of 348 FDA-approved compounds at selected Hotspots. The docking was

carried out in a 5 Å radius sphere centered on the Hotspot. After docking, each Hotspot was characterized by the average LGFE and relative buried surface area (rBSA) for the top twenty molecules, ranked by the LGFE. rBSA is defined as the ratio of the solvent accessible surface area of the ligand alone relative to that of the ligand in the presence of the protein, such that 100% rBSA indicates a fully buried ligand with no solvent accessible surface area (SASA). The SASA of the ligand in both the presence and absence of the protein was based on the conformation of the ligand from the SILCS-MC docking. The 348 compound FDA database was extracted from an initial set of FDA-approved molecules derived from the online databases DrugBank [82] and Drugs@FDA [83]. An initial filter was applied to select only molecules with MW between 250 and 500 Da. To reduce the dimensionality while maintaining the diversity of the molecules in the FDA set, we clustered the dataset with Morgan fingerprints using a radius of 2 and Tanimoto similarity index of 0.3, then selected a representative molecule from each cluster, yielding a total of 380 molecules. The final set of 348 molecules was arrived at by manually removing outliers in the number of rotatable bonds or hydrophobic groups. The FDA database is available in sdf and pdf formats on GitHub at https://github.com/mackerell-lab/FDA-compounds-SILCS-Hotspots-SI. The FDA dataset curation and generation of the pdf table of 2D molecular images was done with the python API for RDKit [84].

Calculation of new analysis features

The Hotspot analysis workflow to calculate features for ML model development consists of three keys steps: cluster adjacent Hotspots within some user-tunable cutoff distance, collect various properties of the individual Hotspots and Hotspot clusters, and then use those features to develop the ML model to identify Hotspots at the binding sites of drug-like molecules. Here we define a Hotspot cluster as containing all the Hotspots within 12 Å of each Hotspot (centroid), because the maximum distance between two neighboring Hotspots in the training set is 11.6 Å. Based on this definition, each individual Hotspot can be a member of multiple Hotspot clusters, though each Hotspot is the centroid of just one Hotspot cluster with the features based on that cluster assigned to the centroid Hotspot.

The new features include the number of protein non-hydrogen atoms in the input PDB file within a user-defined radius of each Hotspot (default 3 Å), the SASA and volume of each Hotspot (using a 3 Å radius for the Hotspots), the SASA and volume of the Hotspot clusters, the distances between Hotspots in the cluster, as well as various statistical measures (e.g. mean, minimum, and maximum values) of the distribution of these properties over the Hotspot cluster (Table 1). As a feature we wanted the calculation of the SASA of a Hotspots to account for the protein flexibility that is included in the SILCS simulations. Accordingly, in addition to using the original crystal structure used for the SILCS simulations for the SASA calculation, an “Exclusion-map HS SASA” was calculated where the solvent-accessibility of the Hotspot (default radius 5 Å) was relative to voxels that were included in the SILCS Exclusion map rather than the standard use of the positions of the protein atoms. The different Hotspot radii (3 Å for use with protein PDB file and 5 Å for use with Exclusion map) adjusts for the smaller size of an Exclusion map relative to a corresponding protein. All SASA calculations used a solvent probe radius of 1.4 Å. Additional features using the Exclusion map were calculated as described in Table 1.

The code to calculate the SASA of Hotspots with respect to the Exclusion map was built on the freeSASA [85] package in python. The freeSASA code was modified to allow for non-default input atomic radii for the Hotspots and Exclusion map voxels. In addition, the SASA of Hotspot clusters was calculated based on the SASA of all the Hotspots in the cluster (default radius 5 Å). The Exclusion map is represented as a set of spheres of radius 1 Å sitting on 1 $Å^3$ grid voxels. To calculate the volume of the Hotspot clusters not within the protein or Exclusion map a Monte Carlo integration algorithm we implemented. The calculation of the SASA and volume of the Hotspot clusters requires substantial CPU time, and so the algorithms were parallelized with numba [86].

Table 1: Names and descriptions of the features calculated by the new SILCS-Hotspots workflow. The radius of each Hotspot for the SASA calculations can be user-defined separately for the protein coordinates and Exclusion map calculations; defaults are 3 Å and 5 Å, respectively. LGFE stands for Ligand Grid Free Energy of the fragments located in each Hotspot and SASA stands for solvent-accessible surface area.

Name	Description
Orig	Mean LGFE of each Hotspot (Original ranking metric).
Min	Minimum LGFE of each Hotspot cluster.
Ave	Average LGFE of each Hotspot cluster.
NFrag	Number of drug-like fragments in each Hotspot.
N_Heavy_Atoms	Number of protein non-hydrogen atoms within 3 Å of each Hotspot.
N_BBone_Atoms	Number of protein backbone atoms within 3 Å of each Hotspot.
PDB_SASA	SASA of protein atoms occluded by each Hotspot.
Excl_SASA	SASA of protein Exclusion map occluded by each Hotspot.
PDB_HS_SASA	SASA of each Hotspot occluded by the protein.
Excl_HS_SASA	SASA of each Hotspot occluded by the Exclusion map.
Adj_PDB_SASA	SASA of protein atoms occluded by each Hotspot cluster.
Adj_PDB_HS_SASA	SASA of each Hotspot cluster occluded by the protein.
Relative_Adj_SASA	The relative SASA of each Hotspot cluster defined as the ratio of SASA of the Hotspot cluster in the presence of the protein PDB to total SASA of the Hotspot cluster without the protein.
Vol	Volume of each Hotspot excluding the volume overlapping with protein atoms.
Excl_Vol	Volume of each Hotspot, excluding the volume overlapping with the SILCS Exclusion map.
MinDist	Minimum distance between each Hotspot and the other Hotspots in the cluster.
MaxDist	Maximum distance between each Hotspot and the other Hotspots in the cluster.
MidDist	Median distance between each Hotspot and the other Hotspots in the cluster.
AvgDist	Average distance between each Hotspot and the other Hotspots in the cluster.

Sum_<feature>	Sum of <feature> over the Hotspot cluster.
Mean_<feature>	Mean of <feature> over the Hotspot cluster. This is sum divided by the number of Hotspots in the cluster.
Min_<feature>	Minimum of <feature> among Hotspots in the cluster. For example, the value of the most favorable LGFE of the Hotspots in the cluster.
Max_<feature>	Maximum of <feature> among Hotspots in the cluster. For example, the value of the Hotspot with largest Volume in the cluster.

Training and validation data set curation

The training set is constructed from the seven protein systems from the previous SILCS-Hotspots paper [70]: Cyclin-dependent kinase 2 (CDK2) in both active and inactive states [87,88], Extracellular-signal-regulated kinase 5 (ERK5) [89], Protein tyrosine phosphatase 1b (PTP1B) [90–93], Androgen receptor [94,95], and three G-protein coupled receptors (GPCRs), namely G protein-coupled receptor 40 (GPR40) [96,97], M2 Muscarinic receptor [98,99], and β2 Adrenergic receptor [100,101]. The validation set is comprised of ten proteins, seven of which we recycle from previous SILCS-MC publications [68,69], namely: P38 mitogen-activated protein kinase [102,103], Farnesoid X bile acid receptor (FXR) [104], β-Secretase 1 (BACE1) [105,106], tRNA methyl transferase (TrmD) [107], Myeloid cell leukemia 1 (MCL1) [108,109], Heat-shock protein 90 kDa (Hsp90) [36], and Thrombin [110]. To those we added the C-terminal domain of the lipid-binding protein angiopoietin-like 4 (ANGPTL4) [111], TEM-1 β-lactamase [112–114], and GPCR γ-aminobutyric acid receptor ($GABA_BR$) in both active and inactive states [115–117].

For each protein system, we identified relevant crystal structures where there is a drug-like ligand bound and aligned these structures to the structure used to generate the SILCS FragMaps. Hotspots within 5 Å of a ligand non-hydrogen atom are classified as a “true hit”. In addition, a Hotspot must be within 12 Å of at least one other Hotspot to be a true hit, and the 12 Å path must be unobstructed by any Exclusion map voxels. In the training set, if a Hotspot is within 5 Å of more than one ligand, it is counted for both ligands to reflect its importance in identifying more than one distinct ligand binding site. The PDB [1] and D3R [118] structures used are listed in Table S1, and the Hotspots considered true hits are listed in Table S2. In each system, there may be several ligands bound in similar positions available in different PDB files, but only one such ligand was selected to represent that binding site. In a few cases, there are Hotspots which are within 5 Å of the ligand but are located on the surface of the protein above the ligand binding site. Figure S1 depicts one such example, Hotspot 25 in the ERK5 system, which is within 5 Å of the ligand but largely solvent-exposed. As one of our criteria of druggable binding sites was that they are partially buried sites, we removed outlying Hotspots with greater than 300 $Å^2$ Exclusion-map HS SASA (Figure S2), as these sites may not be suitable for binding drug-like molecules. This empirical cutoff corresponds to ~42% rBSA.

Evaluation of model performance

To evaluate the developed models, we calculated precision, recall, weighted F_1, and binding site recall using the Hotspots identified as true hits. Evaluating a Hotspot classification model requires

ranking the Hotspots, then selecting a cutoff, such as taking all Hotspots with LGFE < 0 or taking the top N Hotspots. For a given cutoff, precision is the ratio of true hits to the total number of Hotspots up to and including the cutoff, while recall is the ratio of true hits up to and including the cutoff to the total number of experimentally verified hits. For example, if a protein has four total experimentally verified hits, two of which are identified with a cutoff at ten Hotspots, the precision is 2/10 = 0.2 and the recall is 2/4 = 0.5. The weighted F_1 statistic is the population-weighted harmonic mean of precision and recall. This is important because it accounts for the low proportion of Hotspots which are true hits: only 7% of all the Hotspots in the training set are experimentally verified hits and only 2% in the test set. Accordingly, a random predictor would have a precision of ~0.02 for the validation set, which is a useful comparison when evaluating the precision of a model (e.g., 0.2 for the validation set example represents a ten-fold increase over a random predictor). In addition, binding site recall was calculated to compare the performance of the models on the practical problem of identifying at least one Hotspot per ligand. Binding site recall is defined as the ratio of identified ligand binding sites to the total number of experimentally identified ligand binding sites for that protein. A ligand binding site is identified once a single Hotspot within 5 Å of that ligand is identified above a given cutoff. Accordingly, the maximum number of ligand binding sites is equivalent to the total number of experimentally identified ligand binding sites although the total number of Hotspots defined as true hits may be greater than the total number of experimentally identified ligand binding sites. Below the total number of experimentally verified hits is indicated as "# Sites" in the tables.

We note that the calculated performance of the models may underestimate their true performance, since we base our true hits on crystallographically-identified ligand binding sites. It is possible that some of the Hotspots occupy sites for which a ligand indeed exists but has not yet been identified. Accordingly, the number of true hits may actually be higher than is calculated in the present study.

Table 2: Linear SVM hyperparameters. Descriptions of hyperparameters are adapted from the sci-kit learn library documentation [119]. Where multiple hyperparameter values were tested, the bolded parameter value was selected in the final model.

Hyperparameter	Values	Description
C	1e-4, **1e-3**, 1e-2, 1e-1	Regularization strength, which is proportional to 1/C. Regularization provides a way to reduce the final model complexity.
intercept_scaling	1e1, **1e2**, 1e3	Reduce impact of C on intercept fitting.
loss	**hinge**, squared_hinge	The loss function used in training the classification model. Hinge loss is the standard for SVM.
penalty	l2	Regularization penalty, the l2-norm.
fit_intercept	True	The input feature vector includes a scalar intercept term.
dual	auto	Automatically select optimization algorithm where the optimal choice depends on the relative numbers of features versus samples, and some choices of other

		parameters. Auto will be the default in scikit-learn version 1.5.
max_iter	1e8	Maximum number of iterations of the linear solver.
tol	1e-4	Tolerance criterion for convergence of the linear solver.
class_weight	balanced	A weight for the regularization parameter C, in this case inversely proportional to the class proportion.

Machine learning methods

Given the limited size of the dataset, we focused our efforts on Support Vector Machine (SVM) and Random Forest classifier models. Random forest models and SVM with polynomial kernels of degree > 1 resulted in over-training (Table S3). While all models generated reasonable average weighted F1 statistics on the 5-fold cross-validation (CV), there is a significant degradation in performance between the average CV recall and the recall after fitting on the whole training dataset (single-fit) (Table S3). In comparison, the linear kernel SVM had similar recall between a single-fit and the average CV recall (Table S3), so we selected the linear kernel SVM model and fully trained its hyperparameters (Table 2). To optimize the performance of the SVM, we performed standardization ($(\vec{X} - \mu)/\sigma$) of each feature, then performed principal component analysis (PCA) on these features and used the principal components as inputs for all subsequent models. This ensures the inputs are all mutually orthogonal. The hyperparameters were optimized using a grid search of the parameter space described in Table 2. Each round of grid search was performed using 5-fold cross-validation, and the selection of optimal parameters was made based on the weighted F_1 statistic. Subsequently we performed recursive feature elimination [120] to identify the optimal number of input principal components and reduce the risk of overfitting by reducing the dimensionality of the inputs (Figure S3). The first 22 principal components were selected, corresponding to the maximum weighted F_1 in Figure S3. The final model hyperparameters are indicated in Table 2 with bold text. These were used to train the final model on the whole training dataset; all subsequent results in the paper are based on this model. A key output of an SVM model is the Decision Function, defined as the distance a Hotspot lies from the SVM's decision boundary and can be interpreted as the confidence that a given Hotspot corresponds to a true hit and, therefore, likely located within 5 Å of a crystallographic ligand binding site [121,122]. The Decision Function is positive for higher confidence, and negative for confidence that the Hotspot is not a suitable binding site. The machine learning scripts were written using the scikit-learn [119] and pandas [123] python libraries. All 3D molecular renderings were generated using VMD version 1.9.3 [124], and all plots were created with the python library matplotlib [125] using the accessible color sequences of Petroff [126].

Results

The present study involved the development of a ML model to predict the probabilities that SILCS Hotspots are located in druggable binding sites, based on those sites which are occupied by drug-like molecules (MW > 200 Da) as identified in crystallographic studies. The model builds on the

previously reported SILCS Hotspots based on fragment docking into the SILCS FragMaps combined with additional features for each Hotspot used in ML model development targeting the known druggable sites. The training set included seven proteins while the validation set included ten proteins. As presented, the developed ML model predicts those Hotspots with a high probability of defining druggable sites based on a quantitative ranking score that may be applied to new systems.

Of the ten proteins in the validation set, seven were used in previous SILCS-MC benchmarking studies, and as such each contain a single orthosteric binding site [68,69]. In addition, allosteric ligands were identified for the validation set proteins where available. The full details of the structures and ligands used in both the training and validation sets is described in Table S1, but some additional details are given here. For P38 we selected the allosteric inhibitor ligand BIRB 796 bound in PDB 1KV2 [103]. Note that for the purposes of this study BIRB 796 may be only partially allosteric, as it also overlaps with orthosteric site defined by the ligand in PDB 3FLS [102]. We collected four additional systems, ANGPTL4, TEM-1, and $GABA_BR$ in both the active and inactive state. For ANGPTL4, we selected a structure with glycerol bound for the SILCS simulations (PDB: 6U0A) and used a Palmitic acid-bound structure for assessing which Hotspots are in a ligand binding pocket (PDB: 6U1U) [111]. TEM-1 was selected because of its cryptic allosteric binding site [24,113], which is absent in the apo structure we used for the SILCS simulation (PDB: 1JWP) [112]. For the $GABA_BR$, as previously described for the CDK2 system [70], we collected two sets of FragMaps corresponding to the active (PDB: 7CA3, allosteric modulator BHFF) and inactive (PDB: 7CA5, apo) conformations. Each FragMap set was used to identify ligands from separate PDBs (6UO8 and 7C7Q). This allows us to assess if the individual FragMap sets allows the prediction of binding sites from either state of the protein. However, the large interdomain rearrangement of the transmembrane (TM) helices between active and inactive states [115] disallows predicting the allosteric binding site present in the active conformation using the inactive conformation with the an equilibrium MD method such as SILCS.

New Hotspot properties improve the identification of druggable Hotspot clusters

To generate features of model development we calculated numerous properties of individual Hotspots including features based on the Hotspot clusters of which they are the centroid Hotspot. The previously published Hotspot ranking (Orig in Table 1) was based purely on the mean LGFE over all the specific fragments present in each Hotspot [70]. As discussed above a single Hotspot represents a binding site for fragments (MW < 200 Da) which are generally smaller than most drugs. The ranking of all the Hotspots using the mean LGFE, as well as being within 12 Å of at least one other Hotspot, is shown in Figure S4, which highlights that for many proteins in the training set, the mean LGFE has limited predictive power. To evaluate the ability of the LGFE to predict the binding sites for drug-like molecules, the binding site recall was calculated with respect to the crystallographic ligand poses. The mean LGFE ranking captures 40%, 44%, and 80% experimental binding sites in the top 10, 20, and 40 Hotspots, respectively, over the training set protein systems (Table 3). While the mean LGFE score used to rank the original Hotspots is somewhat successful as a predictor of the Hotspot being a drug-like molecule binding site in some

systems, significant improvements can be made by incorporating additional features in ML model development, as shown below.

Table 3: Training set binding site recall in the top 10, 20, and 40 Hotspots. The recalls are reported for three models: Hotspot LGFE, Exclusion-map HS SASA, and the SVM model. Binding site recall is the ratio of unique ligands within 5 Å of an experimentally-validated ligand binding site over the total number of such sites for that protein.

Protein Name	# Sites	Top 10	Top 20	Top 40
LGFE (Original ranking metric)				
CDK2 Active	6	0.67	0.67	0.67
CDK2 Inactive	6	0.33	0.33	0.83
ERK5	2	0.50	0.50	1.00
PTP1B	3	0.33	0.33	1.00
β2 Adrenergic	2	0.00	0.50	0.50
GPR40	2	0.00	0.00	0.00
M2 Muscarinic	2	0.50	0.50	1.00
Androgen	2	0.50	0.50	1.00
Total	**25**	**0.40**	**0.44**	**0.80**
Exclusion-map HS SASA				
CDK2 Active	6	0.50	0.83	0.83
CDK2 Inactive	6	1.00	1.00	1.00
ERK5	2	1.00	1.00	1.00
PTP1B	3	0.33	0.33	1.00
β2 Adrenergic	2	0.50	1.00	1.00
GPR40	2	1.00	1.00	1.00
M2 Muscarinic	2	0.50	1.00	1.00
Androgen	2	1.00	1.00	1.00
Total	**25**	**0.76**	**0.88**	**0.96**
SVM model				
CDK2 Active	6	0.50	0.50	0.83
CDK2 Inactive	6	1.00	1.00	1.00
ERK5	2	1.00	1.00	1.00
PTP1B	3	0.33	0.33	1.00
β2 Adrenergic	2	1.00	1.00	1.00
GPR40	2	0.50	1.00	1.00
M2 Muscarinic	2	1.00	1.00	1.00
Androgen	2	1.00	1.00	1.00
Total	**25**	**0.76**	**0.80**	**0.96**

When designing new features, we considered another limitation in the original ranking where the mean LGFE scores of Hotspots with high solvent exposure are often quite favorable. To account

for the degree of solvent accessibility required to make a binding site more favorable for drug-like molecules as well as consider the size of drug-like molecules, we designed features related to the degree of solvent accessibility of the Hotspot, the volume of the Hotspot not occluded by the protein, the number of Hotspots in a cluster, and the totals of these in each Hotspot cluster. Figure 2 shows the ranking based on Exclusion-map HS SASA for all Hotspots also within 12 Å of at least one other Hotspot. Those Hotspots within 5 Å of a drug-like molecule from crystallographic structures are shown as large circles. The Exclusion-map HS SASA ranking greatly improves the selection of Hotspots close to drug-like molecules. Table 3 shows that the mean binding site recalls have increased over that of the original LGFE Hotspot ranking to 76%, 88%, and 96% for the top 10, 20, and 40 Hotspots, respectively. While accounting for the SASA and presence of at least one adjacent Hotspot greatly improves the identification of druggable Hotspots, there is variability over the training set proteins. For example, with PTP1B or the M2 Muscarinic receptor, these two criteria alone aren't particularly effective. Accordingly, we reasoned that using a machine learning classifier method to combine the information from many features should provide a better ranking. If the model is trained with cross-validation, it could also lead to robust generalization across a range of protein systems.

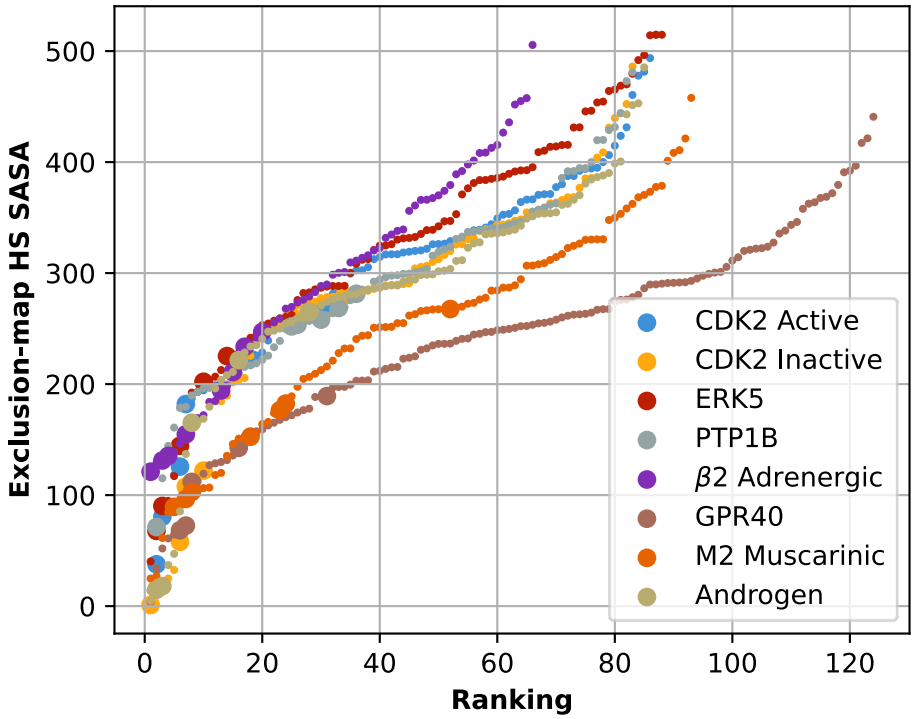


Figure 2: Ranking based on Exclusion-map HS SASA of individual Hotspots with a minimum of one adjacent Hotspot within 12 Å. The larger circles denote Hotspots within 5 Å of a non-hydrogen atom of a drug-like compound bound to the proteins.

Machine learning model improves identification of druggable Hotspots

While the individual feature of Exclusion-map HS SASA, and presence of adjacent Hotspots, contain substantial information about whether a Hotspot is located in a drug binding site, an appropriately selected and trained machine learning model should better integrate the information from a wider range of features and improve the model's accuracy as well as generalizability. Accordingly, we trained several machine learning models using the features listed in Table 1, as shown in the supporting information (Table S3). From that analysis we selected the SVM classifier with a linear kernel as implemented in scikit-learn library [119,121]. The final model improves the predictive power over the untrained features alone, as shown in Figure 3. Figure 3A shows the model's Hotspot ranking for each system and highlights the Hotspots which are within 5 Å of a ligand. Figure 3B presents a precision-recall curve for the training data and includes comparison to two untrained models, the original mean LGFE of all the molecules in the Hotspot, and Hotspot Exclusion-map HS SASA. Precision-recall curves show the change in precision over increasing recall, which corresponds to lowering the level of the cutoff above which a Hotspot is predicted to be a hit. Figure 3C shows the merged ranking of Hotspots from all proteins, for each of the three models, corresponding to Figure 3B. To facilitate easy comparison, the LGFE and Exclusion-map HS SASA were inverted, and then the LGFE, Exclusion-map HS SASA and SVM Decision Function were Min-Max normalized ($(\vec{X} - min)/(max - min)$) so that they all predict maximal druggability at 1 and minimal druggability at 0 (Figure 3C). Figure 3C shows that generally, the SVM model has the greatest density of true hits in the lower rankings; we note that the relative ranking within each metric is important in Figure 3C, not the position of the curves with respect to one another (Figure 3C). Indeed, the SVM model has superior performance to the other models, demonstrated by the larger area under the precision-recall curve (AUC) for the SVM model (0.42) as compared to the LGFE (0.08), Exclusion-map HS SASA (0.29), and the random model (0.07) (Figure 3B). The SVM model's AUC increased six-fold from that of the random model (0.07 to 0.42) (Figure 3B).

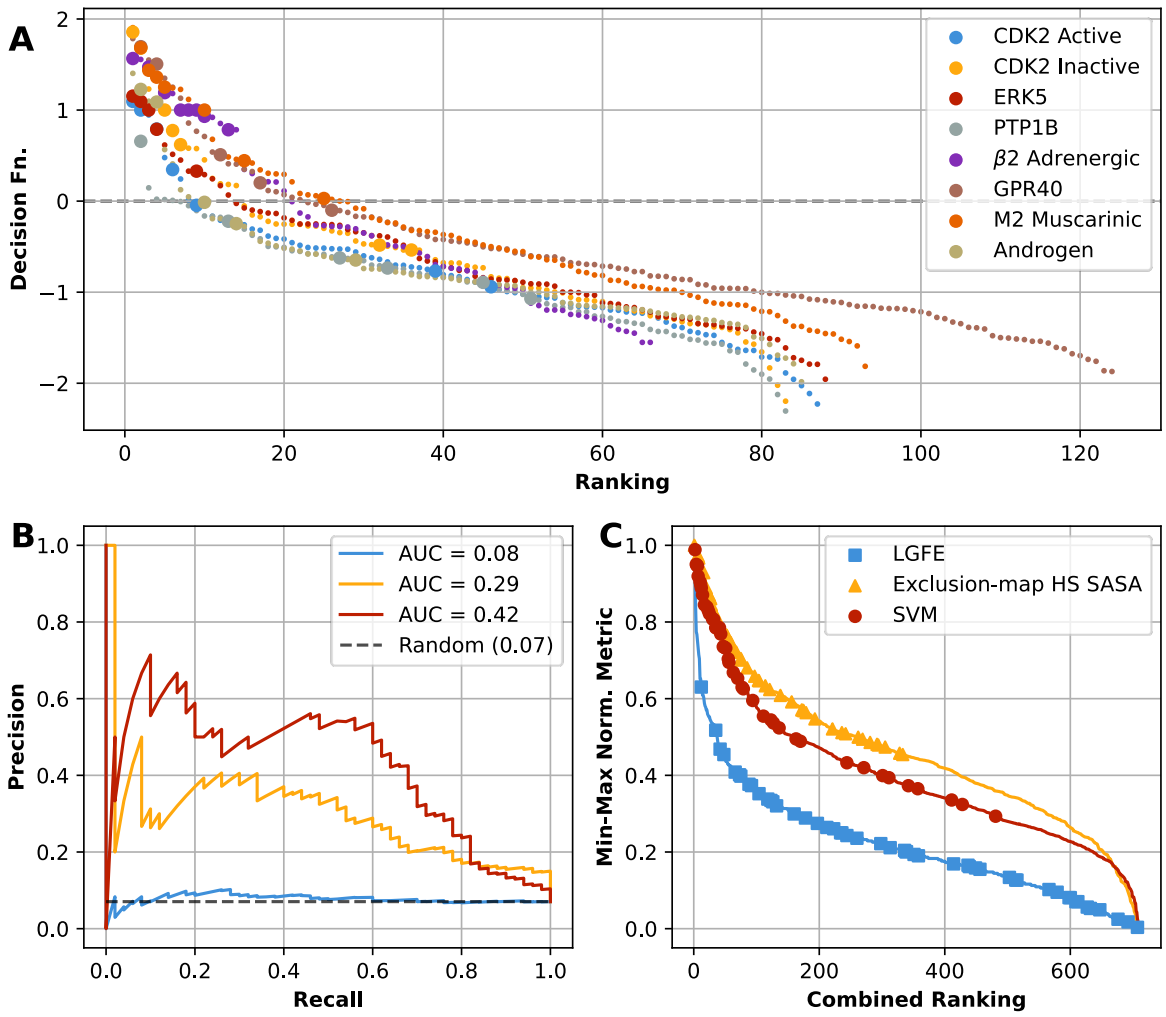


Figure 3: Performance of final model on the training set. A) Ranking of each protein's Hotspots by the final SVM model's Decision Function with Hotspots within 5 Å of the non-hydrogen atoms of known drug-like molecules (true hits) shown as large circles. **B)** Precision-Recall curves of the original LGFE (blue), Exclusion-map HS SASA (yellow), and SVM Decision function (red) models. AUC stands for area under the curve, and the black dashed line reflects the ratio of hits to total Hotspots, or the expected AUC for a random model. **C)** Ranking of all training set Hotspots using the Min-Max normalized ranking metric in which the range for each metric is set from 0 to 1 using $(\vec{X} - Min)/(Min - Max)$. Hotspots within 12 Å of at least one other Hotspot from all proteins are combined and plotted as a continuous curve. Prior to Min-Max normalization the Exclusion-map HS SASA and LGFE were inverted to allow direct comparison to the SVM Decision Function. The markers denote hits, as in panel A).

In practical terms, the model identifies 80% of ligand binding sites in the top 20 Hotspots (Table 3). This is impressive performance given the challenging nature of the problem since the binding sites identified here include both allosteric and orthosteric sites based on ligands exclusively absent in the crystal structures used in the SILCS simulations [70]. In the top 20 Hotspots the SVM model fails to identify three out of twenty ligand sites (Table 3). One is a relatively solvent-

exposed site on the protein PTP1B, and so are unusual in our training set and challenging to the model. The remaining three missing ligands belong the CDK2 kinase in the active state. Two of these missing sites share the same Hotspot ranked 34th by the SVM model (Table S2). The last missing site has no Hotspot within 5 Å (Table S2), as highlighted in the previous paper [70]. Missing this binding site is therefore not a limitation of the ranking method itself but the sampling of that particular pocket using the CDK2 Active structure 3MY5 with the SILCS method. While the system PTP1B, which has largely surface-exposed binding sites, remains challenging even for the SVM model, the model prediction generally improves across all systems (Figure 3B), and may be more generalizable than a single feature such as the Exclusion-map HS SASA, which happens to perform well on this particular dataset. However, an unbiased assessment of the final model must rely on an independent dataset.

Validation of the final SVM model

To validate the final model, we gathered a set of proteins independent of the training set, as discussed in the Methods. The details of the ligands analyzed for each system are listed in Table S1 and Table S2. The results for predicting all Hotspots near crystal ligands using the SVM model are given in Figure 4A, and a comparison of the model's performance to the untrained LGFE and Exclusion-map HS SASA models are given in Figure 4B and Figure 4C. The results for predicting individual binding sites is given in Table 4. There is a six-fold increase in precision-recall AUC between the random model and the SVM model in the validation set (0.02 to 0.12), the same as was in the training set (0.07 to 0.42), which suggests that the model was not overfit to the training data. More practically, the model recalls 65% of ligand binding sites in the top 10, and 88% of sites in the top 20 Hotspots, respectively (Table 4). The SVM model's Decision Function outperforms the untrained models as demonstrated by the increased precision-recall AUC (Figure 4B). Notably, the Exclusion-map HS SASA ranking performs worse in the validation set than in the test set, suggesting that the trained SVM model is more generalizable than either individual feature alone (Figure 4B). Furthermore, although the Exclusion-map HS SASA ranking performed slightly better at binding site recall on the training set (Table 3, top 20), the SVM model performs better than either untrained model on the validation test (Table 4). Overall, the results argue that the model is not over-fitted to our limited training data, and that the model can predict druggable binding sites across a range of proteins with reasonable accuracy.

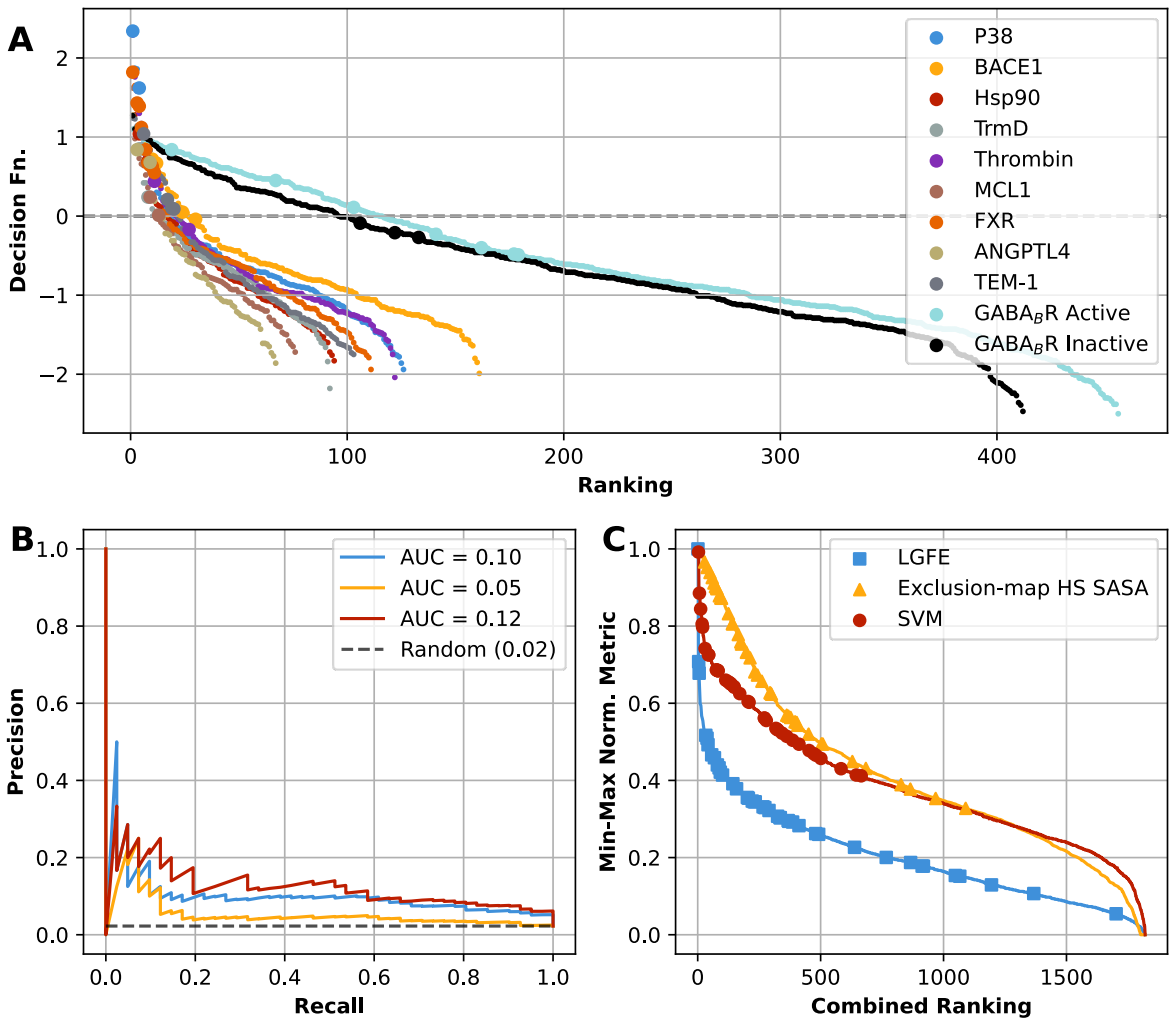


Figure 3: Performance of final model on the training set. A) Ranking of each protein's Hotspots by the final SVM model's Decision Function with Hotspots within 5 Å of the non-hydrogen atoms of known drug-like molecules (true hits) shown as large circles. **B)** Precision-Recall curves of the original LGFE (blue), Exclusion-map HS SASA (yellow), and SVM Decision Function (red) models. AUC stands for area under the curve, and the black dashed line reflects the ratio of hits to total Hotspots, or the expected AUC for a random model. **C)** Ranking of all training set Hotspots using the Min-Max normalized ranking metric in which the range for each metric is set from 0 to 1 using $(\vec{X} - Min)/(Min - Max)$. Hotspots within 12 Å of at least one other Hotspot from all proteins are combined and plotted as a continuous curve. Prior to Min-Max normalization the Exclusion-map HS SASA and LGFE were inverted to allow direct comparison to the SVM Decision Function. The markers denote hits, as in panel A).

While the model performs quite well across most of the validation set, it performs poorly on the heterodimer $GABA_B$ Receptor in both active and inactive states. It captures one of nine true hit Hotspots in the active state and zero of three in the inactive, which corresponds to identifying only one of three ligand binding sites (Table 4). The orthosteric binding site (2C0, Baclofen) was not identified in $GABA_B$R Inactive, despite being identified in the $GABA_B$R Active simulations. In the

simulations of the inactive state, the orthosteric binding site is highly solvent exposed, and the Hotspots' Exclusion-map rBSA values range from 1% to 40%, less than the empirical 42% cutoff used to define the training set (see Methods). This makes this site an outlier compared to the data used to train the model. However, another challenge is that the $GABA_BR$ heterodimer is much larger than the other proteins considered. A total of 416 Hotspots were identified or about four- to five-times the number in the training set systems. To account for this, we ranked the Hotspots near the extracellular part of the $GABA_{B1}$ subunit. From among these 118 Hotspots, a Hotspot near the ligand 2C0 is now ranked in 33rd, or in the top 40 (Table S2). Finally, the missing site in the $GABA_BR$ active state is an allosteric binding site between the two TM domains and directly interacts with lipids in the bilayer during the SILCS GCMC/MD simulations (Figure S5), making this site uniquely challenging to identify with our method. We ranked all the Hotspots in the TM region and found that the first two Hotspots near the ligand are only ranked 50th and 57th, respectively (Table S2). A future improvement of the model could explicitly account for lipid interactions at membrane-protein interfaces.

Table 4: Validation set binding site recall in the top 10, 20, and 40 Hotspots. The recalls are reported for three models, the LGFE, Exclusion-map HS SASA of the Hotspot, and SVM model's Decision Function. Binding site recall is the ratio of the total number of ligand binding sites within 5 Å of a Hotspot in the top N Hotspots. A site is identified when at least one Hotspot corresponding to a ligand is selected in the top N.

Proteins Name	# Sites	Top 10	Top 20	Top 40
		LGFE		
P38	2	0.50	1.00	1.00
BACE1	1	1.00	1.00	1.00
Hsp90	1	1.00	1.00	1.00
TrmD	1	1.00	1.00	1.00
Thrombin	1	1.00	1.00	1.00
MCL1	1	1.00	1.00	1.00
FXR	3	0.67	0.67	1.00
ANGPTL4	1	1.00	1.00	1.00
TEM1	3	0.33	0.33	0.33
$GABA_BR$ Active	2	0.00	0.50	1.00
$GABA_BR$ Inactive	1	0.00	0.00	1.00
Total	**17**	**0.59**	**0.71**	**0.82**
		Exclusion-map HS SASA		
P38	2	1.00	1.00	1.00
BACE1	1	0.00	1.00	1.00
Hsp90	1	1.00	1.00	1.00

TrmD	1	1.00	1.00	1.00
Thrombin	1	0.00	1.00	1.00
MCL1	1	1.00	1.00	1.00
FXR	3	0.67	1.00	1.00
ANGPTL4	1	1.00	1.00	1.00
TEM1	3	0.33	0.33	0.67
$GABA_BR$ Active	2	0.00	0.00	0.00
$GABA_BR$ Inactive	1	0.00	0.00	0.00
Total	**17**	**0.53**	**0.71**	**0.88**
		SVM model		
P38	2	1.00	1.00	1.00
BACE1	1	1.00	1.00	1.00
Hsp90	1	1.00	1.00	1.00
TrmD	1	1.00	1.00	1.00
Thrombin	1	0.00	1.00	1.00
MCL1	1	1.00	1.00	1.00
FXR	3	1.00	1.00	1.00
ANGPTL4	1	1.00	1.00	1.00
TEM1	3	0.33	1.00	1.00
$GABA_BR$ Active	2	0.00	0.50	0.50
$GABA_BR$ Inactive	1	0.00	0.00	0.00
Total	**17**	**0.65**	**0.88**	**0.88**

Model's Decision Function is a predictor of Hotspot druggability

While the SVM model highly ranks most Hotspots corresponding to known drug-like ligand binding sites in the top 20 (Table 4), there are a number of high-ranking Hotspots that do not correspond to known binding sites. Because some may be associated with true drug-like binding sites for which no ligand has yet experimentally been identified, we hypothesized that the most highly-ranked Hotspots should be more druggable than those ranked poorly. To test this hypothesis, we selected two proteins in the validation set, namely TEM-1 and $GABA_BR$ Active, and docked the FDA database of 348 compounds at the Hotspots ranked 1-10, 91-100, and for $GABA_BR$ 391-400. These Hotspots represent the most and least-druggable according to the SVM model's ranking. For each Hotspot we report the mean LGFE and rBSA for the top twenty compounds ranked by LGFE (Table S4). The mean LGFE scaled by mean rBSA (mean LGFE x rBSA), where 100% rBSA is equivalent to 1.0, was used as a measure of Hotspot druggability. This assumes that druggable sites have favorable LGFE scores with high rBSA values, associated with high affinity and with buried sites, respectively. We plotted the final SVM model's Decision Function against the mean LGFE x rBSA for these Hotspots in Figure 5. In general, it shows the expected

anti-correlation between Hotspot predicted druggability, based on larger positive SVM Decision Function values and more negative LGFE x rBSA scores corresponding to druggable sites.

The SVM Decision Function's anti-correlation with the LGFE x rBSA druggability scores accounts for slightly different trends in LGFE and rBSA individually between GABABR and TEM-1. For the TEM-1 Hotspots, the top 10 Hotspots have substantially higher average rBSA and the average LGFE values of Hotspots 91-100 decrease only slightly, whereas in $GABA_BR$ Active the average LGFE score decreases substantially while the average rBSA values decrease slightly (Table S4). The fact that $GABA_BR$ Hotspots appear far more druggable, having more favorable average LGFE and lower rBSA, despite only considering Hotspots 91-100 is due to that system have significantly more Hotspots due to its larger size than the TEM-1 system. Importantly there are large differences between the SVM Decision Function scores between Hotspots 1-10 and 91-100 for both proteins, indicating the ability to discriminate between sites in difference proteins. In addition, it is notable that with both proteins the SVM Decision Function scores for the top Hotspots are similar, ~1.0, indicating that the SVM values may be applied directly to new proteins for the selection of potential druggable sites. Finally, the lack of a stronger anti-correlation between SVM Decision Function scores and the Mean LGFE x rBSA druggability scores may be associated with the concept of druggability being fairly imprecise. For example, some binding sites may have high affinity for just a few ligands, and low affinity for all other ligands, yielding lower druggability score despite the fact that the site is druggable in principle.

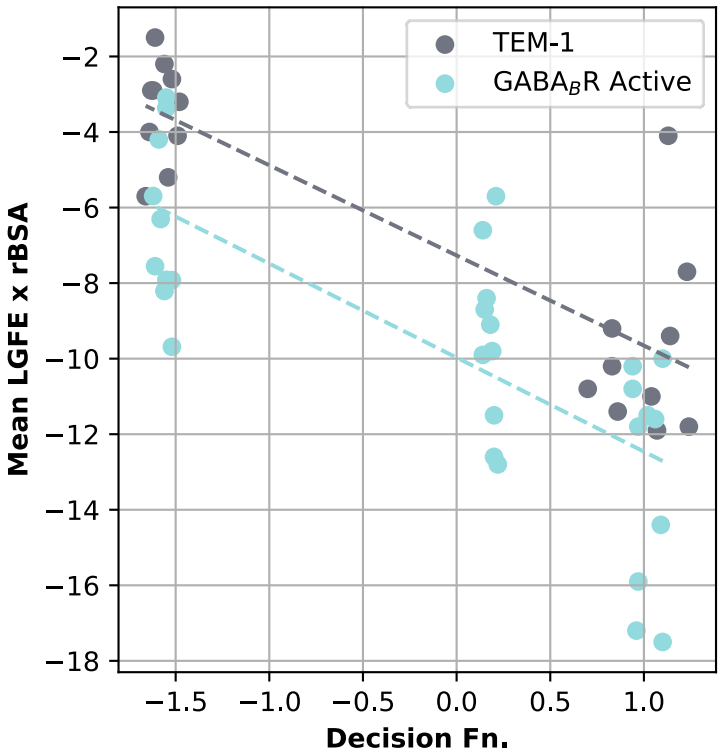


Figure 5: SVM model Decision Function and the Mean LGFE times rBSA for selected Hotspots. For TEM-1 and $GABA_BR$, the Hotspots 1-10 and 91-100 were selected, and for $GABA_BR$ Hotspots 391-400 were also selected. The trendlines show the linear line of best fit.

For TEM-1 Hotspots 1-10 and 91-100 correspond to SVM Decision Function scores of ~1.0 and -1.5, respectively, while Hotspots 1-10, 91-100, and 391-400 correspond to SVM Decision Function scores of ~1.0, 0.2, and -1.5. The discrepancy in the relationship is due to the significantly higher number of Hotspots with $GABA_BR$ versus TEM-1, which biases the overall distribution towards lower ranking SVM Decision Function scores.

Conclusions

We previously presented the SILCS-Hotspots method to leverage the information in SILCS FragMaps to identify a comprehensive set of fragment binding sites. Here we have built upon the previous work and developed a predictive algorithm which identifies the binding sites of larger, drug-like molecules. As a training set, we used the original set of proteins which included a list of Hotspots within 5 Å of a drug-like ligand in a crystal structure of the protein. We first demonstrated that the existing SILCS-Hotspot ranking, based solely on the mean LGFE of each Hotspot that is within 12 Å of at least one other Hotspot, was insufficient to efficiently identify druggable binding sites. Next, use of the Exclusion-map HS SASA of each Hotspot and presence of at least one adjacent Hotspots was shown to substantially improve the ranking. Building on this, a SVM classification model was developed using a wide array of Hotspot and Hotspot cluster properties as features. This led to improved predictions and the final model was validated on a separate set of 9 proteins, on which the model performs quite well. On the problem of identifying at least one Hotspot per ligand binding site, the final model achieves 80% recall in the top 20 Hotspots per protein (20 out of 25 total ligand binding sites total) in the training set, and 88% recall in the top 20 on the validation set (15 out of 17 total sites). By comparing the model's ranking with the predicted affinity and solvent accessibility of members of a chemically-diverse set of FDA-approved compounds, we argue that the model predicts sites which are likely druggable even if they haven't yet been identified through the presence of crystallographic ligands.

In practice, the presented workflow and SVM model offers the capability of identifying novel binding sites for drug-like molecules in proteins, including allosteric sites. This takes advantage of the high information content in the SILCS FragMaps that include contributions from protein flexibility, desolvation and protein-functional group interactions which, in a ligand discovery scenario can be used for database screening and ligand optimization. Notable is the high performance of the SVM model on the validation-set proteins. This is suggested to be due to the use of the physics-based SILCS FragMaps in the initial Hotspots calculation avoiding inherent overtraining effects that may occur with a ML model solely based on data fitting. However, the model has limitations associated with sites adjacent to the lipid bilayer. Future efforts will focus on addressing this issue.

Supporting Information:

Figure S1: Surface-exposed Hotspot 25 in ERK5.

Figure S2: Distribution of Hotspot SASA by protein system.

Figure S3: Class-weighted average of weighted F_1 statistic from Recursive Feature Elimination with 5-fold Cross Validation.

Figure S4: Ranking based on mean LGFE of each Hotspot.

Figure S5: Burial of allosteric binding site between $GABA_BR$ Active TM domains.

Table S1: List of proteins and ligands used for methods validation.
Table S2: Training and validation set Hotspots and ligand distances.
Table S3: Stratified 5-fold Cross-validation training of higher-order SVM Classifier with polynomial or radial basis functions kernels and a Random Forest model.
Table S4. FDA compound screening for selected Hotspots of TEM-1 and $GABA_BR$ Active.

Statements and Declarations

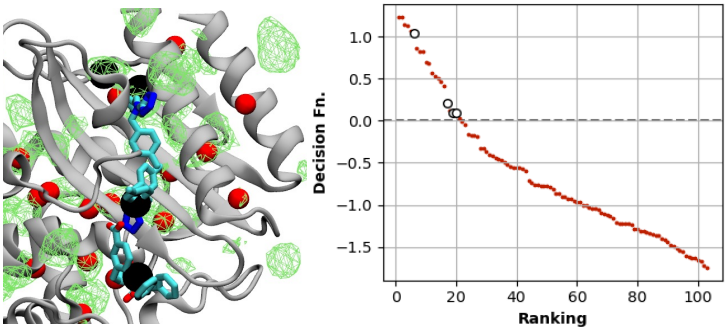
Declaration of Competing Interest

A.D.M. Jr. is co-founder and Chief Scientific Officer of SilcsBio, LLC.

Acknowledgements

The work was funded through National Institutes of Health grant GM131710 to A.D.M. Jr. E.B.N. was supported by the NIH/NCI T32 Training Grant in Cancer Biology T32CA154274 to the University of Maryland, Baltimore. Computational support from the University of Maryland Computer-Aided Drug Design Center is appreciated. The authors acknowledge helpful discussions with Dr. Wenbo Yu.

Table of Contents Figure:



References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000 Jan 1;28(1):235–42.

2. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research. 2022 Jan 7;50(D1):D439–44.

3. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021 Aug;596(7873):590–6.

4. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017 Jan;16(1):19–34.

5. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

6. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021 Aug 20;373(6557):871–6.

7. Pandey M, Fernandez M, Gentile F, Isayev O, Tropsha A, Stern AC, et al. The transformational role of GPU computing and deep learning in drug discovery. Nat Mach Intell. 2022 Mar;4(3):211–21.

8. Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL, et al. Accelerating Molecular Dynamic Simulation on Graphics Processing Units. J Comput Chem. 2009 Apr 30;30(6):864–72.

9. Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. Computational and Structural Biotechnology Journal. 2020 Jan 1;18:417–26.

10. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, et al. Fragment-based identification of druggable “hot spots” of proteins using Fourier domain correlation techniques. Bioinformatics. 2009 Mar 1;25(5):621–7.

11. Ngan CH, Hall DR, Zerbe B, Grove LE, Kozakov D, Vajda S. FTSite: high accuracy detection of ligand binding sites on unbound protein structures. Bioinformatics. 2012 Jan 15;28(2):286–7.

12. Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, et al. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. Nat Protoc. 2015 May;10(5):733–55.

13. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009 Jun 2;10(1):168.

14. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007 Aug 1;23(15):1875–82.

15. Roy A, Zhang Y. Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement. Structure. 2012 Jun 6;20(6):987–97.

16. Roche DB, Tetchner SJ, McGuffin LJ. FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinformatics. 2011 May 16;12(1):160.

17. Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Research. 2010 Jul 1;38(suppl_2):W469–73.

18. Tibaut T, Borišek J, Novič M, Turk D. Comparison of in silico tools for binding site prediction applied for structure-based design of autolysin inhibitors. SAR and QSAR in Environmental Research. 2016 Jul 2;27(7):573–87.

19. Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics. 2013 Oct 15;29(20):2588–95.

20. Huang B. MetaPocket: A Meta Approach to Improve Protein Ligand Binding Site Prediction. OMICS: A Journal of Integrative Biology. 2009 Aug;13(4):325–30.

21. 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 Dec 4;5(12):e1000585.

22. Morrone Xavier M, Sehnem Heck G, Boff de Avila M, Maria Bernhardt Levin N, Oliveira Pintro V, Lemes Carvalho N, et al. SAnDReS a Computational Tool for Statistical Analysis of Docking Results and Development of Scoring Functions. Combinatorial Chemistry & High Throughput Screening. 2016 Dec 1;19(10):801–12.

23. Wu Q, Peng Z, Zhang Y, Yang J. COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Research. 2018 Jul 2;46(W1):W438–42.

24. Vajda S, Beglov D, Wakefield AE, Egbert M, Whitty A. Cryptic binding sites on proteins: definition, detection, and druggability. Curr Opin Chem Biol. 2018 Jun;44:1–8.

25. Schmidtke P, Bidon-Chanal A, Luque FJ, Barril X. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics. 2011 Dec 1;27(23):3276–85.

26. Bowman GR, Geissler PL. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. Proceedings of the National Academy of Sciences. 2012 Jul 17;109(29):11681–6.

27. Bowman GR, Bolin ER, Hart KM, Maguire BC, Marqusee S. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. Proceedings of the National Academy of Sciences. 2015 Mar 3;112(9):2734–9.

28. Cimermancic P, Weinkam P, Rettenmaier TJ, Bichmann L, Keedy DA, Woldeyes RA, et al. CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites. J Mol Biol. 2016 Feb 22;428(4):709–19.

29. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein–ligand docking using GOLD. Proteins: Structure, Function, and Bioinformatics. 2003;52(4):609–23.

30. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. J Med Chem. 2006 Oct 1;49(21):6177–96.

31. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of Computational Chemistry. 2009;30(16):2785–91.

32. 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(2):455–61.

33. Zhang N, Zhao H. Enriching screening libraries with bioactive fragment space. Bioorganic & Medicinal Chemistry Letters. 2016 Aug 1;26(15):3594–7.

34. Seco J, Luque FJ, Barril X. Binding Site Detection and Druggability Index from First Principles. J Med Chem. 2009 Apr 23;52(8):2363–71.

35. Guvench O, MacKerell Jr. AD. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. PLOS Computational Biology. 2009 Jul 10;5(7):e1000435.

36. Congreve M, Chessari G, Tisi D, Woodhead AJ. Recent Developments in Fragment-Based Drug Discovery. J Med Chem. 2008 Jul 1;51(13):3661–80.

37. Kirsch P, Hartman AM, Hirsch AKH, Empting M. Concepts and Core Principles of Fragment-Based Drug Design. Molecules. 2019 Nov 26;24(23):4309.

38. Allen KN, Bellamacina CR, Ding X, Jeffery CJ, Mattos C, Petsko GA, et al. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. J Phys Chem. 1996 Jan 1;100(7):2605–11.

39. Basse N, Kaar JL, Settanni G, Joerger AC, Rutherford TJ, Fersht AR. Toward the rational design of p53-stabilizing drugs: probing the surface of the oncogenic Y220C mutant. Chem Biol. 2010 Jan 29;17(1):46–56.

40. Yang CY, Wang S. Computational Analysis of Protein Hotspots. ACS Med Chem Lett. 2010 Jun 10;1(3):125–9.

41. Tan YS, Śledź P, Lang S, Stubbs CJ, Spring DR, Abell C, et al. Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of polo-like kinase 1. Angew Chem Int Ed Engl. 2012 Oct 1;51(40):10078–81.

42. Huang D, Caflisch A. Small molecule binding to proteins: affinity and binding/unbinding dynamics from atomistic simulations. ChemMedChem. 2011 Sep 5;6(9):1578–80.

43. Bakan A, Nevins N, Lakdawala AS, Bahar I. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. J Chem Theory Comput. 2012 Jul 10;8(7):2435–47.

44. Ghanakota P, Carlson HA. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. J Med Chem. 2016 Dec 8;59(23):10383–99.

45. Alvarez-Garcia D, Barril X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. J Med Chem. 2014 Oct 23;57(20):8530–9.

46. Prakash P, Sayyed-Ahmad A, Gorfe AA. pMD-Membrane: A Method for Ligand Binding Site Identification in Membrane-Bound Proteins. PLOS Computational Biology. 2015 Oct 27;11(10):e1004469.

47. Sayyed-Ahmad A, Gorfe AA. Mixed-Probe Simulation and Probe-Derived Surface Topography Map Analysis for Ligand Binding Site Identification. J Chem Theory Comput. 2017 Apr 11;13(4):1851–61.

48. Ghanakota P, Carlson HA. Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. J Phys Chem B. 2016 Aug 25;120(33):8685–95.

49. Graham SE, Leja N, Carlson HA. MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations. J Chem Inf Model. 2018 Jul 23;58(7):1426–33.

50. Smith RD, Carlson HA. Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics. J Chem Inf Model. 2021 Mar 22;61(3):1287–99.

51. Comitani F, Gervasio FL. Exploring Cryptic Pockets Formation in Targets of Pharmaceutical Interest with SWISH. J Chem Theory Comput. 2018 Jun 12;14(6):3321–31.

52. Borsatto A, Gianquinto E, Rizzi V, Gervasio FL. SWISH-X, an Expanded Approach to Detect Cryptic Pockets in Proteins and at Protein–Protein Interfaces. J Chem Theory Comput [Internet]. 2024 Apr 2 [cited 2024 Apr 9]; Available from: https://doi.org/10.1021/acs.jctc.3c01318

53. Sabanés Zariquiey F, de Souza JV, Bronowska AK. Cosolvent Analysis Toolkit (CAT): a robust hotspot identification platform for cosolvent simulations of proteins to expand the druggable proteome. Sci Rep. 2019 Dec 13;9(1):19118.

711 54. Raman EP, Yu W, Guvench O, MacKerell ADJr. Reproducing Crystal Binding Modes of Ligand
712 Functional Groups Using Site-Identification by Ligand Competitive Saturation (SILCS)
713 Simulations. J Chem Inf Model. 2011 Apr 25;51(4):877–96.

714 55. Raman EP, Yu W, Lakkaraju SK, MacKerell ADJr. Inclusion of Multiple Fragment Types in the
715 Site Identification by Ligand Competitive Saturation (SILCS) Approach. J Chem Inf Model.
716 2013 Dec 23;53(12):3384–98.

717 56. Andreev G, Kovalenko M, Bozdaganyan ME, Orekhov PS. Colabind: A Cloud-Based Approach
718 for Prediction of Binding Sites Using Coarse-Grained Simulations with Molecular Probes. J
719 Phys Chem B. 2024 Apr 4;128(13):3211–9.

720 57. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High
721 performance molecular simulations through multi-level parallelism from laptops to
722 supercomputers. SoftwareX. 2015 Sep 1;1–2:19–25.

723 58. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient,
724 Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput. 2008 Mar
725 1;4(3):435–47.

726 59. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine Microsecond
727 Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J Chem
728 Theory Comput. 2012 May 8;8(5):1542–55.

729 60. Eastman P, Friedrichs MS, Chodera JD, Radmer RJ, Bruns CM, Ku JP, et al. OpenMM 4: A
730 Reusable, Extensible, Hardware Independent Library for High Performance Molecular
731 Simulation. J Chem Theory Comput. 2013 Jan 8;9(1):461–9.

732 61. Best RB, Hummer G. Optimized Molecular Dynamics Force Fields Applied to the Helix−Coil
733 Transition of Polypeptides. J Phys Chem B. 2009 Jul 2;113(26):9004–15.

734 62. Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, et al. Optimization of the Additive
735 CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone φ,
736 ψ and Side-Chain χ1 and χ2 Dihedral Angles. J Chem Theory Comput. 2012 Sep
737 11;8(9):3257–73.

738 63. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: An
739 Improved Force Field for Folded and Intrinsically Disordered Proteins. Nat Methods. 2017
740 Jan;14(1):71–3.

741 64. Robustelli P, Piana S, Shaw DE. Developing a molecular dynamics force field for both folded
742 and disordered protein states. Proceedings of the National Academy of Sciences. 2018
743 May 22;115(21):E4758–66.

65. Tian C, Kasavajhala K, Belfon KAA, Raguette L, Huang H, Migues AN, et al. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. J Chem Theory Comput. 2020 Jan 14;16(1):528–52.

66. Lakkaraju SK, Raman EP, Yu W, MacKerell AD. Sampling of Organic Solutes in Aqueous and Heterogeneous Environments Using Oscillating Excess Chemical Potentials in Grand Canonical-like Monte Carlo-Molecular Dynamics Simulations. J Chem Theory Comput. 2014 Jun 10;10(6):2281–90.

67. Zhao M, Kognole AA, Jo S, Tao A, Hazel A, MacKerell Jr AD. GPU-specific algorithms for improved solute sampling in grand canonical Monte Carlo simulations. Journal of Computational Chemistry. 2023;44(20):1719–32.

68. Ustach VD, Lakkaraju SK, Jo S, Yu W, Jiang W, MacKerell AD. Optimization and Evaluation of Site-Identification by Ligand Competitive Saturation (SILCS) as a Tool for Target-Based Ligand Optimization. J Chem Inf Model. 2019 Jun 24;59(6):3018–35.

69. Goel H, Hazel A, Ustach VD, Jo S, Yu W, MacKerell AD. Rapid and accurate estimation of protein–ligand relative binding affinities using site-identification by ligand competitive saturation. Chem Sci. 2021 Jul 1;12(25):8844–58.

70. MacKerell AD, Jo S, Lakkaraju SK, Lind C, Yu W. Identification and characterization of fragment binding sites for allosteric ligand design using the site identification by ligand competitive saturation hotspots approach (SILCS-Hotspots). Biochim Biophys Acta Gen Subj. 2020 Apr;1864(4):129519.

71. Kognole AA, Hazel A, MacKerell AD. SILCS-RNA: Toward a Structure-Based Drug Design Approach for Targeting RNAs with Small Molecules. J Chem Theory Comput. 2022 Sep 13;18(9):5672–91.

72. Weisel M, Proschak E, Kriegl JM, Schneider G. Form follows function: Shape analysis of protein cavities for receptor-based drug design. PROTEOMICS. 2009;9(2):451–9.

73. Liang J, Woodward C, Edelsbrunner H. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Protein Science. 1998;7(9):1884–97.

74. Johnson DK, Karanicolas J. Druggable Protein Interaction Sites Are More Predisposed to Surface Pocket Formation than the Rest of the Protein Surface. PLOS Computational Biology. 2013 Mar 7;9(3):e1002951.

75. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Research. 2012 Jan;40(D1):D370–6.

76. Lomize AL, Todd SC, Pogozheva ID. Spatial arrangement of proteins in planar and curved membranes by PPM 3.0. Protein Sci. 2022 Jan;31(1):209–20.

77. Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: A web-based graphical user interface for CHARMM. Journal of Computational Chemistry. 2008;29(11):1859–65.

78. Wu EL, Cheng X, Jo S, Rui H, Song KC, Dávila-Contreras EM, et al. CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. Journal of Computational Chemistry. 2014;35(27):1997–2004.

79. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K_a Predictions. J Chem Theory Comput. 2011 Feb 8;7(2):525–37.

80. SILCS: Site Identification by Ligand Competitive Saturation — SilcsBio User Guide [Internet]. [cited 2024 Feb 21]. Available from: https://docs.silcsbio.com/2023/silcs/silcs.html

81. Taylor RD, MacCoss M, Lawson ADG. Rings in Drugs. J Med Chem. 2014 Jul 24;57(14):5845–59.

82. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for “omics” research on drugs. Nucleic Acids Res. 2011 Jan;39(Database issue):D1035-1041.

83. Research C for DE and. Drugs@FDA Data Files. FDA [Internet]. 2024 Mar 19 [cited 2024 Mar 21]; Available from: https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files

84. RDKit: Open-source cheminformatics. [Internet]. Available from: https://www.rdkit.org

85. Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations [Internet]. F1000Research; 2016 [cited 2024 Feb 26]. Available from: https://f1000research.com/articles/5-189

86. Lam SK, Pitrou A, Seibert S. Numba: a LLVM-based Python JIT compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC [Internet]. New York, NY, USA: Association for Computing Machinery; 2015 [cited 2024 Feb 26]. p. 1–6. (LLVM ’15). Available from: https://dl.acm.org/doi/10.1145/2833157.2833162

87. Baumli S, Endicott JA, Johnson LN. Halogen Bonds Form the Basis for Selective P-TEFb Inhibition by DRB. Chemistry & Biology. 2010 Sep 24;17(9):931–6.

88. Wu SY, McNae I, Kontopidis G, McClue SJ, McInnes C, Stewart KJ, et al. Discovery of a Novel Family of CDK Inhibitors with the Program LIDAEUS: Structural Basis for Ligand-Induced Disordering of the Activation Loop. Structure. 2003 Apr 1;11(4):399–410.

811 89. Glatz G, Gógl G, Alexa A, Reményi A. Structural Mechanism for the Specific Assembly and
812 Activation of the Extracellular Signal Regulated Kinase 5 (ERK5) Module*. Journal of
813 Biological Chemistry. 2013 Mar 22;288(12):8596–609.

814 90. Wiesmann C, Barr KJ, Kung J, Zhu J, Erlanson DA, Shen W, et al. Allosteric inhibition of
815 protein tyrosine phosphatase 1B. Nat Struct Mol Biol. 2004 Aug;11(8):730–7.

816 91. Han Y, Belley M, Bayly CI, Colucci J, Dufresne C, Giroux A, et al. Discovery of [(3-bromo-7-
817 cyano-2-naphthyl)(difluoro)methyl]phosphonic acid, a potent and orally active small
818 molecule PTP1B inhibitor. Bioorganic & Medicinal Chemistry Letters. 2008 Jun
819 1;18(11):3200–5.

820 92. Montalibet J, Skorey K, McKay D, Scapin G, Asante-Appiah E, Kennedy BP. Residues Distant
821 from the Active Site Influence Protein-tyrosine Phosphatase 1B Inhibitor Binding*. Journal
822 of Biological Chemistry. 2006 Feb 24;281(8):5258–66.

823 93. Wan ZK, Follows B, Kirincich S, Wilson D, Binnun E, Xu W, et al. Probing acid replacements of
824 thiophene PTP1B inhibitors. Bioorganic & Medicinal Chemistry Letters. 2007 May
825 15;17(10):2913–20.

826 94. Pereira de Jésus-Tran K, Côté PL, Cantin L, Blanchet J, Labrie F, Breton R. Comparison of
827 crystal structures of human androgen receptor ligand-binding domain complexed with
828 various agonists reveals molecular determinants responsible for binding affinity. Protein
829 Science. 2006;15(5):987–99.

830 95. Estébanez-Perpiñá E, Arnold LA, Nguyen P, Rodrigues ED, Mar E, Bateman R, et al. A surface
831 on the androgen receptor that allosterically regulates coactivator binding. Proceedings of
832 the National Academy of Sciences. 2007 Oct 9;104(41):16074–9.

833 96. Srivastava A, Yano J, Hirozane Y, Kefala G, Gruswitz F, Snell G, et al. High-resolution structure
834 of the human GPR40 receptor bound to allosteric agonist TAK-875. Nature. 2014
835 Sep;513(7516):124–7.

836 97. Ho JD, Chau B, Rodgers L, Lu F, Wilbur KL, Otto KA, et al. Structural basis for GPR40 allosteric
837 agonism and incretin stimulation. Nat Commun. 2018 Apr 25;9(1):1645.

838 98. Haga K, Kruse AC, Asada H, Yurugi-Kobayashi T, Shiroishi M, Zhang C, et al. Structure of the
839 human M2 muscarinic acetylcholine receptor bound to an antagonist. Nature. 2012
840 Feb;482(7386):547–51.

841 99. Kruse AC, Ring AM, Manglik A, Hu J, Hu K, Eitel K, et al. Activation and allosteric modulation
842 of a muscarinic acetylcholine receptor. Nature. 2013 Dec;504(7478):101–6.

843 100. Rasmussen SGF, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, et al. Crystal structure
844 of the β2 adrenergic receptor–Gs protein complex. Nature. 2011 Sep 29;477(7366):549–
845 55.

101. Liu X, Ahn S, Kahsai AW, Meng KC, Latorraca NR, Pani B, et al. Mechanism of intracellular allosteric β2AR antagonist revealed by X-ray crystal structure. Nature. 2017 Aug;548(7668):480–4.

102. Goldstein DM, Soth M, Gabriel T, Dewdney N, Kuglstatter A, Arzeno H, et al. Discovery of 6-(2,4-Difluorophenoxy)-2-[3-hydroxy-1-(2-hydroxyethyl)propylamino]-8-methyl-8H-pyrido[2,3-d]pyrimidin-7-one (Pamapimod) and 6-(2,4-Difluorophenoxy)-8-methyl-2-(tetrahydro-2H-pyran-4-ylamino)pyrido[2,3-d]pyrimidin-7(8H)-one (R1487) as Orally Bioavailable and Highly Selective Inhibitors of p38α Mitogen-Activated Protein Kinase. J Med Chem. 2011 Apr 14;54(7):2255–65.

103. Pargellis C, Tong L, Churchill L, Cirillo PF, Gilmore T, Graham AG, et al. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. Nat Struct Mol Biol. 2002 Apr;9(4):268–72.

104. Drug Design Data Resource (D3R). Drug Design Data Resource Grand Challenge 2 Dataset: FXR - Farnesoid X receptor [Internet]. Drug Design Data Resource (D3R); 2017 [cited 2024 Feb 19]. p. 71.5MB. Available from: https://drugdesigndata.org/about/datasets/882

105. Cumming JN, Smith EM, Wang L, Misiaszek J, Durkin J, Pan J, et al. Structure based design of iminohydantoin BACE1 inhibitors: Identification of an orally available, centrally active BACE1 inhibitor. Bioorganic & Medicinal Chemistry Letters. 2012 Apr 1;22(7):2444–9.

106. D3R | Drug Design Data Resource Grand Challenge 4 Dataset: BACE1 [Internet]. [cited 2024 Feb 19]. Available from: https://drugdesigndata.org/about/datasets/2027

107. D3R | Drug Design Data Resource Grand Challenge Dataset: GSK TrmD [Internet]. [cited 2024 Feb 19]. Available from: https://drugdesigndata.org/about/datasets/226

108. Friberg A, Vigil D, Zhao B, Daniels RN, Burke JP, Garcia-Barrantes PM, et al. Discovery of Potent Myeloid Cell Leukemia 1 (Mcl-1) Inhibitors Using Fragment-Based Methods and Structure-Based Design. J Med Chem. 2013 Jan 10;56(1):15–30.

109. Sato M, Arakawa T, Nam YW, Nishimoto M, Kitaoka M, Fushinobu S. Open–close structural change upon ligand binding and two magnesium ions required for the catalysis of *N*-acetylhexosamine 1-kinase. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2015 May 1;1854(5):333–40.

110. Baum B, Muley L, Smolinski M, Heine A, Hangauer D, Klebe G. Non-additivity of Functional Group Contributions in Protein–Ligand Binding: A Comprehensive Study by Crystallography and Isothermal Titration Calorimetry. Journal of Molecular Biology. 2010 Apr 9;397(4):1042–54.

111. Tarver CL. Molecular role of angiopoietin-like 4's carboxy-terminal domain in pancreatic ductal adenocarcinoma progression [Dissertations]. University of Huntsville Alabama; 2019.

112. Wang X, Minasov G, Shoichet BK. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. Journal of Molecular Biology. 2002 Jun 28;320(1):85–95.

113. Horn JR, Shoichet BK. Allosteric Inhibition Through Core Disruption. Journal of Molecular Biology. 2004 Mar 5;336(5):1283–91.

114. Ness S, Martin R, Kindler AM, Paetzel M, Gold M, Jensen SE, et al. Structure-Based Design Guides the Improved Efficacy of Deacylation Transition State Analogue Inhibitors of TEM-1 β-Lactamase. Biochemistry. 2000 May 1;39(18):5312–21.

115. Kim Y, Jeong E, Jeong JH, Kim Y, Cho Y. Structural Basis for Activation of the Heterodimeric $GABA_B$ Receptor. Journal of Molecular Biology. 2020 Nov 6;432(22):5966–84.

116. Shaye H, Ishchenko A, Lam JH, Han GW, Xue L, Rondard P, et al. Structural basis of the activation of a metabotropic GABA receptor. Nature. 2020 Aug;584(7820):298–303.

117. Mao C, Shen C, Li C, Shen DD, Xu C, Zhang S, et al. Cryo-EM structures of inactive and active $GABA_B$ receptor. Cell Res. 2020 Jul;30(7):564–73.

118. D3R | Drug Design Data Resource [Internet]. [cited 2024 Feb 19]. Available from: https://drugdesigndata.org/

119. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON.

120. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002 Jan 1;46(1):389–422.

121. sklearn documentation for SVC [Internet]. [cited 2024 Apr 1]. Available from: https://scikit-learn/stable/modules/generated/sklearn.svm.SVC.html

122. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning [Internet]. 2nd ed. Springer New York, NY; 2009. 745 p. Available from: https://doi.org/10.1007/978-0-387-84858-7

123. The pandas development team. pandas-dev/pandas: Pandas [Internet]. Zenodo; 2023. Available from: https://zenodo.org/record/7741580

124. Humphrey W, Dalke A, Schulten K. VMD – Visual Molecular Dynamics. Journal of Molecular Graphics. 1996;14:33–8.

125. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007;9(3):90–5.

913 126. Petroff MA. Accessible Color Sequences for Data Visualization [Internet]. arXiv; 2024
914 [cited 2024 Mar 21]. Available from: http://arxiv.org/abs/2107.02270

915

Supporting Information

Combined physics- and machine-learning-based method to identify druggable binding sites using SILCS-Hotspots

Erik B. Nordquist,[1,#] Mingtian Zhao,[1,#] Anmol Kumar,[1] Alexander D. MacKerell, Jr.[1*]

[1]Computer Aided Drug Design Center, Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Baltimore, Maryland 21201, United States.

[#]These authors contributed equally to the work.
*Corresponding author: A.D.M. Jr., alex@outerbanks.umaryland.edu

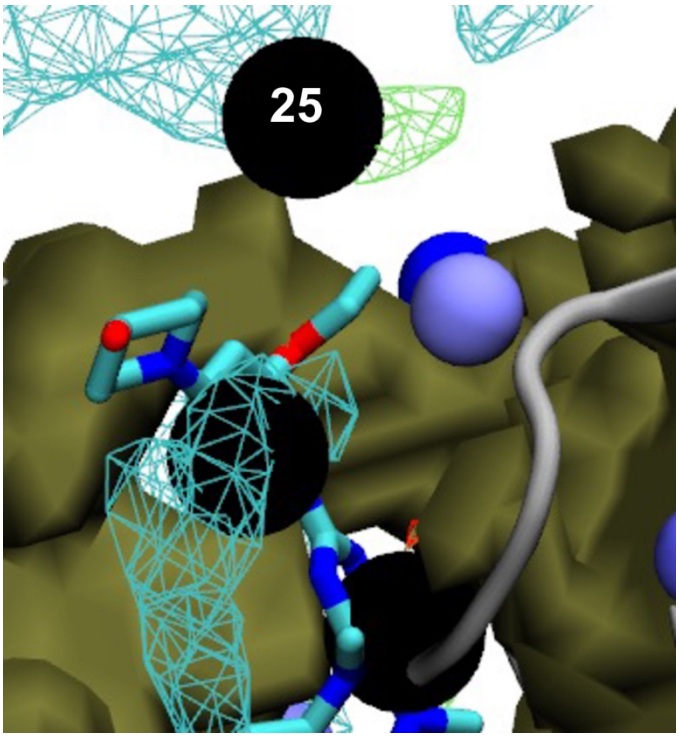


Figure S1: Surface-exposed Hotspot 25 in ERK5. The Hotspots are shown as spheres overlaid on the SILCS exclusion map (tan surface). The Hotspots within 5 Å of ligand 4WG (PDB 5BYY) are black, else the Hotspots are colored by the final model's decision function, with red corresponding to the highest and blue the lowest confidence of being a druggable site. Hotspot 25 (original LGFE-based ranking) is located above and outside of the ligand binding pocket and has a large SASA with respect to the Exclusion map.

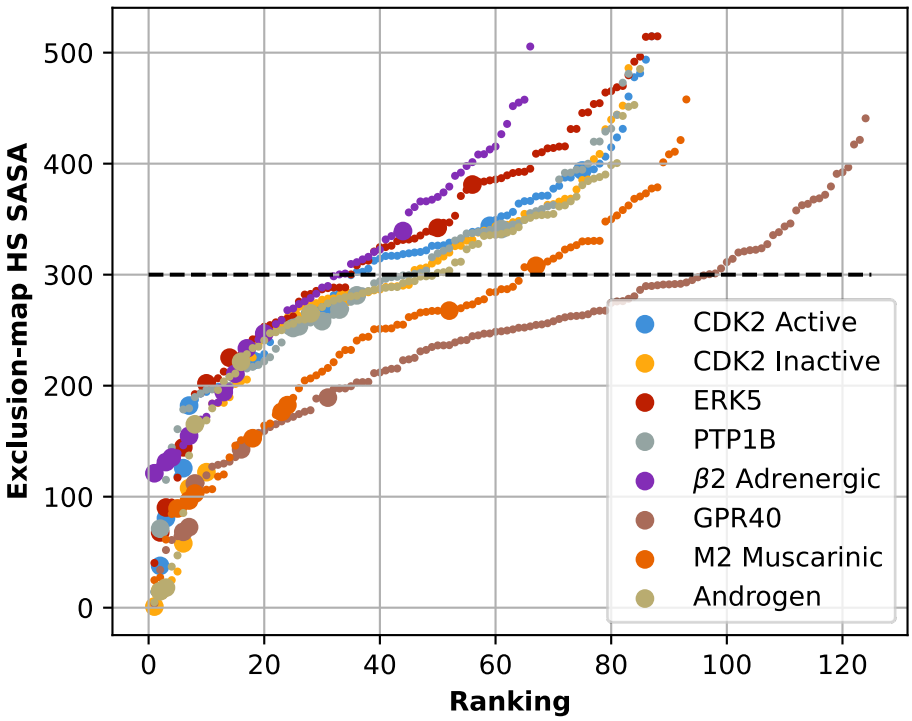


Figure S2. Distribution of Hotspot SASA by protein system. The SASA (Å^2) was calculated with respect to the SILCS Exclusion map for Hotspots of radius 5 Å. The large circles are Hotspots within 5 Å of a crystal ligand's non-hydrogen atoms. The dashed black line indicates the empirical cutoff at 300 Å^2.

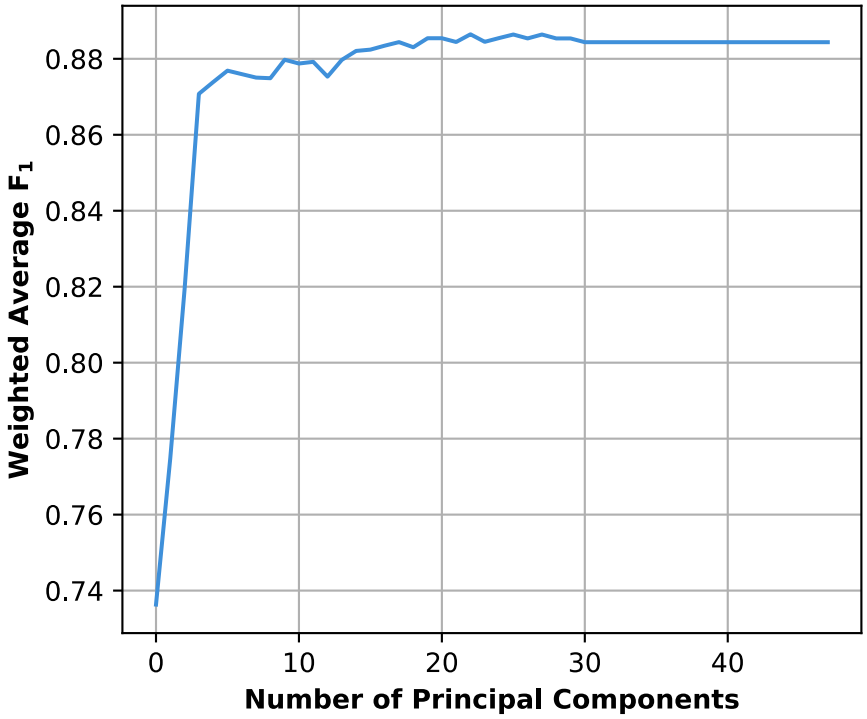


Figure S3. Class-weighted average of weighted F_1 statistic from Recursive Feature Elimination with 5-fold Cross Validation. The weighted F_1 shows the model's performance while including some number of principal components, and the maximum occurs at 22.

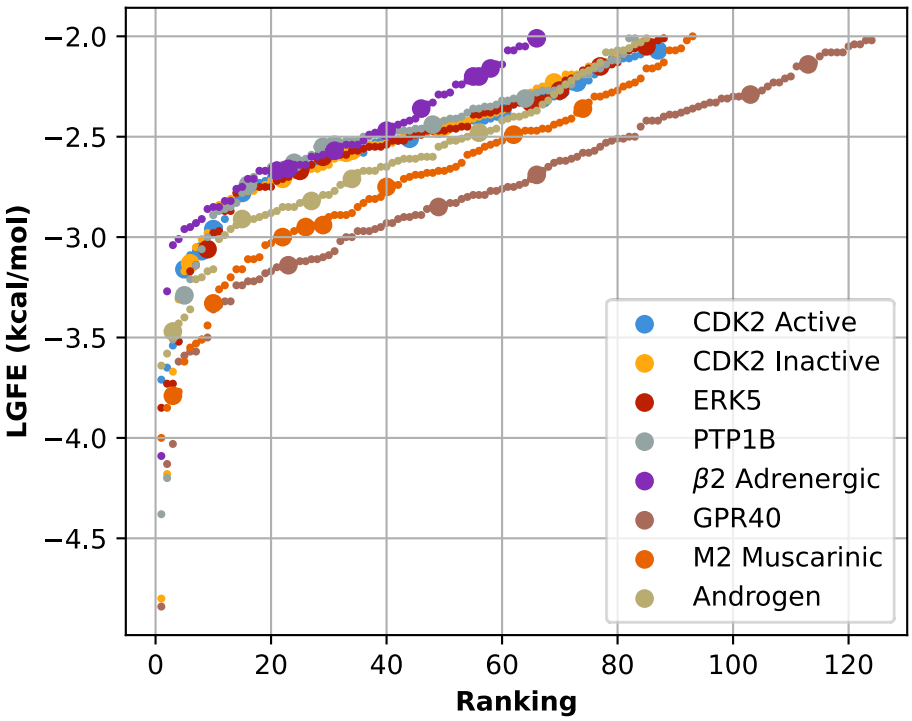


Figure S4: Ranking based on mean LGFE of each Hotspot. This is the mean LGFE of all the fragments clustered within the Hotspot and was the original ranking metric.

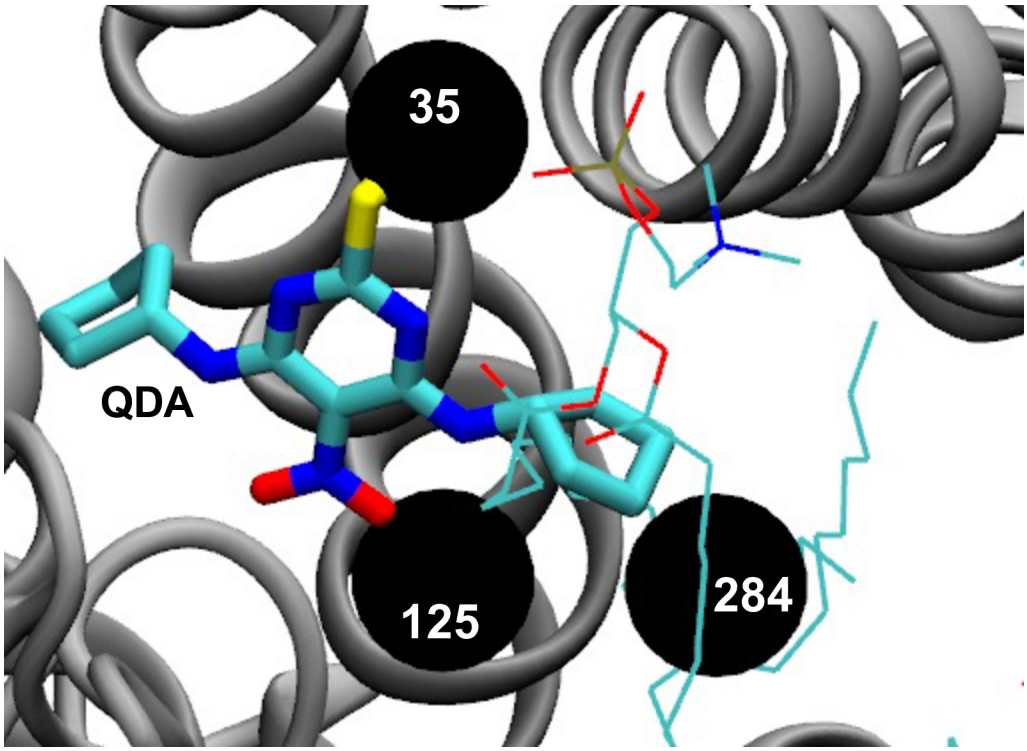


Figure S5: Burial of allosteric binding site between $GABA_BR$ Active TM domains. The Hotspots within 5 Å of one of the ligand non-hydrogen atoms and near the lipid are shown as black spheres. The allosteric ligand QDA is drawn in Licorice style; additional information on QDA is in Table S1. The lipids near to the TM helices are rendered with Lines style. The teal atoms are carbon, the yellow are sulfur, the red are oxygen, the blue are nitrogen, and the pink are fluorine. The protein and lipids are taken from a representative snapshot from the SILCS MD simulations.

Table S1: List of proteins and ligands used for model training and validation. The protein structures used for the SILCS simulations are bolded. Where possible, an apo structure is used for the SILCS simulations. The alignments were done on all backbone non-hydrogen atoms with the residues listed. Where alignment residues are not listed, they are identical to the residues listed for the reference protein (used for SILCS simulations). a) Alignment described in [1]. b) Alignment described in [2]. c) Structures aligned in D3R dataset [3]. d) ASP233 protonated as predicted by PropKa [4]. Some of the data in this table is reproduced from refs [1,2,5]. O stands for Orthosteric, A stands for Allosteric.

Name	PDB/D3R (SILCS)	Alignment residues	RMSD (Å)	Ligand	Notes (Orthosteric/Allosteric), Reference
CDK2 active	**3MY5**			DRB	O, active [6]
CDK2 inactive	**1PW2**				Apo, inactive [7]
	3PXF	1-298	4.6	2AN	A, 2 present [8]
	5FP5	1-298	4.5	1Y6	A, 2 present [9]
	5FP6	1-298	4.0	MFZ	A, 2 present [9]
ERK5	**4IC8**				Apo, inactive [10]
	5BYY	49-389 (27-367)	3.8	4WG	O [11]
	4ZSG	47-389	3.5	4QX	A [11]
PTP1B	**2F6F**				Apo, S295F mutant [12]
	1T48	1-298	2.8	BB3	A [13]
	2NT7	2-298	1.5	9O2	O [14]
	3CWE	1-283	1.1	825	O, phosphonic acid analog [15]
β2 Adrenergic	**3SN6**	6-205, 238-315		P0G	O, active [16]
	5X7D	31-230, 263-340	2.8	CAU	O, carazolol, inactive [17]
	5X7D			8VS	A, inactive [17]
GPR40 (partial agonist)	**4PHU**	15-176, 178-224, 390-453		2YB	A, TAK-875, Site 1, Partial allosteric agonist [18]
GPR40 (full agonist)	**5KW2**	3-31, 39-111, 120-145, 168-210, 214-277		6XQ	A, Lilly, full positive allosteric agonist, Site 2 [19]
	5TZY	2-163, 165-211, 2215-2278	2.0	MK6	A, partial positive allosteric agonist, Site 1. This ligand was compared to the 5KW2 map.
	5TZY	3-31, 39-111, 120-145, 168-210, 2214-2277	1.6	7OS	A, AgoPAM, Full allosteric agonist, Site 2. This ligand was compared to 4PHU map.

M2 Muscarinic	**3UON**	1-198, 201-277		QNB	O, antagonist, inactive form [20]
	4MQT	20-217, 379-455	2.6	2CU	A, LY2119620, positive allosteric modulator, active [21]
	4MQT	20-217, 379-455	2.6	IXO	O, iperoxo, agonist, active [21]
Androgen	**2AM9**			TES	O, testosterone [22]
	2PIX	670-918	0.5	DHT	O, dihydrotestosterone [23]
	2PIX	670-918	0.5	FLA	A, flufenamic acid (inhibitor) [23]
P38	**3FLY**			FLY	A [24]
	1KV2	7-150, 200-300	0.9	B96	A, BIRB 796, partial allosteric [25]
BACE	**4DJW**			0KP	A [26]
	--	--	--	Merck 36-17	Merck 36-17b [27], pose from SILCS-MC docking [2] (no PDB entry)
Hsp90	**2JJC**			LGA	O, Pyrimidin-2-amine (small fragment) [28]
	4YKW	20-220 b	2.1	4ES	O, CS312
TrmD	**4YPW**			4FD	No ref
	4YQ2	1-160 b	0.5	EFY	No ref
Thrombin	**2ZFF**				O [29]
	2ZDA	--	0.5	32U	O, S1 pocket, structure alignment used [29]
MCL1	**4HW3**			19G	O [30]
	4HW2	b	0.6	19H	O [31]
FXR	**1DVWB**				Apo [3]
	1NQQW	250-458c	1.5	--	Roche 016 [3]
	1FGGU	250-458c	1.6	--	Roche 034 [3]
	1WGPH	250-458c	2.4	--	Roche 036 [3]
ANGPTL4	**6U0A**				O, Glycerol [32]
	6U1U	185-400	0.4	PLM	O, Palmitic acid [32]
TEM1	**1JWP**[d]				Apo, M182T mutant [33]
	1ERO	26-181	0.3	BJP	O [34]
	1PZO	26-181	0.3	CBT	A, 2 ligands present [35]
$GABA_BR$ active	**7CA3**			FN0	A, BHFF, active [36]
$GABA_BR$ inactive	**7CA5**				Apo, inactive [36]
	6UO8	Chain A: 165-485, 503-562, Chain B: 380-569	1.9 (7CA3)	QDA	A, GS39783 [37]
	7C7Q	Chain A: 707-861, Chain B:	1.6 (7CA3)	2C0	O, Baclofen [38], GABA analog

7C7Q	600-748 (7CA3) Chain A: 290-480, 495-570 (7CA5)	3.4 (7CA5)	2C0	O, Baclofen [38], GABA analog

Table S2: Training and validation set Hotspots and ligand distances. Distance is the distance to the nearest non-hydrogen atom on that ligand. Ligand names are given in Table S1. Rank refers to the SVM model rank and the original Hotspot LGFE rank is given for comparison. For the validation set, we included only one Hotspot per ligand to avoid over-counting in the test dataset due to it being smaller, although some Hotspots were within 5 Å of multiple ligands. There are some ligands which appear multiple times, as noted in Table S1, which is denoted with a, b.

Training					
Protein	**Distance (Å)**	**Ligand**	**Rank**	**Decision Fn.**	**LGFE Rank**
CDK2 Active	2.7	2AN a	34	-0.49	12
CDK2 Active	--	2AN b	--	--	--
CDK2 Active	4.9	1Y6 a	10	0.23	19
CDK2 Active	1.4	1Y6 a	2	1.34	5
CDK2 Active	4.2	1Y6 a	3	1.25	92
CDK2 Active	2.7	1Y6 b	4	1.24	65
CDK2 Active	1.1	1Y6 b	7	0.27	78
CDK2 Active	4.9	MFZ a	44	-0.78	8
CDK2 Active	2.7	MFZ a	34	-0.49	12
CDK2 Active	1.7	MFZ b	2	1.34	5
CDK2 Active	3	MFZ b	3	1.25	92
CDK2 Inactive	3.5	2AN a	6	0.79	22
CDK2 Inactive	1.5	2AN a	2	1.79	33
CDK2 Inactive	0.9	2AN a	1	2.00	70
CDK2 Inactive	4.2	2AN b	31	-0.51	6
CDK2 Inactive	2	2AN b	6	0.79	22
CDK2 Inactive	3.2	1Y6 a	7	0.57	50
CDK2 Inactive	2.4	1Y6 a	37	-0.64	51
CDK2 Inactive	2.5	1Y6 b	5	1.01	34
CDK2 Inactive	4.4	1Y6 b	1	2.00	70
CDK2 Inactive	1.7	MFZ a	31	-0.51	6
CDK2 Inactive	4.2	MFZ a	6	0.79	22
CDK2 Inactive	1.9	MFZ b	5	1.01	34
ERK5	0.9	4QX	2	1.16	29
ERK5	3.7	4QX	1	1.19	71
ERK5	0.9	4WG	10	0.22	8
ERK5	2.5	4WG	1	1.19	71
ERK5	2.6	4WG	5	0.83	78
ERK5	0.8	4WG	3	1.02	86
PTP1B	1	BB3	2	1.15	5
PTP1B	3	BB3	40	-0.58	16
PTP1B	1.6	BB3	3	0.39	50
PTP1B	0.6	902	19	-0.10	24
PTP1B	2.3	902	52	-0.86	30
PTP1B	1.9	902	32	-0.47	32
PTP1B	1.3	825	19	-0.10	24
PTP1B	4.8	825	32	-0.47	32
β2 Adrenergic	1.7	CAU	3	1.11	20
β2 Adrenergic	1.5	CAU	13	0.37	31
β2 Adrenergic	1.3	CAU	2	1.18	41

β2 Adrenergic	3.5	CAU	9	0.64	47
β2 Adrenergic	2	CAU	10	0.63	57
β2 Adrenergic	3.3	CAU	5	0.85	67
β2 Adrenergic	1.2	8VS	7	0.77	58
β2 Adrenergic	2.1	8VS	8	0.72	60
GPR40 (5KW2)	1.2	MK6	12	-0.05	27
GPR40 (5KW2)	1.1	MK6	1	1.81	57
GPR40 (5KW2)	0.9	MK6	2	1.29	61
GPR40 (4PHU)	1.5	70S	10	0.03	12
GPR40 (4PHU)	2.2	70S	5	0.38	29
M2 Muscarinic	2.3	2CU	10	0.72	23
M2 Muscarinic	3.2	2CU	25	-0.21	28
M2 Muscarinic	2	2CU	19	0.14	41
M2 Muscarinic	3.2	2CU	5	1.00	75
M2 Muscarinic	2.1	IXO	8	0.92	3
M2 Muscarinic	0.8	IXO	3	1.20	10
M2 Muscarinic	3.8	IXO	2	1.36	27
Androgen	0.8	DHT	3	1.41	26
Androgen	3.6	DHT	1	1.57	57
Androgen	0.9	FLA	9	0.19	3
Androgen	1.2	FLA	15	-0.13	17
Androgen	1.6	FLA	27	-0.49	34
		Validation			
P38	1.2	B96	1	2.34	19
P38	1.7	B96	4	1.62	3
P38	3.1	B96	9	0.70	50
BACE1	3.1	Merck36 17b	30	-0.04	4
BACE1	0.8	Merck36 17b	12	0.67	14
BACE1	1.0	Merck36 17b	10	0.72	21
BACE1	4.3	Merck36 17b	24	0.05	29
Hsp90	1.1	4ES	4	1.04	2
Hsp90	1.3	4ES	8	0.69	58
TrmD	1.0	EFY	4	0.85	1
TrmD	1.9	EFY	5	0.85	23
TrmD	1.1	EFY	8	0.24	11
Thrombin	1.0	32U	27	-0.17	2
Thrombin	1.5	32U	11	0.44	5
MCL1	1.9	ADP	13	0.01	7
MCL1	0.6	ADP	9	0.24	10
FXR	0.5	1wpgh	3	1.43	3
FXR	4.1	1nqqw	5	1.12	10
FXR	1.7	1wpgh	4	1.39	11
FXR	3.1	1wgph	11	0.55	20
FXR	1.5	1nqqw	10	0.63	25
FXR	1.0	1fggu	8	0.67	32
FXR	4.1	1wpgh	1	1.82	58
FXR	5.0	1wpgh	7	0.84	62
FXR	1.6	1fggu	9	0.64	72
ANGPTL4	0.9	PLM	3	0.84	1

ANGPTL4	1.0	PLM	9	0.68	6
TEM1	1.6	BJP	20	0.09	95
TEM1	1.6	CBT a	19	0.10	71
TEM1	3.2	CBT b	6	1.04	16
TEM1	3.7	CBT b	17	0.21	10
$GABA_BR$ Active	4.9	QDA	177	-0.48	177
$GABA_BR$ Active	0.5	QDA	162	-0.40	35
$GABA_BR$ Active	1.9	QDA	141	-0.23	125
$GABA_BR$ Active	3.0	QDA	179	-0.49	284
$GABA_BR$ Active	3.0	2C0	67	0.45	21
$GABA_BR$ Active	3.1	2C0	19	0.84	12
$GABA_BR$ Active	4.8	2C0	103	0.11	97
$GABA_BR$ Inactive	3.3	2C0	133	-0.27	47
$GABA_BR$ Inactive	4.9	2C0	106	-0.09	39
$GABA_BR$ Inactive	4.0	2C0	122	-0.21	186

Table S3: Stratified 5-fold cross-validation training of higher-order SVM Classifier with polynomial or radial basis functions kernels and a Random Forest model. These models were all trained with class_weight = 'balanced', max_iter = 1e6, and tol = 1e-4. The reported metrics are mean ± sem over the 5-fold CV. Weighted F_1, precision, and recall are defined based on the Hotspots near crystal ligands as described in the Methods section. Precision is the ratio of predicted hits to total Hotspots above some cutoff, and recall is the ratio of predicted hits to the total true hits. Weighted F_1 is the population-weighted harmonic mean of precision and recall. Single-fit recall is the recall after training on the whole dataset. The RF model was optimized over the following hyperparameter space, with the selected values bolded: n_estimators = [**10**, 50, 100], max_depth = [2, **10**, 50, 100], min_samples_split = [**2**, 10, 50, 100], min_samples_leaf = [2, 10, **50**, 100], class_weight = **balanced**, bootstrap = **True**, max_features = [**'sqrt'**, 'log2', None]. The hyperparameters for the linear kernel are fully described in Table 2 of the main text.

Model	C (SVM)	Weighted F_1	Precision	Recall	Single-fit recall
Linear kernel	1e-2, **1e-3**, 1e-4	0.88 ± 0.03	0.31 ± 0.08	0.72 ± 0.16	0.74
Polynomial degree 2	**1**, 1e-2, 1e-4	0.91 ± 0.03	0.45 ± 0.21	0.44 ± 0.14	0.76
Polynomial degree 3	**1**, 1e-2, 1e-4	0.92 ± 0.02	0.47 ± 0.19	0.38 ± 0.10	0.74
Polynomial degree 4	**1**, 1e-2, 1e-4	0.93 ± 0.00	0.55 ± 0.11	0.42 ± 0.10	0.76
Radial basis functions	**1**, 1e-2, 1e-4	0.88 ± 0.03	0.25 ± 0.14	0.38 ± 0.17	0.98
RF Classifier	--	0.84 ± 0.02	0.17 ± 0.06	0.50 ± 0.19	0.88

Table S4. FDA compound screening for selected Hotspots of TEM-1 and $GABA_BR$ Active. The Hotspots selected for each protein system are ranked 1-10 and 91-100. The results are average LGFE and %rBSA for the top 20 compounds ranked by LGFE. %rBSA is the relative buried surface area expressed as a percentage. For more details regarding the docking and the set of compounds used, see the Methods section.

			TEM -1		
Rank	**Hotspot**	**Decision Fn.**	**Mean LGFE**	**Mean %rBSA**	**LGFE x %rBSA**
1	3	1.24	-11.8	100	-11.8
2	13	1.23	-7.8	99	-7.7
3	28	1.14	-9.6	98	-9.4
4	74	1.13	-5.2	79	-4.1
5	9	1.07	-12.0	99	-11.9
6	16	1.04	-11.1	99	-11.0
7	5	0.86	-11.4	100	-11.4
8	26	0.83	-10.3	99	-10.2
9	14	0.83	-9.4	98	-9.2
10	25	0.7	-11.0	98	-10.8
	Mean	**1.01**	**-9.9**	**97**	**-9.6**
91	68	-1.48	-9.0	36	-3.2
92	4	-1.49	-11.5	36	-4.1
93	85	-1.52	-8.5	30	-2.6
94	44	-1.54	-11.9	44	-5.2
95	75	-1.56	-6.4	34	-2.2
96	42	-1.61	-5.0	29	-1.5
97	66	-1.62	-8.3	35	-2.9
98	57	-1.63	-8.2	35	-2.9
99	6	-1.64	-12.0	33	-4.0
100	24	-1.66	-13.9	41	-5.7
	Mean	**-1.58**	**-9.5**	**35**	**-3.3**

			$GABA_BR$ Active		
Rank	**Hotspot**	**Decision Fn.**	**Mean LGFE**	**Mean %rBSA**	**LGFE x rBSA**
1	91	1.10	-10.3	97	-10.0
2	28	1.10	-17.5	100	-17.5
3	414	1.09	-14.4	100	-14.4
4	15	1.06	-11.7	99	-11.6
5	121	1.02	-11.5	100	-11.5
6	202	0.97	-11.8	100	-11.8
7	3	0.97	-16.1	99	-15.9
8	141	0.96	-17.7	97	-17.2
9	59	0.94	-11.0	98	-10.8
10	45	0.94	-10.2	100	-10.2
	Mean	**1.02**	**-13.2**	**99**	**-13.1**
91	70	0.22	-13.8	93	-12.8
92	451	0.21	-5.8	99	-5.7
93	72	0.20	-13.6	93	-12.6
94	80	0.20	-11.6	99	-11.5
95	169	0.19	-10.1	97	-9.8

96	277	0.18	-12.3	74	-9.1
97	391	0.16	-8.6	98	-8.4
98	217	0.15	-8.7	100	-8.7
99	307	0.14	-9.0	73	-6.6
100	139	0.14	-10.0	99	-9.9
	Mean	**0.18**	**-10.4**	**93**	**-9.7**
391	366	-1.52	-9.8	99	-9.7
392	306	-1.52	-8.3	95.3	-7.9
393	422	-1.55	-7.6	44.3	-3.4
394	457	-1.55	-8.5	92.6	-7.9
395	394	-1.55	-6.2	49.7	-3.1
396	152	-1.56	-9.7	85	-8.2
397	287	-1.58	-9.6	65.6	-6.3
398	385	-1.59	-8.1	51.6	-4.2
399	442	-1.61	-8.1	93.6	-7.6
400	424	-1.62	-6.1	92.7	-5.7
	Mean	**-1.57**	**-8.2**	**77**	**-6.4**

References

1. Ustach VD, Lakkaraju SK, Jo S, Yu W, Jiang W, MacKerell AD. Optimization and Evaluation of Site-Identification by Ligand Competitive Saturation (SILCS) as a Tool for Target-Based Ligand Optimization. J Chem Inf Model. 2019 Jun 24;59(6):3018–35.

2. Goel H, Hazel A, Ustach VD, Jo S, Yu W, MacKerell AD. Rapid and accurate estimation of protein–ligand relative binding affinities using site-identification by ligand competitive saturation. Chem Sci. 2021 Jul 1;12(25):8844–58.

3. Drug Design Data Resource (D3R). Drug Design Data Resource Grand Challenge 2 Dataset: FXR - Farnesoid X receptor [Internet]. Drug Design Data Resource (D3R); 2017 [cited 2024 Feb 19]. p. 71.5MB. Available from: https://drugdesigndata.org/about/datasets/882

4. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K_a Predictions. J Chem Theory Comput. 2011 Feb 8;7(2):525–37.

5. MacKerell AD, Jo S, Lakkaraju SK, Lind C, Yu W. Identification and characterization of fragment binding sites for allosteric ligand design using the site identification by ligand competitive saturation hotspots approach (SILCS-Hotspots). Biochim Biophys Acta Gen Subj. 2020 Apr;1864(4):129519.

6. Baumli S, Endicott JA, Johnson LN. Halogen Bonds Form the Basis for Selective P-TEFb Inhibition by DRB. Chemistry & Biology. 2010 Sep 24;17(9):931–6.

7. Wu SY, McNae I, Kontopidis G, McClue SJ, McInnes C, Stewart KJ, et al. Discovery of a Novel Family of CDK Inhibitors with the Program LIDAEUS: Structural Basis for Ligand-Induced Disordering of the Activation Loop. Structure. 2003 Apr 1;11(4):399–410.

8. Betzi S, Alam R, Martin M, Lubbers DJ, Han H, Jakkaraj SR, et al. Discovery of a Potential Allosteric Ligand Binding Site in CDK2. ACS Chem Biol. 2011 May 20;6(5):492–501.

9. Ludlow RF, Verdonk ML, Saini HK, Tickle IJ, Jhoti H. Detection of secondary binding sites in proteins using fragment screening. Proceedings of the National Academy of Sciences. 2015 Dec 29;112(52):15910–5.

10. Glatz G, Gógl G, Alexa A, Reményi A. Structural Mechanism for the Specific Assembly and Activation of the Extracellular Signal Regulated Kinase 5 (ERK5) Module*. Journal of Biological Chemistry. 2013 Mar 22;288(12):8596–609.

11. Chen H, Tucker J, Wang X, Gavine PR, Phillips C, Augustin MA, et al. Discovery of a novel allosteric inhibitor-binding site in ERK5: comparison with the canonical kinase hinge ATP-binding site. Acta Crystallographica Section D. 2016;72(5):682–93.

12. Montalibet J, Skorey K, McKay D, Scapin G, Asante-Appiah E, Kennedy BP. Residues Distant from the Active Site Influence Protein-tyrosine Phosphatase 1B Inhibitor Binding*. Journal of Biological Chemistry. 2006 Feb 24;281(8):5258–66.

13. Wiesmann C, Barr KJ, Kung J, Zhu J, Erlanson DA, Shen W, et al. Allosteric inhibition of protein tyrosine phosphatase 1B. Nat Struct Mol Biol. 2004 Aug;11(8):730–7.

14. Wan ZK, Follows B, Kirincich S, Wilson D, Binnun E, Xu W, et al. Probing acid replacements of thiophene PTP1B inhibitors. Bioorganic & Medicinal Chemistry Letters. 2007 May 15;17(10):2913–20.

15. Han Y, Belley M, Bayly CI, Colucci J, Dufresne C, Giroux A, et al. Discovery of [(3-bromo-7-cyano-2-naphthyl)(difluoro)methyl]phosphonic acid, a potent and orally active small molecule PTP1B inhibitor. Bioorganic & Medicinal Chemistry Letters. 2008 Jun 1;18(11):3200–5.

16. Rasmussen SGF, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, et al. Crystal structure of the β2 adrenergic receptor–Gs protein complex. Nature. 2011 Sep 29;477(7366):549–55.

17. Liu X, Ahn S, Kahsai AW, Meng KC, Latorraca NR, Pani B, et al. Mechanism of intracellular allosteric β2AR antagonist revealed by X-ray crystal structure. Nature. 2017 Aug;548(7668):480–4.

18. Srivastava A, Yano J, Hirozane Y, Kefala G, Gruswitz F, Snell G, et al. High-resolution structure of the human GPR40 receptor bound to allosteric agonist TAK-875. Nature. 2014 Sep;513(7516):124–7.

19. Ho JD, Chau B, Rodgers L, Lu F, Wilbur KL, Otto KA, et al. Structural basis for GPR40 allosteric agonism and incretin stimulation. Nat Commun. 2018 Apr 25;9(1):1645.

20. Haga K, Kruse AC, Asada H, Yurugi-Kobayashi T, Shiroishi M, Zhang C, et al. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. Nature. 2012 Feb;482(7386):547–51.

21. Kruse AC, Ring AM, Manglik A, Hu J, Hu K, Eitel K, et al. Activation and allosteric modulation of a muscarinic acetylcholine receptor. Nature. 2013 Dec;504(7478):101–6.

22. Pereira de Jésus-Tran K, Côté PL, Cantin L, Blanchet J, Labrie F, Breton R. Comparison of crystal structures of human androgen receptor ligand-binding domain complexed with various agonists reveals molecular determinants responsible for binding affinity. Protein Science. 2006;15(5):987–99.

23. Estébanez-Perpiñá E, Arnold LA, Nguyen P, Rodrigues ED, Mar E, Bateman R, et al. A surface on the androgen receptor that allosterically regulates coactivator binding. Proceedings of the National Academy of Sciences. 2007 Oct 9;104(41):16074–9.

24. Goldstein DM, Soth M, Gabriel T, Dewdney N, Kuglstatter A, Arzeno H, et al. Discovery of 6-(2,4-Difluorophenoxy)-2-[3-hydroxy-1-(2-hydroxyethyl)propylamino]-8-methyl-8H-pyrido[2,3-d]pyrimidin-7-one (Pamapimod) and 6-(2,4-Difluorophenoxy)-8-methyl-2-(tetrahydro-2H-pyran-4-ylamino)pyrido[2,3-d]pyrimidin-7(8H)-one (R1487) as Orally Bioavailable and Highly Selective Inhibitors of p38α Mitogen-Activated Protein Kinase. J Med Chem. 2011 Apr 14;54(7):2255–65.

25. Pargellis C, Tong L, Churchill L, Cirillo PF, Gilmore T, Graham AG, et al. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. Nat Struct Mol Biol. 2002 Apr;9(4):268–72.

26. Cumming JN, Smith EM, Wang L, Misiaszek J, Durkin J, Pan J, et al. Structure based design of iminohydantoin BACE1 inhibitors: Identification of an orally available, centrally active BACE1 inhibitor. Bioorganic & Medicinal Chemistry Letters. 2012 Apr 1;22(7):2444–9.

27. D3R | Drug Design Data Resource Grand Challenge 4 Dataset: BACE1 [Internet]. [cited 2024 Feb 19]. Available from: https://drugdesigndata.org/about/datasets/2027

28. Congreve M, Chessari G, Tisi D, Woodhead AJ. Recent Developments in Fragment-Based Drug Discovery. J Med Chem. 2008 Jul 1;51(13):3661–80.

29. Baum B, Muley L, Smolinski M, Heine A, Hangauer D, Klebe G. Non-additivity of Functional Group Contributions in Protein–Ligand Binding: A Comprehensive Study by Crystallography and Isothermal Titration Calorimetry. Journal of Molecular Biology. 2010 Apr 9;397(4):1042–54.

30. Friberg A, Vigil D, Zhao B, Daniels RN, Burke JP, Garcia-Barrantes PM, et al. Discovery of Potent Myeloid Cell Leukemia 1 (Mcl-1) Inhibitors Using Fragment-Based Methods and Structure-Based Design. J Med Chem. 2013 Jan 10;56(1):15–30.

31. Sato M, Arakawa T, Nam YW, Nishimoto M, Kitaoka M, Fushinobu S. Open–close structural change upon ligand binding and two magnesium ions required for the catalysis of *N*-acetylhexosamine 1-kinase. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2015 May 1;1854(5):333–40.

32. Tarver CL. Molecular role of angiopoietin-like 4's carboxy-terminal domain in pancreatic ductal adenocarcinoma progression [Dissertations]. University of Huntsville Alabama; 2019.

33. Wang X, Minasov G, Shoichet BK. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. Journal of Molecular Biology. 2002 Jun 28;320(1):85–95.

34. Ness S, Martin R, Kindler AM, Paetzel M, Gold M, Jensen SE, et al. Structure-Based Design Guides the Improved Efficacy of Deacylation Transition State Analogue Inhibitors of TEM-1 β-Lactamase,. Biochemistry. 2000 May 1;39(18):5312–21.

35. Horn JR, Shoichet BK. Allosteric Inhibition Through Core Disruption. Journal of Molecular Biology. 2004 Mar 5;336(5):1283–91.

36. Kim Y, Jeong E, Jeong JH, Kim Y, Cho Y. Structural Basis for Activation of the Heterodimeric $GABA_B$ Receptor. Journal of Molecular Biology. 2020 Nov 6;432(22):5966–84.

37. Shaye H, Ishchenko A, Lam JH, Han GW, Xue L, Rondard P, et al. Structural basis of the activation of a metabotropic GABA receptor. Nature. 2020 Aug;584(7820):298–303.

38. Mao C, Shen C, Li C, Shen DD, Xu C, Zhang S, et al. Cryo-EM structures of inactive and active $GABA_B$ receptor. Cell Res. 2020 Jul;30(7):564–73.