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

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 ML 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 67% and 89% 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. Historically, computational binding pocket identification was first carried out using the protein molecular surface defined with the LJ potential and a grid of lattice points sampling the space around that surface.[9] Standard methods still often use geometric analysis,[10–12] in addition to molecular docking, and/or machine-learning.[13] When a representative structure is available and the binding pocket is relatively well-defined, methods including FTMap[14–16] and Fpocket[17] are effective, as well as the widely-used methods related to common CADD software packages, such as SiteMap[18,19] (Glide/Schrödinger),[20] SiteFinder[21] (MOE/Chemical Computing Group), or

AutoLigand[22] (AutoDock).[23] Some methods employ template based modeling to predict binding sites when only a sequence is known.[24–27] PepSite uses 3D grids of position-specific scoring matrices to efficiently identify linear peptide binding sites across the proteome, an interesting approach for a highly-specialized class of ligand-protein interactions.[28] There are many machine-/deep-learning models[13,29] that incorporate geometry, sequence-homology, structural features, molecular docking, and/or consensus to predict ligand binding sites.[30–36] The recently published AlphaFold 3 model claims to predict protein-ligand interactions with higher fidelity than standard docking methods,[37] although the web server available for non-commercial researchers only predicts sites for nineteen common cofactors like ATP and citric acid. To remain highly computationally efficient, methods reliant on static structures necessarily neglect protein backbone flexibility, thus cannot capture protein allostery or cryptic binding sites.[38–42] In addition, the traditional molecular docking approaches used in available methods,[43,20,23,44,45] while efficiently sampling known ligand-protein interactions,[16,34] 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 MD simulations, and of particular interest, all-atom cosolute MD simulations.[46,47] Alternatively, a key example of a natural, non-cosolute approach to incorporating dynamics into site prediction is to utilize enhanced sampling or coarse grained simulations to sample pocket openings, and include the resulting dynamics in the inputs to a ML model, such as the method CryptoSite.[42] On the other hand, cosolute methods are conceptually similar to experimental fragment-based drug design[48,49] wherein proteins are co-crystallized with various small solutes to determine their binding sites.[50] In general, cosolute methods involve solvating the target biomolecule with various small molecules and performing molecular simulations to analyze the distribution of the molecules over the course of the simulation. This approach is widely-employed[51–56] including by MDmix,[46,57] pMD-Membrane,[58,59] Mix-MD,[60–62] SWISH and SWISH-X,[63,64] Cosolvent Analysis Toolkit (CAT),[65] and SILCS.[47,66,67] The coarse grain MD cosolute method Colabind was recently released,[68] 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,[69–72] combined with consistent improvements in the accuracy of all-atom force fields,[73–77] 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[78,79] 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.[80,81] In a previous paper, a method was presented for identifying a comprehensive set of fragment binding sites, or Hotspots, on proteins,[82] and subsequently applied to RNA.[83] 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.

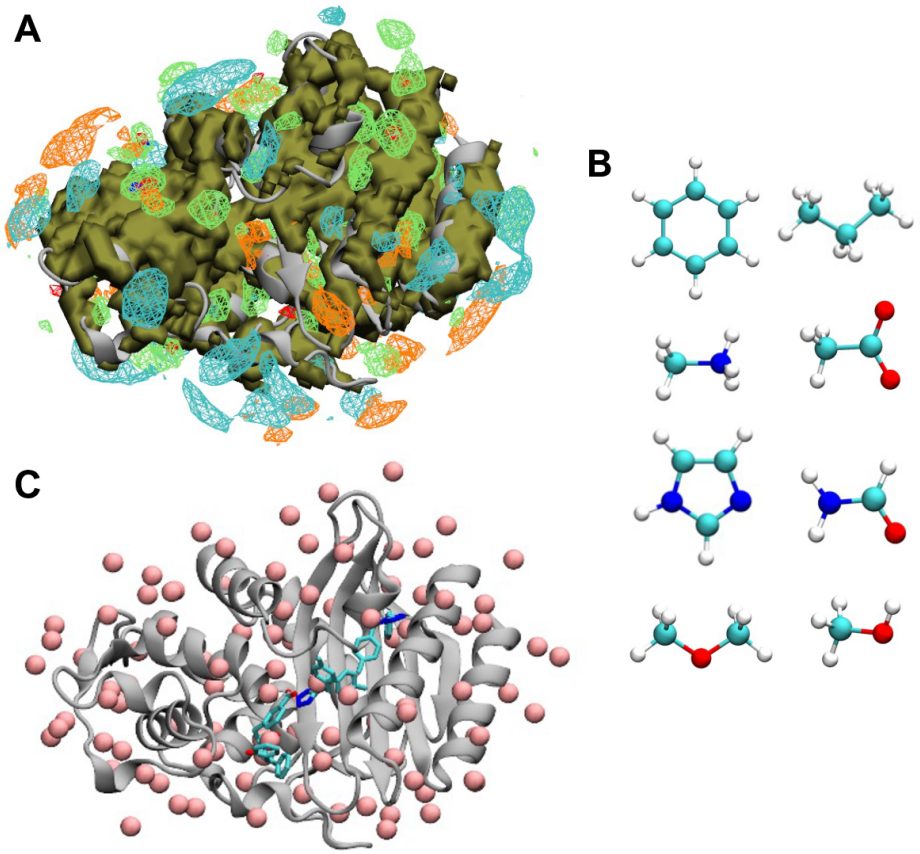

Figure 1: Example SILCS FragMap and Hotspots and depiction of the SILCS solutes. A) TEM-1 β-lactamase is rendered in NewCartoon style (PDB: 1JWP), with the various FragMaps contoured at -1.2 kcal/mol. The green map corresponds to generic apolar carbons (propane and benzene carbon), the red corresponds 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 Hotspots 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[84–86] 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 67% and 89% 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.[47,67] 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,[87,88] 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.[89,90] The CHARMM-GUI webserver was also used to generate small missing loops (<12 amino acids) and to adjust the protonation state of titratable residues.[89,90] The protonation state of titratable residues at pH 7.0 was determined using PropKa3.[91] The FragMaps were obtained from our previous study[82] that were performed using SILCS software version 2019 (SilcsBio LLC) and Gromacs version 2019, except for ANGPTL4, TEM-1, NKG2D, and $GABA_BR$, for which SILCS software version 2023[92] and Gromacs version 2022 were used.[69,70] The SILCS simulations are based on a published GCMC/MD approach[78] that has not been changed beyond porting the GCMC code to GPUs[79] that is implemented in version 2023. The computations for each set of SILCS FragMap using version 2023, were carried out in parallel on ten compute nodes each with 1 GPU (e.g. GTX 980, GTX 1080Ti, RTX 2080Ti) and eight CPU threads (e.g. AMD Ryzen 7 1700, AMD EPYC 7551P), and require between ~1-7 days to complete depending on the system size. The full simulation boxes in this study contain between ~35,000 and ~190,000 atoms.

After calculating the FragMaps, we performed the SILCS-Hotspots calculation as described in our previous work.[82] The Hotspots calculation consists of comprehensively docking a library 90 mono- and bicyclic fragments[93] 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.[92] The SILCS-Hotspots docking performed for this study utilized a GPU implementation of SILCS-MC docking.[94] The SILCS-Hotspots calculations generated ~6,000 to ~65,000 independent SILCS-MC jobs that each run for ~15 sec total and can be scheduled to run in parallel on a given cluster.

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[95] and Drugs@FDA.[96] 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.[97]

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 in the presence of the protein (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). The protein-derived features are similar to those used in previous ML models.[98,99] As a feature we wanted the calculation of the SASA of a Hotspot in the presence of the protein 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[100] 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 was implemented. The calculation of the SASA and volume of the Hotspot clusters requires substantial CPU time, and so the algorithms were parallelized with numba.[101]

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:[82] Cyclin-dependent kinase 2 (CDK2) in both active and inactive states,[102,103] Extracellular-signal-regulated kinase 5 (ERK5),[104] Protein tyrosine phosphatase 1b (PTP1B),[105–108] Androgen receptor,[109,110] and three G-protein coupled receptors (GPCRs), namely G protein-coupled receptor 40 (GPR40),[111,112] M2 Muscarinic receptor,[113,114] and β2 Adrenergic receptor.[115,116] The validation set is comprised of eleven proteins, seven of which we recycle from previous SILCS-MC publications.[80,81] namely: P38 mitogen-activated protein kinase,[117,118] Farnesoid X bile acid receptor (FXR),[119] β-Secretase 1 (BACE1),[120,121] tRNA methyl transferase (TrmD),[122] Myeloid cell leukemia 1 (MCL1),[123,124] Heat-shock protein 90 kDa (Hsp90),[48] and Thrombin.[125] To those we added the C-terminal domain of the lipid-binding protein angiopoietin-like 4 (ANGPTL4),[126] TEM-1 β-lactamase,[127–129] Natural killer group 2D receptor (NKG2D),[130,131] and GPCR γ-aminobutyric acid receptor ($GABA_BR$) in both active and inactive states.[132–134]

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[135] 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 that 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 were assumed to 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.

We used the proteins TEM-1 and NKG2D, both containing cryptic sites, to benchmark our method against three alternative methods, namely CryptoSite,[42] SiteMap[18,19] and SiteFinder.[21] Note that previously the SILCS-Hotspots approach was also benchmarked against FTMap and Fpocket. These proteins are in common between our validation set and a recent method employing SiteMap and SiteFinder to identify cryptic sites, which found that both SiteMap and SiteFinder struggled to identify the cryptic sites on these two proteins.[136] We used the free, online CryptoSite server at https://modbase.compbio.ucsf.edu/cryptosite to obtain the results of the predictions using the apo structures of each protein listed in Table S1. The results took ~ 7 hours, although the site and original publication notes that on average there can be a total time of 1-2 days depending on the server load.[42]

Table 2: Linear SVM hyperparameters. Descriptions of hyperparameters are adapted from the sci-kit learn library documentation.[137] 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 nonlinear kernels resulted in over-training (Table S3). While all models generated reasonable average weighted F_1 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[138] to identify the optimal number of input principal components and reduce the risk of overfitting by reducing the dimensionality of the inputs (Figure S3A). The first 22 principal components were selected, corresponding to the maximum weighted F_1 in Figure S3A. The distribution of the data in the first two principal components is given in Figure S3B, indicating that the two classes are somewhat linearly separable. 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.[139,140] The Decision Function is positive for higher confidence, and negative for confidence that the Hotspot is not a suitable binding site. The ML scripts were written using the scikit-learn version 1.3.0[137] and pandas 2.0.3[141] python libraries. All 3D molecular renderings were generated using VMD version 1.9.3,[142] and all plots were created with the python library matplotlib[143] using the accessible color sequences of Petroff.[144]

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 eleven 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 eleven proteins in the validation set, seven were used in previous SILCS-MC benchmarking studies, and as such each contain a single orthosteric binding site.[80,81] 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.[118] 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.[117] We collected five additional systems, ANGPTL4, TEM-1, NKG2D, 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).[126] TEM-1 was selected because of its cryptic allosteric binding site,[38,128] which is absent in the apo structure we used for the SILCS simulation (PDB: 1JWP).[127] Similarly, NKG2D was selected for a cryptic allosteric site.[130,131] For the $GABA_BR$, as previously described for the CDK2 system,[82] 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[132] 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.[82] 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 ML 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.

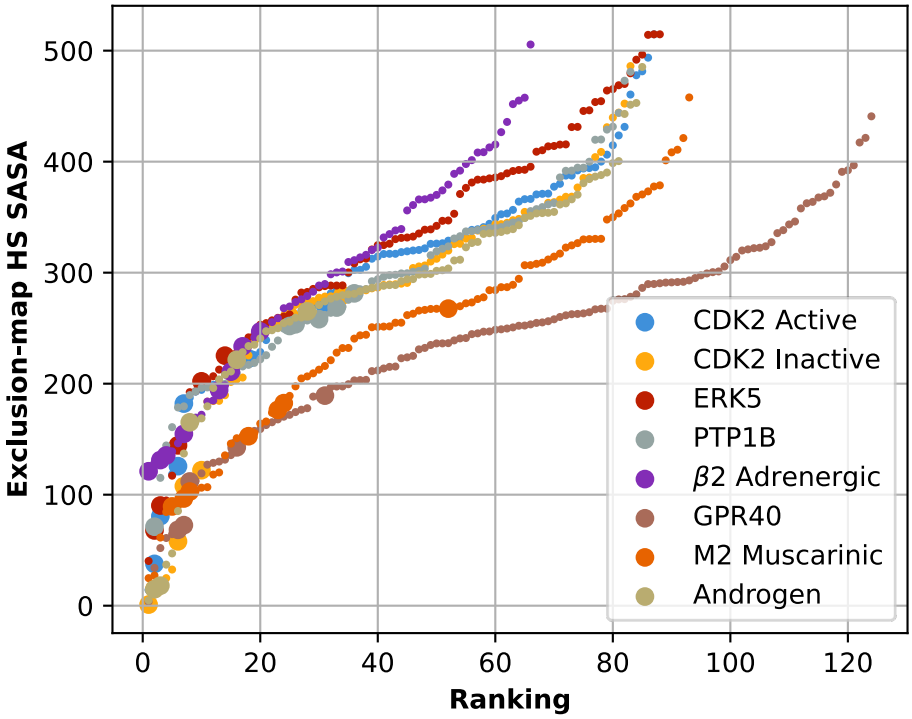

Figure 2: Ranking based on Exclusion-map HS SASA of individual Hotspots with a minimum of one adjacent Hotspot within 12 Å. The larger circles denote Hotspots within 5 Å of a non-hydrogen atom of a drug-like compound bound to the proteins.

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 ML 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 ML 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.[137,139] 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).

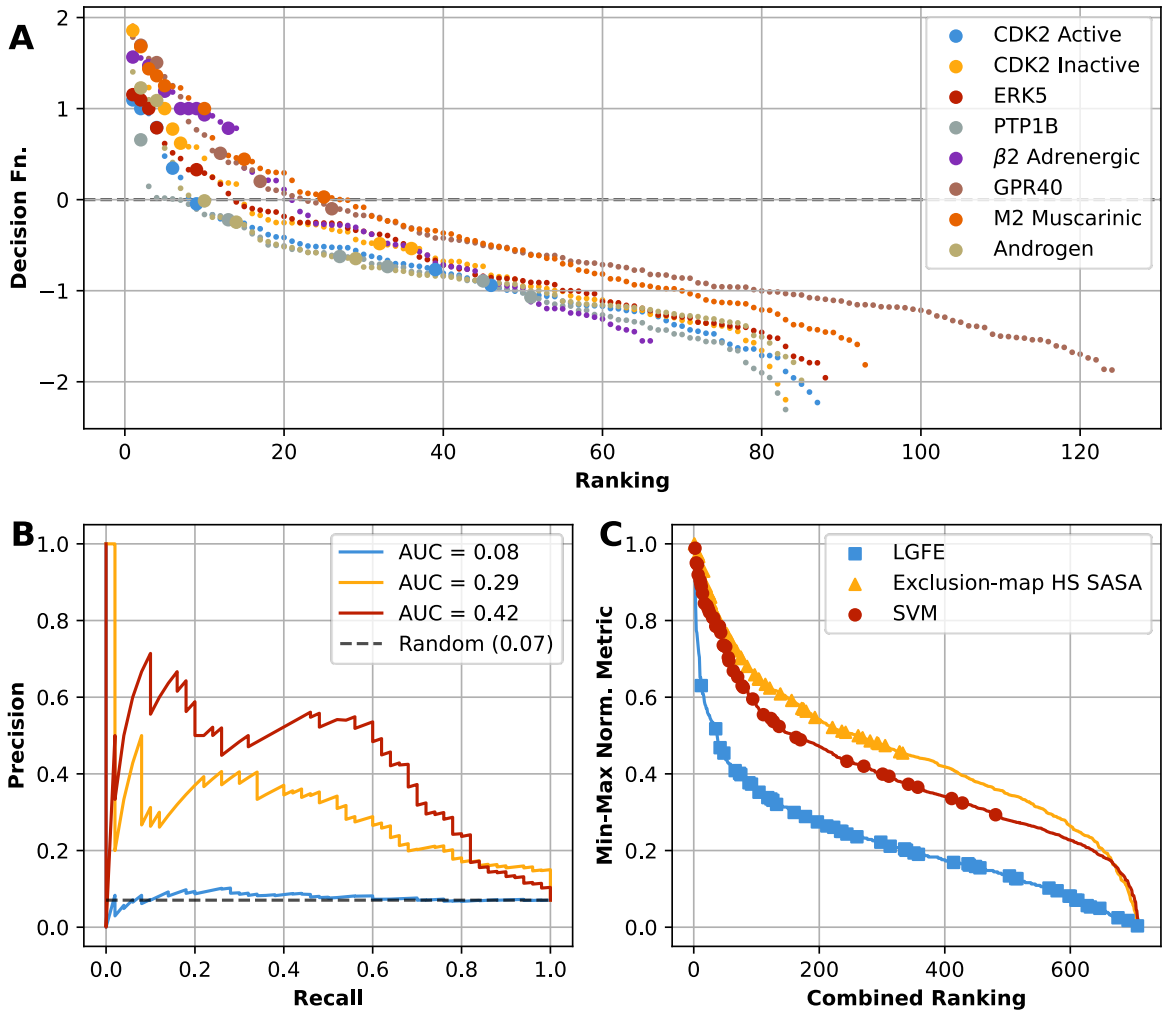

Figure 3: Performance of final model on the training set. A) Ranking of each protein's Hotspots by the final SVM model's Decision Function with Hotspots within 5 Å of the non-hydrogen atoms of known drug-like molecules (true hits) shown as large circles. **B)** Precision-Recall curves of the original 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 large 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.[82] In the top 20 Hotspots the SVM model fails to identify three out of twenty-five 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.[82] 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 67% of ligand binding sites in the top 10, and 89% 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.

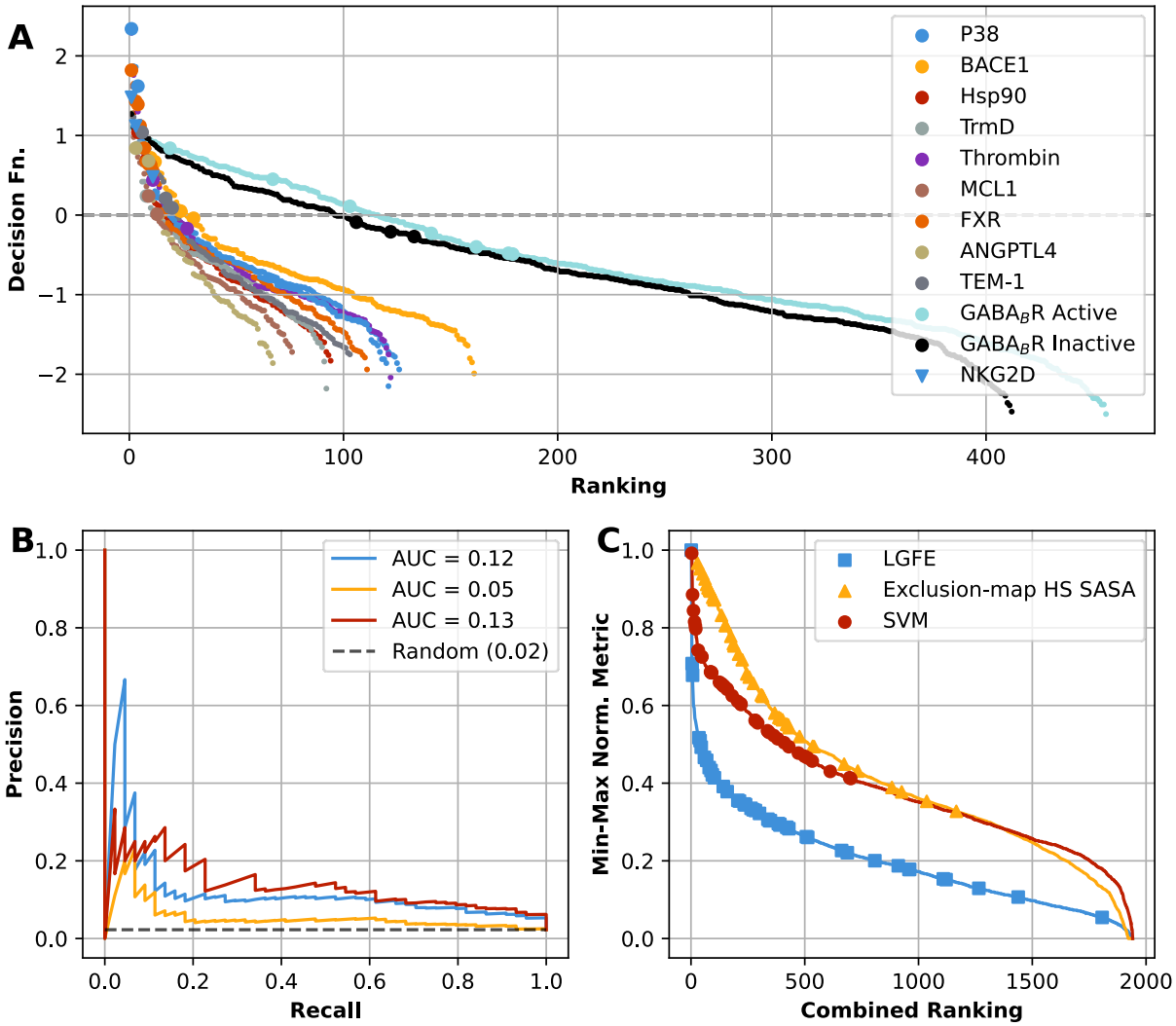

Figure 4: Performance of final model on the validation set. A) Ranking of each protein's Hotspots by the final SVM model's Decision Function with Hotspots within 5 Å of the non-hydrogen atoms of known drug-like molecules (true hits) shown as large circles. **B)** Precision-Recall curves of the original 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 large 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 $GABA_BR$ Inactive, despite being identified in the $GABA_BR$ 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, since this burial is not explicitly accounted for in the highly-predictive Exclusion map surface area calculations.

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
NKG2D	1	1.00	1.00	1.00
Total	**18**	**0.61**	**0.72**	**0.83**
		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
NKG2D	1	1.00	1.00	1.00
Total	**18**	**0.56**	**0.72**	**0.78**
		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
NKG2D	1	1.00	1.00	1.00
Total	**18**	**0.67**	**0.89**	**0.89**

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 mean 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.

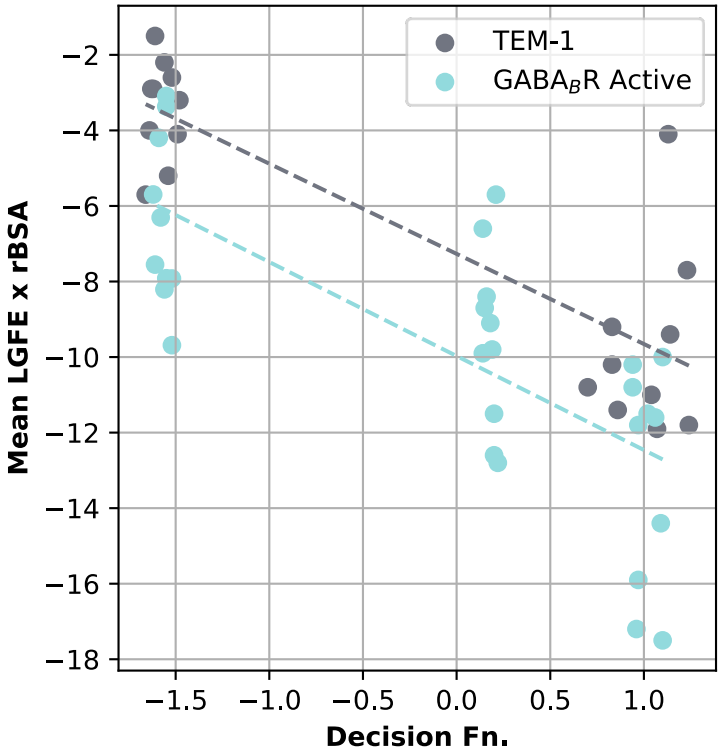

Figure 5: SVM model Decision Function and the Mean LGFE times rBSA for selected Hotspots. For TEM-1 and GABA$_B$R, the Hotspots 1-10 and 91-100 were selected, and for GABA$_B$R Hotspots 391-400 were also selected. The trendlines show the linear line of best fit. For TEM-1 Hotspots 1-10 and 91-100 correspond 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$_B$R versus TEM-1, which biases the overall distribution towards lower ranking SVM Decision Function scores.

Comparison to existing methods of cryptic binding site prediction

In our previous work introducing the SILCS-Hotspots method, we compared the Hotspots generated against the fragment binding sites identified by FTMap[16] and Fpocket,[17] and found that SILCS-Hotspots identifies more Hotspots near the crystallographic sites than the other methods.[82] To give a sense of the performance of the model against other available cryptic binding site identification methods, we selected two proteins in our validation set, TEM-1 and NKG2D, to compare with CryptoSite.[42] These cryptic sites were selected because they were recently identified[136] as being particularly challenging to SiteMap (Schrödinger, Inc.)[18,19] and SiteFinder (Chemical Computing Group).[21] CryptoSite successfully identified the cryptic site in NKG2D (Figure S6). As noted in the original CryptoSite paper, it identifies the residues involved in the disruption of a core region upon ligand binding to the cryptic site of TEM-1, although the scores of ~0.06-0.08 are below the typical CrytoSite cutoff score of 0.1 (Figure S6).[42] These results suggest that both CryptoSite and SILCS-Hotspots perform better than either SiteMap or SiteFinder at identifying cryptic sites. It should be noted that CryptoSite requires more

computation than SiteMap/SiteFinder, and similarly SILCS-Hotspots requires more than CryptoSite associated with the computational requirements of the initial SILCS Simulations. The SILCS-Hotspots method is not intended to be used as a standalone tool, but as part of the integrated SILCS workflow with methods for site identification, pharmacophore discovery and lead optimization.

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 89% recall in the top 20 on the validation set (16 out of 18 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 may have limitations associated with sites adjacent to the lipid bilayer, such as the site observed in $GABA_BR$ Active state. Future efforts will focus on addressing this issue, such as by directly accounting for burial in lipids and by constructing a training set of sites at protein-bilayer interfaces. Furthermore, while the model has been tested on a reasonably diverse test set of proteins including challenging cryptic sites, more extensive testing is necessary to conclude the model will generalize to exotic systems. We expect that this relatively simple classification model with the physical insights from SILCS sampling will tend to generalize well.

Supporting Information:

Figure S1: Surface-exposed Hotspot 25 in ERK5.
Figure S2: Distribution of Hotspot SASA by protein system.
Figure S3. Analysis of the recursive feature elimination and the top two principal components (PCs) of the training set.
Figure S4: Ranking based on mean LGFE of each Hotspot.
Figure S5: Burial of allosteric binding site between $GABA_BR$ Active TM domains.
Figure S6: CryptoSite predictions for NKG2D (A) and TEM-1 (B).

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.

Data and Software Availability

Information about the training and validation set, including the crystallographic ligands and the adjacent Hotspots, is provided in Table S1 and Table S2. The compounds used to perform the FDA analysis in sdf and pdf file formats, as well as all the data in training and test data sets in csv format, are provided free on GitHub at https://github.com/mackerell-lab/FDA-compounds-SILCS-Hotspots-SI.

References

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(2) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database:

Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Research* **2022**, *50* (D1), D439–D444. https://doi.org/10.1093/nar/gkab1061.

(3) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596* (7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1.

(4) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A Comprehensive Map of Molecular Drug Targets. *Nat Rev Drug Discov* **2017**, *16* (1), 19–34. https://doi.org/10.1038/nrd.2016.230.

(5) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(6) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871–876. https://doi.org/10.1126/science.abj8754.

(7) Pandey, M.; Fernandez, M.; Gentile, F.; Isayev, O.; Tropsha, A.; Stern, A. C.; Cherkasov, A. The Transformational Role of GPU Computing and Deep Learning in Drug Discovery. *Nat Mach Intell* **2022**, *4* (3), 211–221. https://doi.org/10.1038/s42256-022-00463-x.

(8) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating Molecular Dynamic Simulation on Graphics Processing Units. *J Comput Chem* **2009**, *30* (6), 864–872. https://doi.org/10.1002/jcc.21209.

(9) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857. https://doi.org/10.1021/jm00145a002.

(10) Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites. *Bioinformatics* **2005**, *21* (9), 1908–1916. https://doi.org/10.1093/bioinformatics/bti315.

(11) Siragusa, L.; Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. BioGPS: Navigating Biological Space to Predict Polypharmacology, off-Targeting, and Selectivity. *Proteins: Structure, Function, and Bioinformatics* **2015**, *83* (3), 517–532. https://doi.org/10.1002/prot.24753.

(12) Gagliardi, L.; Rocchia, W. SiteFerret: Beyond Simple Pocket Identification in Proteins. *J. Chem. Theory Comput.* **2023**, *19* (15), 5242–5259. https://doi.org/10.1021/acs.jctc.2c01306.
(13) Zhao, J.; Cao, Y.; Zhang, L. Exploring the Computational Methods for Protein-Ligand Binding Site Prediction. *Computational and Structural Biotechnology Journal* **2020**, *18*, 417–426. https://doi.org/10.1016/j.csbj.2020.02.008.
(14) Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-Based Identification of Druggable "hot Spots" of Proteins Using Fourier Domain Correlation Techniques. *Bioinformatics* **2009**, *25* (5), 621–627. https://doi.org/10.1093/bioinformatics/btp036.
(15) Ngan, C.-H.; Hall, D. R.; Zerbe, B.; Grove, L. E.; Kozakov, D.; Vajda, S. FTSite: High Accuracy Detection of Ligand Binding Sites on Unbound Protein Structures. *Bioinformatics* **2012**, *28* (2), 286–287. https://doi.org/10.1093/bioinformatics/btr651.
(16) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat Protoc* **2015**, *10* (5), 733–755. https://doi.org/10.1038/nprot.2015.043.
(17) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10* (1), 168. https://doi.org/10.1186/1471-2105-10-168.
(18) Halgren, T. New Method for Fast and Accurate Binding-Site Identification and Analysis. *Chem Biol Drug Des* **2007**, *69* (2), 146–148. https://doi.org/10.1111/j.1747-0285.2007.00483.x.
(19) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49* (2), 377–389. https://doi.org/10.1021/ci800324m.
(20) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.* **2006**, *49* (21), 6177–6196. https://doi.org/10.1021/jm051256o.
(21) *Finding Druggable Binding Pockets Using SiteFinder*. https://video.chemcomp.com/watch/2VtMGBYvvMkumZqo8A3yJN?custom_id= (accessed 2024-07-28).
(22) Harris, R.; Olson, A. J.; Goodsell, D. S. Automated Prediction of Ligand-Binding Sites in Proteins. *Proteins* **2008**, *70* (4), 1506–1517. https://doi.org/10.1002/prot.21645.
(23) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.
(24) Capra, J. A.; Singh, M. Predicting Functionally Important Residues from Sequence Conservation. *Bioinformatics* **2007**, *23* (15), 1875–1882. https://doi.org/10.1093/bioinformatics/btm270.
(25) Roy, A.; Zhang, Y. Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement. *Structure* **2012**, *20* (6), 987–997. https://doi.org/10.1016/j.str.2012.03.009.

(26) Roche, D. B.; Tetchner, S. J.; McGuffin, L. J. FunFOLD: An Improved Automated Method for the Prediction of Ligand Binding Residues Using 3D Models of Proteins. *BMC Bioinformatics* **2011**, *12* (1), 160. https://doi.org/10.1186/1471-2105-12-160.
(27) Wass, M. N.; Kelley, L. A.; Sternberg, M. J. E. 3DLigandSite: Predicting Ligand-Binding Sites Using Similar Structures. *Nucleic Acids Research* **2010**, *38* (suppl_2), W469–W473. https://doi.org/10.1093/nar/gkq406.
(28) Trabuco, L. G.; Lise, S.; Petsalaki, E.; Russell, R. B. PepSite: Prediction of Peptide-Binding Sites from Protein Surfaces. *Nucleic Acids Research* **2012**, *40* (W1), W423–W427. https://doi.org/10.1093/nar/gks398.
(29) 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**, *27* (7), 573–587. https://doi.org/10.1080/1062936X.2016.1217271.
(30) Yang, J.; Roy, A.; Zhang, Y. Protein–Ligand Binding Site Recognition Using Complementary Binding-Specific Substructure Comparison and Sequence Profile Alignment. *Bioinformatics* **2013**, *29* (20), 2588–2595. https://doi.org/10.1093/bioinformatics/btt447.
(31) Huang, B. MetaPocket: A Meta Approach to Improve Protein Ligand Binding Site Prediction. *OMICS: A Journal of Integrative Biology* **2009**, *13* (4), 325–330. https://doi.org/10.1089/omi.2009.0045.
(32) Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A. Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLOS Computational Biology* **2009**, *5* (12), e1000585. https://doi.org/10.1371/journal.pcbi.1000585.
(33) Morrone Xavier, M.; Sehnem Heck, G.; Boff de Avila, M.; Maria Bernhardt Levin, N.; Oliveira Pintro, V.; Lemes Carvalho, N.; Filgueira de Azevedo, W. SAnDReS a Computational Tool for Statistical Analysis of Docking Results and Development of Scoring Functions. *Combinatorial Chemistry & High Throughput Screening* **2016**, *19* (10), 801–812.
(34) 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**, *46* (W1), W438–W442. https://doi.org/10.1093/nar/gky439.
(35) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Improving Detection of Protein-Ligand Binding Sites with 3D Segmentation. *Sci Rep* **2020**, *10* (1), 5035. https://doi.org/10.1038/s41598-020-61860-z.
(36) Trisciuzzi, D.; Siragusa, L.; Baroni, M.; Cruciani, G.; Nicolotti, O. An Integrated Machine Learning Model To Spot Peptide Binding Pockets in 3D Protein Screening. *J. Chem. Inf. Model.* **2022**, *62* (24), 6812–6824. https://doi.org/10.1021/acs.jcim.2c00583.
(37) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.;

Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.

(38) Vajda, S.; Beglov, D.; Wakefield, A. E.; Egbert, M.; Whitty, A. Cryptic Binding Sites on Proteins: Definition, Detection, and Druggability. *Curr Opin Chem Biol* **2018**, *44*, 1–8. https://doi.org/10.1016/j.cbpa.2018.05.003.

(39) Schmidtke, P.; Bidon-Chanal, A.; Luque, F. J.; Barril, X. MDpocket: Open-Source Cavity Detection and Characterization on Molecular Dynamics Trajectories. *Bioinformatics* **2011**, *27* (23), 3276–3285. https://doi.org/10.1093/bioinformatics/btr550.

(40) Bowman, G. R.; Geissler, P. L. Equilibrium Fluctuations of a Single Folded Protein Reveal a Multitude of Potential Cryptic Allosteric Sites. *Proceedings of the National Academy of Sciences* **2012**, *109* (29), 11681–11686. https://doi.org/10.1073/pnas.1209309109.

(41) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. Discovery of Multiple Hidden Allosteric Sites by Combining Markov State Models and Experiments. *Proceedings of the National Academy of Sciences* **2015**, *112* (9), 2734–2739. https://doi.org/10.1073/pnas.1417811112.

(42) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J Mol Biol* **2016**, *428* (4), 709–719. https://doi.org/10.1016/j.jmb.2016.01.029.

(43) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein–Ligand Docking Using GOLD. *Proteins: Structure, Function, and Bioinformatics* **2003**, *52* (4), 609–623. https://doi.org/10.1002/prot.10465.

(44) Trott, O.; Olson, A. J. 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), 455–461. https://doi.org/10.1002/jcc.21334.

(45) Zhang, N.; Zhao, H. Enriching Screening Libraries with Bioactive Fragment Space. *Bioorganic & Medicinal Chemistry Letters* **2016**, *26* (15), 3594–3597. https://doi.org/10.1016/j.bmcl.2016.06.013.

(46) Seco, J.; Luque, F. J.; Barril, X. Binding Site Detection and Druggability Index from First Principles. *J. Med. Chem.* **2009**, *52* (8), 2363–2371. https://doi.org/10.1021/jm801385d.

(47) Guvench, O.; MacKerell Jr., A. D. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLOS Computational Biology* **2009**, *5* (7), e1000435. https://doi.org/10.1371/journal.pcbi.1000435.

(48) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent Developments in Fragment-Based Drug Discovery. *J. Med. Chem.* **2008**, *51* (13), 3661–3680. https://doi.org/10.1021/jm8000373.

(49) Kirsch, P.; Hartman, A. M.; Hirsch, A. K. H.; Empting, M. Concepts and Core Principles of Fragment-Based Drug Design. *Molecules* **2019**, *24* (23), 4309. https://doi.org/10.3390/molecules24234309.

(50) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J. Phys. Chem.* **1996**, *100* (7), 2605–2611. https://doi.org/10.1021/jp952516o.

(51) Basse, N.; Kaar, J. L.; Settanni, G.; Joerger, A. C.; Rutherford, T. J.; Fersht, A. R. Toward the Rational Design of P53-Stabilizing Drugs: Probing the Surface of the Oncogenic Y220C Mutant. *Chem Biol* **2010**, *17* (1), 46–56. https://doi.org/10.1016/j.chembiol.2009.12.011.
(52) Yang, C.-Y.; Wang, S. Computational Analysis of Protein Hotspots. *ACS Med. Chem. Lett.* **2010**, *1* (3), 125–129. https://doi.org/10.1021/ml100026a.
(53) Tan, Y. S.; Śledź, P.; Lang, S.; Stubbs, C. J.; Spring, D. R.; Abell, C.; Best, R. B. 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**, *51* (40), 10078–10081. https://doi.org/10.1002/anie.201205676.
(54) Huang, D.; Caflisch, A. Small Molecule Binding to Proteins: Affinity and Binding/Unbinding Dynamics from Atomistic Simulations. *ChemMedChem* **2011**, *6* (9), 1578–1580. https://doi.org/10.1002/cmdc.201100237.
(55) Bakan, A.; Nevins, N.; Lakdawala, A. S.; Bahar, I. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J Chem Theory Comput* **2012**, *8* (7), 2435–2447. https://doi.org/10.1021/ct300117j.
(56) Ghanakota, P.; Carlson, H. A. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J. Med. Chem.* **2016**, *59* (23), 10383–10399. https://doi.org/10.1021/acs.jmedchem.6b00399.
(57) 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**, *57* (20), 8530–8539. https://doi.org/10.1021/jm5010418.
(58) Prakash, P.; Sayyed-Ahmad, A.; Gorfe, A. A. pMD-Membrane: A Method for Ligand Binding Site Identification in Membrane-Bound Proteins. *PLOS Computational Biology* **2015**, *11* (10), e1004469. https://doi.org/10.1371/journal.pcbi.1004469.
(59) Sayyed-Ahmad, A.; Gorfe, A. A. Mixed-Probe Simulation and Probe-Derived Surface Topography Map Analysis for Ligand Binding Site Identification. *J. Chem. Theory Comput.* **2017**, *13* (4), 1851–1861. https://doi.org/10.1021/acs.jctc.7b00130.
(60) Ghanakota, P.; Carlson, H. A. Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *J. Phys. Chem. B* **2016**, *120* (33), 8685–8695. https://doi.org/10.1021/acs.jpcb.6b03515.
(61) Graham, S. E.; Leja, N.; Carlson, H. A. MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations. *J. Chem. Inf. Model.* **2018**, *58* (7), 1426–1433. https://doi.org/10.1021/acs.jcim.8b00265.
(62) Smith, R. D.; Carlson, H. A. Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics. *J Chem Inf Model* **2021**, *61* (3), 1287–1299. https://doi.org/10.1021/acs.jcim.0c01002.
(63) Comitani, F.; Gervasio, F. L. Exploring Cryptic Pockets Formation in Targets of Pharmaceutical Interest with SWISH. *J. Chem. Theory Comput.* **2018**, *14* (6), 3321–3331. https://doi.org/10.1021/acs.jctc.8b00263.
(64) Borsatto, A.; Gianquinto, E.; Rizzi, V.; Gervasio, F. L. SWISH-X, an Expanded Approach to Detect Cryptic Pockets in Proteins and at Protein–Protein Interfaces. *J. Chem. Theory Comput.* **2024**. https://doi.org/10.1021/acs.jctc.3c01318.
(65) Sabanés Zariquiey, F.; de Souza, J. V.; Bronowska, A. K. Cosolvent Analysis Toolkit (CAT): A Robust Hotspot Identification Platform for Cosolvent Simulations of Proteins to Expand

the Druggable Proteome. *Sci Rep* **2019**, *9* (1), 19118. https://doi.org/10.1038/s41598-019-55394-2.

(66) Raman, E. P.; Yu, W.; Guvench, O.; MacKerell, A. D. Jr. Reproducing Crystal Binding Modes of Ligand Functional Groups Using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations. *J. Chem. Inf. Model.* **2011**, *51* (4), 877–896. https://doi.org/10.1021/ci100462t.

(67) Raman, E. P.; Yu, W.; Lakkaraju, S. K.; MacKerell, A. D. Jr. Inclusion of Multiple Fragment Types in the Site Identification by Ligand Competitive Saturation (SILCS) Approach. *J. Chem. Inf. Model.* **2013**, *53* (12), 3384–3398. https://doi.org/10.1021/ci4005628.

(68) Andreev, G.; Kovalenko, M.; Bozdaganyan, M. E.; Orekhov, P. S. Colabind: A Cloud-Based Approach for Prediction of Binding Sites Using Coarse-Grained Simulations with Molecular Probes. *J. Phys. Chem. B* **2024**, *128* (13), 3211–3219. https://doi.org/10.1021/acs.jpcb.3c07853.

(69) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(70) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447. https://doi.org/10.1021/ct700301q.

(71) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8* (5), 1542–1555. https://doi.org/10.1021/ct200909j.

(72) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9* (1), 461–469. https://doi.org/10.1021/ct300857j.

(73) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix−Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113* (26), 9004–9015. https://doi.org/10.1021/jp901540t.

(74) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone φ, ψ and Side-Chain X1 and X2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273. https://doi.org/10.1021/ct300400x.

(75) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat Methods* **2017**, *14* (1), 71–73. https://doi.org/10.1038/nmeth.4067.

(76) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proceedings of the National Academy of Sciences* **2018**, *115* (21), E4758–E4766. https://doi.org/10.1073/pnas.1800690115.

(77) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone

Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16* (1), 528–552. https://doi.org/10.1021/acs.jctc.9b00591.
(78) Lakkaraju, S. K.; Raman, E. P.; Yu, W.; MacKerell, A. D. 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**, *10* (6), 2281–2290. https://doi.org/10.1021/ct500201y.
(79) Zhao, M.; Kognole, A. A.; Jo, S.; Tao, A.; Hazel, A.; MacKerell Jr, A. D. GPU-Specific Algorithms for Improved Solute Sampling in Grand Canonical Monte Carlo Simulations. *Journal of Computational Chemistry* **2023**, *44* (20), 1719–1732. https://doi.org/10.1002/jcc.27121.
(80) Ustach, V. D.; Lakkaraju, S. K.; Jo, S.; Yu, W.; Jiang, W.; MacKerell, A. D. Optimization and Evaluation of Site-Identification by Ligand Competitive Saturation (SILCS) as a Tool for Target-Based Ligand Optimization. *J. Chem. Inf. Model.* **2019**, *59* (6), 3018–3035. https://doi.org/10.1021/acs.jcim.9b00210.
(81) Goel, H.; Hazel, A.; Ustach, V. D.; Jo, S.; Yu, W.; MacKerell, A. D. Rapid and Accurate Estimation of Protein–Ligand Relative Binding Affinities Using Site-Identification by Ligand Competitive Saturation. *Chem. Sci.* **2021**, *12* (25), 8844–8858. https://doi.org/10.1039/D1SC01781K.
(82) MacKerell, A. D.; Jo, S.; Lakkaraju, S. K.; 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**, *1864* (4), 129519. https://doi.org/10.1016/j.bbagen.2020.129519.
(83) Kognole, A. A.; Hazel, A.; MacKerell, A. D. SILCS-RNA: Toward a Structure-Based Drug Design Approach for Targeting RNAs with Small Molecules. *J Chem Theory Comput* **2022**, *18* (9), 5672–5691. https://doi.org/10.1021/acs.jctc.2c00381.
(84) Weisel, M.; Proschak, E.; Kriegl, J. M.; Schneider, G. Form Follows Function: Shape Analysis of Protein Cavities for Receptor-Based Drug Design. *PROTEOMICS* **2009**, *9* (2), 451–459. https://doi.org/10.1002/pmic.200800092.
(85) 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), 1884–1897. https://doi.org/10.1002/pro.5560070905.
(86) Johnson, D. K.; Karanicolas, J. Druggable Protein Interaction Sites Are More Predisposed to Surface Pocket Formation than the Rest of the Protein Surface. *PLOS Computational Biology* **2013**, *9* (3), e1002951. https://doi.org/10.1371/journal.pcbi.1002951.
(87) Lomize, M. A.; Pogozheva, I. D.; Joo, H.; Mosberg, H. I.; Lomize, A. L. OPM Database and PPM Web Server: Resources for Positioning of Proteins in Membranes. *Nucleic Acids Research* **2012**, *40* (D1), D370–D376. https://doi.org/10.1093/nar/gkr703.
(88) Lomize, A. L.; Todd, S. C.; Pogozheva, I. D. Spatial Arrangement of Proteins in Planar and Curved Membranes by PPM 3.0. *Protein Science* **2022**, *31* (1), 209–220. https://doi.org/10.1002/pro.4219.
(89) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *Journal of Computational Chemistry* **2008**, *29* (11), 1859–1865. https://doi.org/10.1002/jcc.20945.

(90) Wu, E. L.; Cheng, X.; Jo, S.; Rui, H.; Song, K. C.; Dávila-Contreras, E. M.; Qi, Y.; Lee, J.; Monje-Galvan, V.; Venable, R. M.; Klauda, J. B.; Im, W. CHARMM-GUI Membrane Builder toward Realistic Biological Membrane Simulations. *Journal of Computational Chemistry* **2014**, *35* (27), 1997–2004. https://doi.org/10.1002/jcc.23702.
(91) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K_a Predictions. *J. Chem. Theory Comput.* **2011**, *7* (2), 525–537. https://doi.org/10.1021/ct100578z.
(92) SilcsBio, LLC. *SILCS: Site Identification by Ligand Competitive Saturation — SilcsBio User Guide*. https://docs.silcsbio.com/ (accessed 2024-02-21).
(93) Taylor, R. D.; MacCoss, M.; Lawson, A. D. G. Rings in Drugs. *J. Med. Chem.* **2014**, *57* (14), 5845–5859. https://doi.org/10.1021/jm4017625.
(94) Zhao, M.; Yu, W.; MacKerell, A. D. Jr. Enhancing SILCS-MC via GPU Acceleration and Ligand Conformational Optimization with Genetic and Parallel Tempering Algorithms. *J. Phys. Chem. B* **2024**, *128* (30), 7362–7375. https://doi.org/10.1021/acs.jpcb.4c03045.
(95) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A Comprehensive Resource for "omics" Research on Drugs. *Nucleic Acids Res* **2011**, *39* (Database issue), D1035-1041. https://doi.org/10.1093/nar/gkq1126.
(96) Research, C. for D. E. and. Drugs@FDA Data Files. *FDA* **2024**.
(97) RDKit: Open-Source Cheminformatics. https://www.rdkit.org.
(98) Xiong, G.; Shen, C.; Yang, Z.; Jiang, D.; Liu, S.; Lu, A.; Chen, X.; Hou, T.; Cao, D. Featurization Strategies for Protein–Ligand Interactions and Their Applications in Scoring Function Development. *WIREs Computational Molecular Science* **2022**, *12* (2), e1567. https://doi.org/10.1002/wcms.1567.
(99) Zhang, Y.; Li, S.; Meng, K.; Sun, S. Machine Learning for Sequence and Structure-Based Protein–Ligand Interaction Prediction. *J. Chem. Inf. Model.* **2024**, *64* (5), 1456–1472. https://doi.org/10.1021/acs.jcim.3c01841.
(100) Mitternacht, S. FreeSASA: An Open Source C Library for Solvent Accessible Surface Area Calculations. F1000Research February 18, 2016. https://doi.org/10.12688/f1000research.7931.1.
(101) Lam, S. K.; Pitrou, A.; Seibert, S. Numba: A LLVM-Based Python JIT Compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*; LLVM '15; Association for Computing Machinery: New York, NY, USA, 2015; pp 1–6. https://doi.org/10.1145/2833157.2833162.
(102) Baumli, S.; Endicott, J. A.; Johnson, L. N. Halogen Bonds Form the Basis for Selective P-TEFb Inhibition by DRB. *Chemistry & Biology* **2010**, *17* (9), 931–936. https://doi.org/10.1016/j.chembiol.2010.07.012.
(103) Wu, S. Y.; McNae, I.; Kontopidis, G.; McClue, S. J.; McInnes, C.; Stewart, K. J.; Wang, S.; Zheleva, D. I.; Marriage, H.; Lane, D. P.; Taylor, P.; Fischer, P. M.; Walkinshaw, M. D. Discovery of a Novel Family of CDK Inhibitors with the Program LIDAEUS: Structural Basis for Ligand-Induced Disordering of the Activation Loop. *Structure* **2003**, *11* (4), 399–410. https://doi.org/10.1016/S0969-2126(03)00060-1.
(104) 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**, *288* (12), 8596–8609. https://doi.org/10.1074/jbc.M113.452235.

(105) Wiesmann, C.; Barr, K. J.; Kung, J.; Zhu, J.; Erlanson, D. A.; Shen, W.; Fahr, B. J.; Zhong, M.; Taylor, L.; Randal, M.; McDowell, R. S.; Hansen, S. K. Allosteric Inhibition of Protein Tyrosine Phosphatase 1B. *Nat Struct Mol Biol* **2004**, *11* (8), 730–737. https://doi.org/10.1038/nsmb803.

(106) Han, Y.; Belley, M.; Bayly, C. I.; Colucci, J.; Dufresne, C.; Giroux, A.; Lau, C. K.; Leblanc, Y.; McKay, D.; Therien, M.; Wilson, M.-C.; Skorey, K.; Chan, C.-C.; Scapin, G.; Kennedy, B. P. 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**, *18* (11), 3200–3205. https://doi.org/10.1016/j.bmcl.2008.04.064.

(107) Montalibet, J.; Skorey, K.; McKay, D.; Scapin, G.; Asante-Appiah, E.; Kennedy, B. P. Residues Distant from the Active Site Influence Protein-Tyrosine Phosphatase 1B Inhibitor Binding*. *Journal of Biological Chemistry* **2006**, *281* (8), 5258–5266. https://doi.org/10.1074/jbc.M511546200.

(108) Wan, Z.-K.; Follows, B.; Kirincich, S.; Wilson, D.; Binnun, E.; Xu, W.; Joseph-McCarthy, D.; Wu, J.; Smith, M.; Zhang, Y.-L.; Tam, M.; Erbe, D.; Tam, S.; Saiah, E.; Lee, J. Probing Acid Replacements of Thiophene PTP1B Inhibitors. *Bioorganic & Medicinal Chemistry Letters* **2007**, *17* (10), 2913–2920. https://doi.org/10.1016/j.bmcl.2007.02.043.

(109) Pereira de Jésus-Tran, K.; Côté, P.-L.; 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), 987–999. https://doi.org/10.1110/ps.051905906.

(110) Estébanez-Perpiñá, E.; Arnold, L. A.; Nguyen, P.; Rodrigues, E. D.; Mar, E.; Bateman, R.; Pallai, P.; Shokat, K. M.; Baxter, J. D.; Guy, R. K.; Webb, P.; Fletterick, R. J. A Surface on the Androgen Receptor That Allosterically Regulates Coactivator Binding. *Proceedings of the National Academy of Sciences* **2007**, *104* (41), 16074–16079. https://doi.org/10.1073/pnas.0708036104.

(111) Srivastava, A.; Yano, J.; Hirozane, Y.; Kefala, G.; Gruswitz, F.; Snell, G.; Lane, W.; Ivetac, A.; Aertgeerts, K.; Nguyen, J.; Jennings, A.; Okada, K. High-Resolution Structure of the Human GPR40 Receptor Bound to Allosteric Agonist TAK-875. *Nature* **2014**, *513* (7516), 124–127. https://doi.org/10.1038/nature13494.

(112) Ho, J. D.; Chau, B.; Rodgers, L.; Lu, F.; Wilbur, K. L.; Otto, K. A.; Chen, Y.; Song, M.; Riley, J. P.; Yang, H.-C.; Reynolds, N. A.; Kahl, S. D.; Lewis, A. P.; Groshong, C.; Madsen, R. E.; Conners, K.; Lineswala, J. P.; Gheyi, T.; Saflor, M.-B. D.; Lee, M. R.; Benach, J.; Baker, K. A.; Montrose-Rafizadeh, C.; Genin, M. J.; Miller, A. R.; Hamdouchi, C. Structural Basis for GPR40 Allosteric Agonism and Incretin Stimulation. *Nat Commun* **2018**, *9* (1), 1645. https://doi.org/10.1038/s41467-017-01240-w.

(113) Haga, K.; Kruse, A. C.; Asada, H.; Yurugi-Kobayashi, T.; Shiroishi, M.; Zhang, C.; Weis, W. I.; Okada, T.; Kobilka, B. K.; Haga, T.; Kobayashi, T. Structure of the Human M2 Muscarinic Acetylcholine Receptor Bound to an Antagonist. *Nature* **2012**, *482* (7386), 547–551. https://doi.org/10.1038/nature10753.

(114) Kruse, A. C.; Ring, A. M.; Manglik, A.; Hu, J.; Hu, K.; Eitel, K.; Hübner, H.; Pardon, E.; Valant, C.; Sexton, P. M.; Christopoulos, A.; Felder, C. C.; Gmeiner, P.; Steyaert, J.; Weis, W. I.;

Garcia, K. C.; Wess, J.; Kobilka, B. K. Activation and Allosteric Modulation of a Muscarinic Acetylcholine Receptor. *Nature* **2013**, *504* (7478), 101–106. https://doi.org/10.1038/nature12735.

(115) Rasmussen, S. G. F.; DeVree, B. T.; Zou, Y.; Kruse, A. C.; Chung, K. Y.; Kobilka, T. S.; Thian, F. S.; Chae, P. S.; Pardon, E.; Calinski, D.; Mathiesen, J. M.; Shah, S. T. A.; Lyons, J. A.; Caffrey, M.; Gellman, S. H.; Steyaert, J.; Skiniotis, G.; Weis, W. I.; Sunahara, R. K.; Kobilka, B. K. Crystal Structure of the B2 Adrenergic Receptor–Gs Protein Complex. *Nature* **2011**, *477* (7366), 549–555. https://doi.org/10.1038/nature10361.

(116) Liu, X.; Ahn, S.; Kahsai, A. W.; Meng, K.-C.; Latorraca, N. R.; Pani, B.; Venkatakrishnan, A. J.; Masoudi, A.; Weis, W. I.; Dror, R. O.; Chen, X.; Lefkowitz, R. J.; Kobilka, B. K. Mechanism of Intracellular Allosteric β2AR Antagonist Revealed by X-Ray Crystal Structure. *Nature* **2017**, *548* (7668), 480–484. https://doi.org/10.1038/nature23652.

(117) Goldstein, D. M.; Soth, M.; Gabriel, T.; Dewdney, N.; Kuglstatter, A.; Arzeno, H.; Chen, J.; Bingenheimer, W.; Dalrymple, S. A.; Dunn, J.; Farrell, R.; Frauchiger, S.; La Fargue, J.; Ghate, M.; Graves, B.; Hill, R. J.; Li, F.; Litman, R.; Loe, B.; McIntosh, J.; McWeeney, D.; Papp, E.; Park, J.; Reese, H. F.; Roberts, R. T.; Rotstein, D.; San Pablo, B.; Sarma, K.; Stahl, M.; Sung, M.-L.; Suttman, R. T.; Sjogren, E. B.; Tan, Y.; Trejo, A.; Welch, M.; Weller, P.; Wong, B. R.; Zecic, H. 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**, *54* (7), 2255–2265. https://doi.org/10.1021/jm101423y.

(118) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of P38 MAP Kinase by Utilizing a Novel Allosteric Binding Site. *Nat Struct Mol Biol* **2002**, *9* (4), 268–272. https://doi.org/10.1038/nsb770.

(119) Drug Design Data Resource (D3R). Drug Design Data Resource Grand Challenge 2 Dataset: FXR - Farnesoid X Receptor, 2017, 71.5MB. https://doi.org/10.15782/D6RP4P.

(120) Cumming, J. N.; Smith, E. M.; Wang, L.; Misiaszek, J.; Durkin, J.; Pan, J.; Iserloh, U.; Wu, Y.; Zhu, Z.; Strickland, C.; Voigt, J.; Chen, X.; Kennedy, M. E.; Kuvelkar, R.; Hyde, L. A.; Cox, K.; Favreau, L.; Czarniecki, M. F.; Greenlee, W. J.; McKittrick, B. A.; Parker, E. M.; Stamford, A. W. Structure Based Design of Iminohydantoin BACE1 Inhibitors: Identification of an Orally Available, Centrally Active BACE1 Inhibitor. *Bioorganic & Medicinal Chemistry Letters* **2012**, *22* (7), 2444–2449. https://doi.org/10.1016/j.bmcl.2012.02.013.

(121) *D3R | Drug Design Data Resource Grand Challenge 4 Dataset: BACE1*. https://drugdesigndata.org/about/datasets/2027 (accessed 2024-02-19).

(122) *D3R | Drug Design Data Resource Grand Challenge Dataset: GSK TrmD*. https://drugdesigndata.org/about/datasets/226 (accessed 2024-02-19).

(123) Friberg, A.; Vigil, D.; Zhao, B.; Daniels, R. N.; Burke, J. P.; Garcia-Barrantes, P. M.; Camper, D.; Chauder, B. A.; Lee, T.; Olejniczak, E. T.; Fesik, S. W. Discovery of Potent Myeloid Cell Leukemia 1 (Mcl-1) Inhibitors Using Fragment-Based Methods and Structure-Based Design. *J. Med. Chem.* **2013**, *56* (1), 15–30. https://doi.org/10.1021/jm301448p.

(124) Sato, M.; Arakawa, T.; Nam, Y.-W.; 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**, *1854* (5), 333–340. https://doi.org/10.1016/j.bbapap.2015.01.011.
(125) 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**, *397* (4), 1042–1054. https://doi.org/10.1016/j.jmb.2010.02.007.
(126) Tarver, C. L. Molecular Role of Angiopoietin-like 4's Carboxy-Terminal Domain in Pancreatic Ductal Adenocarcinoma Progression. Dissertations, University of Huntsville Alabama, 2019.
(127) Wang, X.; Minasov, G.; Shoichet, B. K. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-Offs. *Journal of Molecular Biology* **2002**, *320* (1), 85–95. https://doi.org/10.1016/S0022-2836(02)00400-X.
(128) Horn, J. R.; Shoichet, B. K. Allosteric Inhibition Through Core Disruption. *Journal of Molecular Biology* **2004**, *336* (5), 1283–1291. https://doi.org/10.1016/j.jmb.2003.12.068.
(129) Ness, S.; Martin, R.; Kindler, A. M.; Paetzel, M.; Gold, M.; Jensen, S. E.; Jones, J. B.; Strynadka, N. C. J. Structure-Based Design Guides the Improved Efficacy of Deacylation Transition State Analogue Inhibitors of TEM-1 β-Lactamase. *Biochemistry* **2000**, *39* (18), 5312–5321. https://doi.org/10.1021/bi992505b.
(130) Li, P.; Morris, D. L.; Willcox, B. E.; Steinle, A.; Spies, T.; Strong, R. K. Complex Structure of the Activating Immunoreceptor NKG2D and Its MHC Class I–like Ligand MICA. *Nat Immunol* **2001**, *2* (5), 443–451. https://doi.org/10.1038/87757.
(131) Thompson, A. A.; Harbut, M. B.; Kung, P.-P.; Karpowich, N. K.; Branson, J. D.; Grant, J. C.; Hagan, D.; Pascual, H. A.; Bai, G.; Zavareh, R. B.; Coate, H. R.; Collins, B. C.; Côte, M.; Gelin, C. F.; Damm-Ganamet, K. L.; Gholami, H.; Huff, A. R.; Limon, L.; Lumb, K. J.; Mak, P. A.; Nakafuku, K. M.; Price, E. V.; Shih, A. Y.; Tootoonchi, M.; Vellore, N. A.; Wang, J.; Wei, N.; Ziff, J.; Berger, S. B.; Edwards, J. P.; Gardet, A.; Sun, S.; Towne, J. E.; Venable, J. D.; Shi, Z.; Venkatesan, H.; Rives, M.-L.; Sharma, S.; Shireman, B. T.; Allen, S. J. Identification of Small-Molecule Protein–Protein Interaction Inhibitors for NKG2D. *Proceedings of the National Academy of Sciences* **2023**, *120* (18), e2216342120. https://doi.org/10.1073/pnas.2216342120.
(132) Kim, Y.; Jeong, E.; Jeong, J.-H.; Kim, Y.; Cho, Y. Structural Basis for Activation of the Heterodimeric $GABA_B$ Receptor. *Journal of Molecular Biology* **2020**, *432* (22), 5966–5984. https://doi.org/10.1016/j.jmb.2020.09.023.
(133) Shaye, H.; Ishchenko, A.; Lam, J. H.; Han, G. W.; Xue, L.; Rondard, P.; Pin, J.-P.; Katritch, V.; Gati, C.; Cherezov, V. Structural Basis of the Activation of a Metabotropic GABA Receptor. *Nature* **2020**, *584* (7820), 298–303. https://doi.org/10.1038/s41586-020-2408-4.
(134) Mao, C.; Shen, C.; Li, C.; Shen, D.-D.; Xu, C.; Zhang, S.; Zhou, R.; Shen, Q.; Chen, L.-N.; Jiang, Z.; Liu, J.; Zhang, Y. Cryo-EM Structures of Inactive and Active $GABA_B$ Receptor. *Cell Res* **2020**, *30* (7), 564–573. https://doi.org/10.1038/s41422-020-0350-5.
(135) *D3R | Drug Design Data Resource*. https://drugdesigndata.org/ (accessed 2024-02-19).

(136) Ge, Y.; Pande, V.; Seierstad, M. J.; Damm-Ganamet, K. L. Exploring the Application of SiteMap and Site Finder for Focused Cryptic Pocket Identification. *J. Phys. Chem. B* **2024**, *128* (26), 6233–6245. https://doi.org/10.1021/acs.jpcb.4c00664.
(137) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.
(138) Guyon, I.; Weston, J.; Barnhill, S. Gene Selection for Cancer Classification Using Support Vector Machines. 34.
(139) Sklearn Documentation for SVC. https://scikit-learn/stable/modules/generated/sklearn.svm.SVC.html (accessed 2024-04-01).
(140) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer New York, NY, 2009.
(141) The pandas development team. Pandas-Dev/Pandas: Pandas, 2023. https://doi.org/10.5281/zenodo.7741580.
(142) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.
(143) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
(144) Petroff, M. A. Accessible Color Sequences for Data Visualization. arXiv February 28, 2024. https://doi.org/10.48550/arXiv.2107.02270.

Table of Contents Figure: