Title: **CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites**

Authors: Peter Cimermancic[a,b,*], Patrick Weinkam[a], T. Justin Rettenmaier[c,d], Leon Bichmann[a], Daniel A. Keedy[a], Rahel A. Woldeyes[a,c], Dina Schneidmann[a], Omar N. A. Demerdash[f], Julie C. Mitchell[g], James A. Wells[d,e], James S. Fraser[a], Andrej Sali[a,d,*]

Affiliations:

[a] Departments of Bioengineering and Therapeutic Sciences,

[b] Graduate Group in Biological and Medical Informatics,

[c] Graduate Group in Chemistry and Chemical Biology,

[d] Pharmaceutical Chemistry, and

[e] Cellular and Molecular Pharmacology, and California Institute for Quantitative Biosciences (QB3), University of California, San Francisco, San Francisco, CA 94158, USA.

[f] Department of Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, CA 94720, USA.

[g] Departments of Biochemistry and Mathematics, University of Wisconsin – Madison,

*Corresponding authors:

1700 4th Street, Byers Hall 501A, University of California, San Francisco, San Francisco, CA 94158; tel 415-696-1455; peter.cimermancic@ucsf.edu.

1700 4th Street, Byers Hall 503B, University of California, San Francisco, San Francisco, CA 94158; tel 415-514-4227; web http://salilab.org; sali@salilab.org.

Abstract

Many proteins have small molecule-binding pockets that are not easily detectable in the ligand-free structures. These cryptic sites require a conformational change to become apparent; a cryptic site can therefore be defined as a site that forms a pocket in a *holo* structure, but not in the *apo* structure. Because many proteins appear to lack druggable pockets, understanding and accurately identifying cryptic sites could expand the set of drug targets. Previously, cryptic sites were identified experimentally by fragment-based ligand discovery, and computationally by long molecular dynamics simulations and fragment docking. Here, we begin by constructing a set of structurally defined *apo-holo* pairs with cryptic sites. Next, we comprehensively characterize the cryptic sites in terms of their sequence, structure, and dynamics attributes. We find that cryptic sites tend to be as conserved in evolution as traditional binding pockets, but are less hydrophobic and more flexible. Relying on this characterization, we use machine learning to predict cryptic sites with relatively high accuracy (for our benchmark, the true positive and false positive rates are 73% and 29%, respectively). We then predict cryptic sites in the entire structurally characterized human proteome (11,201 structures, covering 23% of all residues in the proteome). CryptoSite increases the size of the potentially "druggable" human proteome from ~40% to ~78% of disease-associated proteins. Finally, to demonstrate the utility of our approach in practice, we experimentally validate a cryptic site in protein tyrosine phosphatase 1B using a covalent ligand and NMR spectroscopy. The CryptoSite web server is available at http://salilab.org/cryptosite.

Keywords: cryptic binding sites, protein dynamics, undruggable proteins, machine learning

Highlights:

- *Bona fide* cryptic sites identified by comparison of *apo* and *holo* protein structures.
- Features distinguishing cryptic sites and binding pockets identified.
- Efficient and accurate prediction of cryptic sites developed.
- Cryptic sites predicted for all human proteins of known structure.
- The "druggable" human proteome may be larger than previously estimated.

Introduction

Biological function often involves binding of proteins to other molecules, including small ligands and macromolecules. Usually, these interactions occur at defined binding sites in the protein structure (1). Knowledge of binding site location has a number of applications (2). For example, in drug discovery, binding site localization is often the starting point followed by virtual screening or *de novo* ligand design (3); in cell biology, it facilitates prediction of protein substrates, especially when the target protein cannot be reliably related to homologs of known function (4).

Binding sites, particularly those for small molecules, are often located in exposed concave pockets, which provide an increased surface area that in turn maximizes intra-molecular interactions (5). A concave pocket can already exist in a ligand-free structure of a protein; such binding sites are called here binding pockets. Sometimes, however, a binding site is flat in the absence of a ligand and only forms in the presence of a ligand (*ie*, induced fit) or only opens transiently for short periods of time (*ie*, conformational selection); such binding sites are called cryptic sites (**Fig. 1*A***) (6-11).

Many computational methods have been developed to localize binding pockets on proteins. These methods are based on a variety of principles (12): (*i*) concavity of the protein surface, (*ii*) energy functions including van der Waals terms, (*iii*) geometrical and physico-chemical similarity to known binding pockets, and (*iv*) composite approaches that use a combination of different features (13-15). Unfortunately, only ~60% of protein structures were judged to have pockets larger than 250 $Å^3$ (many of which may not be druggable), and could potentially be subjected to ligand discovery based on binding pocket knowledge (16, 17).

In contrast to binding pockets, cryptic sites are not easily detectable in a ligand-free structure of a protein because they by definition require ligand-induced conformational changes to become apparent. For example, large and flat interfaces between interacting proteins were considered undruggable, although several examples of protein interfaces undergoing a conformational change coupled with binding a small molecule were recently described (18, 19). Similarly, allosterically regulated sites are sometimes not apparent in the absence of a small-molecule allosteric regulator (e.g., p38 MAP kinase (7) and TEM1 β-lactamase (9)).

Currently, the only approaches to cryptic site discovery are exhaustive site-directed small-molecule tethering by experiment (20-22), long time-scale molecular dynamics simulations by computation (6, 8, 23, 24), flexible docking (25, 26), and computational tools for identification of small-molecule binding hot spots (10, 27-30). All of these approaches are time-consuming, expensive, and/or not always successful. Therefore, there is a need for an accurate, automated, and efficient method to predict the location of cryptic pockets in a given ligand-free protein

structure. Such a method would offer several advantages. First, a cryptic site may be the only suitable binding site on the target protein; for example, when activation is required and thus the active site cannot be targeted, the active site is not druggable, or active site ligands need to be avoided due to adverse off-target effects. Second, binding sites may be discovered on structures determined or computed at only moderate resolution.

Here, we analyze known cryptic sites and develop a method for predicting cryptic site locations to address a number of questions: What are the sequence, structure, and dynamics attributes of a cryptic site, especially in comparison to binding pockets? Can we accurately, automatically, and efficiently predict cryptic sites? How common are cryptic sites? Are they common enough to significantly expand the druggable proteome? Can we predict cryptic sites in specific proteins of clinical significance?

Results and Discussion

Our method development and analysis proceeded according to **Figure 1*B***. In outline, we started by creating a representative dataset of 84 known examples of cryptic binding sites, 92 binding pockets, and 705 concave surface patches from the Protein Data Bank (31) and the MOAD database (32) (**Methods**, **SI Text**, and **Table S1**). We selected cryptic sites and binding pockets whose ligands are biologically relevant (32). Next, we designed a set of 30 features that describe sequence, structure, and dynamics of individual residues and their neighbors (**SI Text** and **Table S2**), based on the crystal structures (15, 33, 34). We then compared these attributes between the three types of a site to better understand the underlying characteristics of each site. Based on these comparisons, we expanded the set of features for proteins containing cryptic sites to 58 (**Table S2**), describing their crystal structures as well as their alternative conformations obtained by molecular dynamics simulations using AllosMod (35) (**SI Text**). Next, we put to test 11 supervised machine-learning algorithms (36, 37) to classify residues as belonging to a cryptic site or not; the accuracy of the best predictive model was assessed using leave-one-out cross-validation on a training set as well as using an independent test set. We then predicted cryptic sites in the entire structurally characterized human proteome. Finally, we focused on a detailed characterization of protein tyrosine phosphatase 1B (PTP1B), a protein that is involved in the insulin signaling pathway and is considered a validated therapeutic target for treatment of type 2 diabetes (38).

Pocket formation at a cryptic site is driven by small changes in the structure, resulting in a conformationally conserved cryptic site regardless of the ligand type. First, we set out to analyze structural changes needed for a binding pocket formation at a cryptic site. The dataset of cryptic sites reveals mostly minor structural changes required for formation of a detectable

pocket. The all-atom RMSD of cryptic binding sites between *apo* and *holo* conformations ranges between 0.45 Å and 22.45 Å (**Fig. S1*A***) with 67% *apo-holo* pairs differing less than 3 Å in RMSD. The only two *apo-holo* pairs whose differences in RMSD exceed 10 Å are calcium ATPase and calmodulin (PDB IDs: 1SU4–3FGO and 1CLL–1CTR, respectively). Loop movement is the most prominent type of conformational changes (observed in 45% of the binding sites), followed by side-chain rotation (18%), domain motion (17%), displacement of secondary structure elements (16%), and N- or C-terminus flexibility (4%).

To determine whether or not a cryptic site assumes the same bound conformation irrespective of the ligand type, we computed similarities between cryptic site conformations in a protein bound to at least 5 different ligands (58 proteins). Interestingly, only 26% of such cases have an average RMSD exceeding 2 Å (**Fig. S1*B***), even though the average Tanimoto distance (calculated by Open Babel (39), **SI Text**) is low (0.8). This finding suggests that the conformation of a given cryptic site generally does not depend strongly on the ligand type (similar analysis of binding pockets yields 9% of cases with an average RMSD exceeding 2 Å, and an average Tanimoto distance of 0.7). Moreover, the magnitude of the conformational difference within a group of *holo* structures is not significantly correlated with ligand similarity (the correlation coefficient between the all-atom binding site RMSD and Tanimoto distance is 0.01; **Figs. S1*C*** and **S1*D***). Finally, the average RMSD of 1.7 Å between bound cryptic binding sites is significantly lower than the average RMSD of 3.0 Å between the unbound and bound conformations ($P = 1.4\times10^{-3}$, based on two-sample Kolmogorov-Smirnov statistics). Thus, the bound form of the cryptic site is surprisingly conformationally conserved with respect to the ligand type (the average RMSD values of bound conformations of cryptic sites and binding pockets are 1.7 and 2.0 Å, respectively). These observations are consistent with a limited number of protein conformational states as well as with the variability in allosterically regulated proteins, where the binding of the effector alters the conformational distribution between two or more conformational states (40). Indeed, 24 of the 58 cryptic sites are found in proteins that are known to be allosterically regulated, with 17 of the 24 annotated as effector binding sites (41). 20 of the remaining 34 cryptic sites are found on proteins with two or more different binding sites that may or may not be allosteric. The remaining 14 cryptic sites occur on enzymes with flexible active sites and receptors for large hydrophobic ligands, where cryptic site residues modulate binding site accessibility (*e.g.*, the “portal” hypothesis for glycolipid transfer protein, lactoglobulin, and adipocyte lipid binding protein) (42). In other words, a cryptic site does not convert from flat to concave to accommodate a number of different ligands; rather, cryptic sites may have evolved the ability to convert from flat to concave to modulate ligand-binding kinetics, specificity, affinity, and allostery.

Cryptic sites are as flexible as random concave surface patches, but evolutionarily as conserved as binding pockets. Next, we analyzed the differences between the sequence, structure, and dynamics attributes of cryptic sites, binding pockets, and concave surface patches. While the differences between cryptic sites and binding pockets are generally small, 4 characteristics distinguish a cryptic site from a binding pocket and/or a concave surface patch: First, a cryptic site predominantly localizes at concave protein regions, even though the site itself is not as concave in the unbound form as a binding pocket. For example, while the average number of protruding atoms at a cryptic site and a binding pocket is 170 and 183 ($P = 8.0\times10^{-3}$) and the average convexity value is 2.4 and 1.9 ($P = 0.8$), the average pocket score is 0.07 and 0.42 ($P = 1.7\times10^{-31}$), respectively (**Table S3**). Second, a cryptic site tends to be less hydrophobic than a binding pocket, due mostly to an increased frequency of charged residues (arginine in particular, $P = 1.8\times10^{-5}$) (**Fig. S2*A* and Table S3**). Third, a cryptic site is more flexible than a binding pocket, as indicated by significantly higher normalized B-factors (**Fig. S2*B***). Finally, cryptic site residues are evolutionarily as conserved as those of a binding pocket (**Fig. S2*C***), suggesting a similar degree of evolutionary pressure and selection on the function of many of these two types of binding sites. Evolutionarily conserved residues have been previously associated with low B-factors (43-45); low B-factors are an indicator of residue rigidity. Both evolutionarily conserved residues and residues with low B-factors are often found in functionally important regions of a protein, including binding pockets (13, 46). In contrast to binding pockets, cryptic sites conserve conformational flexibility to convert from flat to concave. We found no statistically significant differences between properties of ligands of cryptic sites and binding pockets (**Fig. S3**).

Molecular dynamics simulations based on a simplified energy landscape, sequence conservation, and fragment docking are sufficient to predict cryptic sites. To test if cryptic sites could be predicted accurately, automatically, and efficiently, we used the dataset of *apo* structures with cryptic sites to train 10 different machine-predictive models for the prediction of cryptic site residues, based on the extended set of 58 features (**Table S2**); the datasets with binding pockets and concave surface patches were not used as training sets for machine learning. The optimal predictive model and its parameter values were selected by maximizing the sensitivity (true positive rate) and the specificity (true negative rate) of cryptic site residue prediction, using leave-one-out cross validation on the training set of proteins with 84 cryptic binding sites (**SI Text, Fig. S4*A***). The optimal predictive model is a support vector machine (SVM) with a quadratic kernel function. By removing redundant and irrelevant features using greedy-forward selection that maximizes the AUC and by testing the statistical significance of the improvement in the prediction accuracy, we selected 3 features, resulting in the AUC of 0.77 (**Fig. S4*B-D***).

Although an SVM operates as a "black-box", the relative importance of different features can be inferred from the order of selection, and may be informative about the cryptic site characteristics (47). We find the average pocket score from the molecular dynamics simulations is the most informative single feature according to greedy-forward selection (AUC = 0.73) as well as the two-sample Kolmogorov-Smirnov test ($P = 4.3\times10^{-138}$) (**Fig. S2*D*** and **Table S2**). This feature alone is almost as informative as a subset of 30 crystal structure features combined (AUC = 0.74) (**Table S2**). Therefore, molecular dynamics simulations on a simplified energy landscape, which is significantly more computationally efficient than a traditional all-atom molecular dynamics simulation (6), often provides sufficient information for localizing cryptic sites. The second feature added to the subset of the 3 features by the greedy-forward approach was sequence conservation (AUC = 0.74). Cryptic site residues are significantly more conserved than the rest of a protein ($P = 3.4\times10^{-67}$). The third feature, likelihood of binding small-molecule fragments (**SI Text**), also significantly improves the accuracy of the model (AUC = 0.77). Despite the relatively small magnitude of the increase in the accuracy, the improvement of adding two additional features to the single most informative feature is statistically significant (**Figs. S4*C* and S4*D***). In summary, a cryptic site can be predicted relatively accurately based primarily on pocket formation in molecular dynamics simulations, evolutionary conservation, and likelihood of binding small-molecule fragments. Independent predictions based on different molecular dynamics trajectories are highly similar, with the cross-correlation coefficient larger than 0.9 and the average residue score difference of the most variable decile smaller than 0.04 (**Fig. S5*A***); the predicted scores vary the most for residues that reside on α-helices or β-sheets and are adjacent to flexible parts of a protein. Similarly, predictions for a subunit on its own or in the context of a biological assembly are also highly similar, except for the subunit-subunit interface residues (**Fig. S5*B***).

CryptoSite accurately localizes over 96% of cryptic binding sites, outperforming other computational methods. To assess the performance of our predictive model, we applied it to the training set using leave-one-out cross validation as well as to the test set of 14 *apo* structures with one or more known cryptic sites that were not used during the training or any of the analyses above. The prediction capability of the SVM model is satisfactory; we measure an overall AUC of 0.83, with respective true positive and false positive rates of 79% and 29% at the residue score threshold of 0.05 (**Fig. 2*A***). At higher score thresholds of 0.1 and 0.15, the respective true positive and false positive rates are 15% and 55%, and 6% and 28% (in other words, in an experimental test of a prediction, on average 7.6, 5.9, and 4.9 residues with the predicted residues score higher than 0.05, 0.1, and 0.15, respectively, would need to be tested to find at least one true cryptic site residue – a significant improvement over the need to test 19 randomly chosen residues for the same outcome). CryptoSite can also be applied to low-resolution atomic structures and comparative models, in addition to high-resolution X-ray structures, without a large

loss of accuracy. For example, the average cross-correlation between cryptic site predictions for a high-resolution X-ray structure and its comparative model based on a template with at least 50% sequence identity is approximately 0.7 (**SI Text, Fig. S6, Table S4**). To further dissect the performance of the learning algorithm, we evaluated predictions for individual proteins from our training and test sets (**Fig. 2*B***). We define a prediction of a cryptic site to be accurate when at least one third of its residues are identified (sensitivity > 33%). Predictions above this threshold can arguably guide small-molecule tethering experiments and more detailed molecular dynamics simulations. Remarkably, all 14 proteins in the test set and 75 out of 79 proteins in the cross-validation/training set have all of their cryptic sites identified accurately, resulting in 96% recall (**Tables S1** and **S5**); even for 50% sensitivity, the recall is still 88%. The predictions are particularly accurate when a large and hydrophobic ligand binds to a cryptic site. For example, we identified 98% of cryptic site residues in the acyl-CoA binding site of the fatty acid responsive transcription factor and 89% of cryptic site residues in the lipid-binding site of β-lactoglobulin (**Fig. S7**). Our predictive model also accurately predicted cryptic sites in 18 out of 20 proteins (including the proteins from the cross-validation set) that undergo domain movements to expose small-molecule binding sites. For example, more than half of the cryptic site residues of GluR2 receptor (100%), exportin 1 (68%), and biotin carboxylase were predicted correctly (56%) (**Figs. 2*C* and S7**).

Our predictive model also accurately predicts known allosteric cryptic sites in TEM-1 β-lactamase that are buried in the *apo* conformation (60%) (**Figs. 2*C* and S7*D***) and were previously studied using extensive molecular dynamics simulations in explicit solvent and Markov state models (6, 24). Moreover, both molecular dynamics simulations (23, 25, 48) and CryptoSite also successfully predicted known binding sites at difficult-to-drug protein-protein interaction interfaces, including in interleukin-2 (specificity of 79%), Bcl-X_L (73%), FK506-binding protein (FKBP12; 73%), HPV regulatory protein E2 (50%), and cell division protein ZipA (60%) (**Figs. 2*C*, S7*E*, and S8*C***). Finally, we used our testing set to benchmark CryptoSite against FTFlex (28, 49), a computational solvent mapping approach for prediction of small molecule-binding hot spots that takes into account side chain flexibility. CryptoSite is more accurate (**Figs. 2*A* and S8*A***), especially when a cryptic site is buried (TEM-1 β-lactamase and β-lactoglobulin) or resides in a large protein (exportin 1; 68%) (**Figs. 2 and S8**). In conclusion, CryptoSite is as accurate as approaches based on extensive molecular dynamics simulations, but significantly faster (a calculation on an average sized protein takes 1-2 days on our webserver) and completely automated. In comparison to approaches of similar efficiency (25, 28), CryptoSite is generally more accurate, particularly when the location of a cryptic site is buried in the *apo* state.

False negatives result from large rearrangements. Next, we analyze false negatives and false positives (defined based on the cryptic sites annotated in MOAD). Our predictive model failed to predict most cryptic sites that undergo large conformational changes and whose pockets are difficult to sample with current molecular dynamics approaches, and partial sites that require binding to another protein chain to become functional (**Fig. S9**). In particular, we failed at predicting the cryptic site for stabilizing substrates (*eg*, cyclopiazonic acid) in Ca-ATPase (sensitivity of 6%) that resides at the interface between three domains, two of which are ~50 Å apart in the *apo* conformation (**Fig. S9*A***). Similarly, we also failed at predicting an allosteric site in the thumb site of HCV RNA polymerase (sensitivity of 0%), a site between two chains of kynurenine aminotransferase II (sensitivity of 17%), and an allosteric site in PTP1B (sensitivity of 29%) (**Fig. S9**). In the future, inadequate sampling in AllosMod will be addressed by using multiple input structures and/or restraints from experimental data (*e.g.*, small-angle X-ray scattering profiles (50), chemical cross-links (51), hydrogen/deuterium exchange with mass spectrometry, and electron microscopy density maps (52)).

A false positive prediction can be an unknown cryptic site. While it is difficult to be certain that a predicted cryptic site does not bind a ligand, potential false positives include high-scoring isolated residues or terminal regions of truncated proteins, which may not be as flexible in full-length proteins. However, our benchmark probably overestimates the false positive rate, because some predicted cryptic sites are in fact true binding sites, even though they are not annotated as such in the MOAD database (*e.g.*, proteins that bind peptides or other proteins). For example, our predictive model identifies the binding site for the light chain of coagulation factor VII in the heavy chain of coagulation factor VII; the binding site for guanine-nucleotide exchange factor DBS in CDC42 protein; the dimer interfaces in fructose-1,6-bisphosphate aldolase and estrogen-related receptor γ; the docking site for its N-terminal motif in Bcl-X_L; and the phosphate binding site in acid-β-glucosidase **(Figs. 2*C*** and **S7)**. Excluding protein-protein interface residues from the prediction of cryptic sites may reduce the number of false positives; however, the improvement appears to be modest, case-dependent, and comes at a cost of ignoring cryptic sites that are located at such interfaces (**Fig. S5*C***). In summary, the analysis of successes and failures demonstrates the potential of our approach to guide the experimental identification of new sites in difficult small-molecule targets.

The druggable proteome is significantly larger than estimated previously. Given the overall accuracy of our approach (above), a large number of predicted cryptic sites that are not yet annotated as such in our benchmark might also indicate that there are many cryptic sites yet to be discovered. If so, our predictive model could facilitate finding novel binding sites in "undruggable" proteins, and hence expand the druggable proteome space. It has been suggested that the human proteome of approximately 20,000 proteins contains ~3,000 proteins associated

with disease and ~3,000 druggable proteins, with the overlap between the two sets of only ~600 – 1,500 (16, 53, 54). To predict how much cryptic binding sites expand the druggable proteome space, we first applied a faster version of our predictive model (based on a subset of features that are not extracted from molecular dynamics simulations, resulting in the speedup factor of 1000 and AUC of 0.74) on 4,421 human proteins with at least one domain of known structure (11,201 structures in total). Next, we counted the numbers of cryptic sites and pockets in each structure (**SI Text**). Pockets were predicted in ~1,900 (43%) proteins, and cryptic sites were predicted in ~3,300 (74%) proteins. Among the 1,420 disease-associated proteins of known structure, 40% have pockets in their crystal structures (in agreement with the previous estimate that the fraction of proteins that are both disease-associated and druggable is 20-50% (55)). In contrast to pockets, cryptic sites were predicted in 72% of the disease-associated proteins, 38% of which have no apparent pockets (**Fig. 3*A***). However, some of the predictions may be false positives (the sites may in fact not bind any ligands). Moreover, for some sites, it may be very difficult to find a ligand (even if it does exist), and even if the ligand is found, it may not be a drug because it does not target the disease-modifying function of a protein or because it does not meet clinical development criteria. Nevertheless, the prediction of cryptic sites on the disease-associated proteins of known structure indicates that small molecules might be used to target significantly more disease-associated proteins than were previously thought druggable.

If cryptic sites are more abundant than previously estimated, why does high-throughput screening not identify them more often than it does? It has been shown that small-molecule libraries are biased towards traditional drug targets, such as G protein-coupled receptors, ion channels, and kinases, while they are not as suitable for antimicrobial targets and those identified from genomic studies (56). It is conceivable that the existing libraries are also less suitable for cryptic sites. Moreover, cryptic sites may tend to bind ligands more weakly than binding pockets, due to the need to compensate for the free energy of site formation (57), and may thus be ranked lower on the high-throughput screening lists. Therefore, different approaches based on larger and more diverse chemical libraries, including small fragments (20, 58, 59), peptides, peptidomimetics, and natural products may be needed for more efficient discovery of cryptic site ligands. A case in point is the discovery of a number of ligands for cryptic allosteric sites and cryptic sites at protein-protein interfaces, such as IL-2, caspases, kinase PDK1, and PTP1B, by fragment-based tethering (20-22, 59). Our data suggests that cryptic sites are much more prevalent than previously expected. However, while such sites do provide additional opportunities for drug discovery, they may not ultimately lead to drugs.

Experimental characterization of a predicted cryptic site in PTP1B by NMR spectroscopy. Finally, to demonstrate the practical utility of our approach, we focused on the clinically significant protein PTP1B. Targeting PTP1B with small molecules has been challenging due to the lack of

specificity and bioavailability of substrate mimetics as well as the presence of only a single known allosteric pocket (38, 59, 60). In addition to identifying 4 of the 14 residues in the known allosteric cryptic site (59), our predictive model also suggested two additional putative cryptic sites, a site near the N-terminus and a site relatively close to the active site (**Fig. 3*B***). The latter site is interesting for several reasons. First, the predicted cryptic site residues form an internal cavity (between residues Ile 67 and Phe 95) in crystal structures of PTP1B that is large enough to accommodate a small molecule (volume of ~150 $Å^3$). Our molecular dynamics simulations suggest that small conformational changes in the cavity-forming loops could make the cavity accessible to the solvent and expand its size (up to 430 $Å^3$). Second, the site is in proximity of two cysteine residues, Cys 92 and Cys 121, that could be targeted covalently in small-molecule fragment screening by tethering (20). In fact, Cys 121 is an already known target of a covalent small-molecule modifier and an allosteric inhibitor of PTP1B, ABDF, but its mechanism of action remains unclear (61). Third, this cryptic site in PTP1B differs from the corresponding region in the closely related tyrosine-protein phosphatase non-receptor type 2 (TCPTP) at, for example, position 97 (glutamate instead of leucine). This difference between PTP homologs could be exploited to develop selective inhibitors that avoid the serious adverse effects associated with TCPTP inhibition in mice (59). Finally, the cryptic site may be allosterically coupled to the catalytic site; examining contacts between pairs of residues (35) suggests extensive coupling between the cryptic and catalytic sites (**Fig. S10*A***).

We experimentally studied the binding of ABDF to PTP1B to determine whether or not it involves the putative cryptic site. Although PTP1B has three other surface-exposed cysteine residues, ABDF covalently attaches specifically to the side chain of Cys 121 (**Fig. 3*B*** and **Fig. S10*B***). The Cys 121 side chain points towards the interior of the unlabeled protein, so binding of ABDF likely requires a conformational change in the protein. We were unable to obtain a crystal structure of ABDF-labeled PTP1B, in agreement with other reports that ABDF-labeled PTP1B, unlike *apo* PTP1B, is recalcitrant to crystallization (61). To determine whether or not the covalent label causes specific local conformational changes or globally perturbs the protein, we collected ^{1}H, ^{15}N TROSY HSQC NMR spectra of both *apo* and ABDF-labeled protein (**SI Text** and **Fig. S10*C-F***). Using previously published backbone resonance assignments (62), we observed no perturbation of chemical shifts for a number of residues distal to the predicted cryptic site, indicating that the effects are local and that the protein remains folded. In contrast, a cluster of residues nearby the predicted cryptic site were significantly perturbed (**Fig. 3*B*** and **Fig. S10*C-F***). Many other residues near the predicted cryptic site that would need to move for ligand binding, including the adjacent β-sheet and Cys 121 loop, were unassigned due to resonance broadening, which is indicative of conformational exchange. Collectively, these results point to structural flexibility in the vicinity of the predicted cryptic site and the specific perturbation of residues surrounding the predicted binding pocket, validating our prediction.

To conclude, we describe cryptic sites and a method that accurately, automatically, and efficiently predicts their locations in protein structures. Our results support the hypothesis of ubiquitous cryptic sites and suggest many new small-molecule protein targets, including those that are associated with diseases. Moreover, we illustrate how chemical tethering can be used to validate cryptic site predictions by discovering cryptic site ligands. Cryptic sites can also be characterized by experimental techniques that measure protein dynamics, such as NMR spectroscopy and room-temperature X-ray crystallography (63, 64), as well as by discovery of ligands through virtual screening against conformations with pockets computed by AllosMod or molecular dynamics simulations. Our approach provides a convenient first step for such characterizations.

Materials and Methods

We started by finding cryptic sites in the Protein Data Bank (PDB) (65, 66), as follows. First, we gathered structures of protein-ligand complexes as well as structures of proteins in ligand-free (unbound) conformations. We define binding residues as the residues with at least one atom within 5 Å from any atom of a ligand in the bound conformation (a binding site). Second, we removed the redundant protein occurrences in the dataset by applying sequence identity threshold of 40% (**SI Text**). Finally, we evaluated each binding site in the unbound conformation using pocket scores based on two pocket-detection algorithms, Fpocket and ConCavity (13, 14). Binding sites with bad pocket sores in the unbound conformation and good pocket scores in the bound conformation were defined as cryptic sites, whereas those with good pocket scores in both conformations were defined as binding pockets (**Tables S1** and **S5**). More details and methods are available in **SI Text**. The web server for predicting cryptic binding sites is available at http://salilab.org/cryptosite (username: *reviewer*, password: *reviewerpw*).

Acknowledgments

The authors thank Hao Fan, Marcus Fischer, Nir London, Avner Schlessinger, and other members of Sali lab for their comments and feedback. P.C. is supported by a Howard Hughes Predoctoral Fellowship; T.J.R. is supported by a predoctoral fellowship from the NIH (F31 CA180378) and the Krevans Fellowship; L.B. is supported by Bayer Science and Education Foundation; D.A.K. is supported by an A.P. Giannini Foundation Postdoctoral Research Fellowship; R.A.W. is supported by a National Science Foundation Graduate Research Fellowship; J.S.F. is supported by the National Institutes of Health (DP5 OD009180, R21 GM110580, P30 DK063720) and National Science Foundation (STC-1231306); A.S. is supported

by the National Institutes of Health (R01 GM083960, U54 RR022220, U54 GM094662, P01 AI091575, and U01 GM098256).

References

1. Nisius B, Sha F, & Gohlke H (2012) Structure-based computational analysis of protein binding sites for function and druggability prediction. *Journal of biotechnology* 159(3):123-134.
2. Campbell SJ, Gold ND, Jackson RM, & Westhead DR (2003) Ligand binding: functional site location, similarity and docking. *Current opinion in structural biology* 13(3):389-395.
3. Laurie AT & Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 21(9):1908-1916.
4. Hermann JC*, et al.* (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature* 448(7155):775-779.
5. Laskowski RA, Luscombe NM, Swindells MB, & Thornton JM (1996) Protein clefts in molecular recognition and function. *Protein science : a publication of the Protein Society* 5(12):2438-2452.
6. Bowman GR & Geissler PL (2012) Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proceedings of the National Academy of Sciences of the United States of America* 109(29):11681-11686.
7. Diskin R, Engelberg D, & Livnah O (2008) A novel lipid binding site formed by the MAP kinase insert in p38 alpha. *Journal of molecular biology* 375(1):70-79.
8. Durrant JD & McCammon JA (2011) Molecular dynamics simulations and drug discovery. *BMC biology* 9:71.
9. Horn JR & Shoichet BK (2004) Allosteric inhibition through core disruption. *Journal of molecular biology* 336(5):1283-1291.
10. Lexa KW & Carlson HA (2011) Full protein flexibility is essential for proper hot-spot mapping. *Journal of the American Chemical Society* 133(2):200-202.
11. Mitternacht S & Berezovsky IN (2011) Binding leverage as a molecular basis for allosteric regulation. *PLoS computational biology* 7(9):e1002148.
12. Henrich S*, et al.* (2010) Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of molecular recognition : JMR* 23(2):209-219.
13. Capra JA, Laskowski RA, Thornton JM, Singh M, & Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS computational biology* 5(12):e1000585.
14. Le Guilloux V, Schmidtke P, & Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics* 10:168.
15. Rossi A, Marti-Renom MA, & Sali A (2006) Localization of binding sites in protein structures by optimization of a composite scoring function. *Protein science : a publication of the Protein Society* 15(10):2366-2380.
16. Hopkins AL & Groom CR (2002) The druggable genome. *Nature reviews. Drug discovery* 1(9):727-730.
17. Sheridan RP, Maiorov VN, Holloway MK, Cornell WD, & Gao YD (2010) Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *Journal of chemical information and modeling* 50(11):2029-2040.
18. Arkin MR & Wells JA (2004) Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature reviews. Drug discovery* 3(4):301-317.
19. Wells JA & McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 450(7172):1001-1009.
20. Hardy JA & Wells JA (2004) Searching for new allosteric sites in enzymes. *Current opinion in structural biology* 14(6):706-715.
21. Ostrem JM, Peters U, Sos ML, Wells JA, & Shokat KM (2013) K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* 503(7477):548-551.
22. Sadowsky JD*, et al.* (2011) Turning a protein kinase on or off from a single allosteric site via disulfide trapping. *Proceedings of the National Academy of Sciences of the United States of America* 108(15):6056-6061.

23. Johnson DK & Karanicolas J (2013) Druggable protein interaction sites are more predisposed to surface pocket formation than the rest of the protein surface. *PLoS computational biology* 9(3):e1002951.
24. Bowman GR, Bolin ER, Hart KM, Maguire BC, & Marqusee S (2015) Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proceedings of the National Academy of Sciences of the United States of America* 112(9):2734-2739.
25. Loving KA, Lin A, & Cheng AC (2014) Structure-based druggability assessment of the mammalian structural proteome with inclusion of light protein flexibility. *PLoS computational biology* 10(7):e1003741.
26. Bakan A, Nevins N, Lakdawala AS, & Bahar I (2012) Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *Journal of chemical theory and computation* 8(7):2435-2447.
27. Brenke R*, et al.* (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* 25(5):621-627.
28. Grove LE, Hall DR, Beglov D, Vajda S, & Kozakov D (2013) FTFlex: accounting for binding site flexibility to improve fragment-based identification of druggable hot spots. *Bioinformatics* 29(9):1218-1219.
29. Kozakov D*, et al.* (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature protocols* 10(5):733-755.
30. Ngan CH*, et al.* (2012) FTMAP: extended protein mapping with user-selected probe molecules. *Nucleic acids research* 40(Web Server issue):W271-275.
31. Bernstein FC*, et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology* 112(3):535-542.
32. Benson ML*, et al.* (2008) Binding MOAD, a high-quality protein-ligand database. *Nucleic acids research* 36(Database issue):D674-678.
33. Demerdash ON, Daily MD, & Mitchell JC (2009) Structure-based predictive models for allosteric hot spots. *PLoS computational biology* 5(10):e1000531.
34. Zhu X & Mitchell JC (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins* 79(9):2671-2683.
35. Weinkam P, Chen YC, Pons J, & Sali A (2013) Impact of mutations on the allosteric conformational equilibrium. *Journal of molecular biology* 425(3):647-661.
36. Pedregosa F*, et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825−2830.
37. Schaul T*, et al.* (2010) PyBrain. *Journal of Machine Learning Research* 11:743-746.
38. Combs AP (2010) Recent advances in the discovery of competitive protein tyrosine phosphatase 1B inhibitors for the treatment of diabetes, obesity, and cancer. *Journal of medicinal chemistry* 53(6):2333-2344.
39. O'Boyle NM*, et al.* (2011) Open Babel: An open chemical toolbox. *Journal of cheminformatics* 3:33.
40. Gunasekaran K, Ma B, & Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins* 57(3):433-443.
41. Huang Z*, et al.* (2011) ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic acids research* 39(Database issue):D663-669.
42. Jenkins AE, Hockenberry JA, Nguyen T, & Bernlohr DA (2002) Testing of the portal hypothesis: analysis of a V32G, F57G, K58G mutant of the fatty acid binding protein of the murine adipocyte. *Biochemistry* 41(6):2022-2027.
43. Schlessinger A & Rost B (2005) Protein flexibility and rigidity predicted from sequence. *Proteins* 61(1):115-126.
44. Shih CH, Chang CM, Lin YS, Lo WC, & Hwang JK (2012) Evolutionary information hidden in a single protein structure. *Proteins* 80(6):1647-1657.
45. Swapna LS, Bhaskara RM, Sharma J, & Srinivasan N (2012) Roles of residues in the interface of transient protein-protein complexes before complexation. *Scientific reports* 2:334.
46. Bartova I, Koca J, & Otyepka M (2008) Functional flexibility of human cyclin-dependent kinase-2 and its evolutionary conservation. *Protein science : a publication of the Protein Society* 17(1):22-33.

47. Martens D, Huysmans J, Setiono R, Vanthienen J, & Baesens B (2008) Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. *Studies in Computational Intelligence* 80:33-63.
48. Brown SP & Hajduk PJ (2006) Effects of conformational dynamics on predicted protein druggability. *ChemMedChem* 1(1):70-72.
49. Kozakov D*, et al.* (2011) Structural conservation of druggable hot spots in protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America* 108(33):13528-13533.
50. Weinkam P*, et al.* (TBD) Mapping protein allosteric mechanisms with small angle X-ray scattering profiles. *submitted*.
51. Molnar KS*, et al.* (2014) Cys-scanning disulfide crosslinking and Bayesian modeling probe the transmembrane signaling mechanism of the histine kinase, PhoQ. *submitted*.
52. Liao M, Cao E, Julius D, & Cheng Y (2013) Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* 504(7478):107-112.
53. Overington JP, Al-Lazikani B, & Hopkins AL (2006) How many drug targets are there? *Nature reviews. Drug discovery* 5(12):993-996.
54. Russ AP & Lampel S (2005) The druggable genome: an update. *Drug discovery today* 10(23-24):1607-1610.
55. Schmidtke P & Barril X (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of medicinal chemistry* 53(15):5858-5867.
56. Hert J, Irwin JJ, Laggner C, Keiser MJ, & Shoichet BK (2009) Quantifying biogenic bias in screening libraries. *Nature chemical biology* 5(7):479-483.
57. Mobley DL & Dill KA (2009) Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure* 17(4):489-498.
58. Makley LN & Gestwicki JE (2013) Expanding the number of 'druggable' targets: non-enzymes and protein-protein interactions. *Chemical biology & drug design* 81(1):22-32.
59. Wiesmann C*, et al.* (2004) Allosteric inhibition of protein tyrosine phosphatase 1B. *Nature structural & molecular biology* 11(8):730-737.
60. Krishnan N*, et al.* (2014) Targeting the disordered C terminus of PTP1B with an allosteric inhibitor. *Nature chemical biology* 10(7):558-566.
61. Hansen SK*, et al.* (2005) Allosteric inhibition of PTP1B activity by selective modification of a non-active site cysteine residue. *Biochemistry* 44(21):7704-7712.
62. Meier S*, et al.* (2002) Backbone resonance assignment of the 298 amino acid catalytic domain of protein tyrosine phosphatase 1B (PTP1B). *Journal of biomolecular NMR* 24(2):165-166.
63. Fraser JS*, et al.* (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proceedings of the National Academy of Sciences of the United States of America* 108(39):16247-16252.
64. Fischer M, Shoichet BK, & Fraser JS (2015) One Crystal, Two Temperatures: Cryocooling Penalties Alter Ligand Binding to Transient Protein Sites. *Chembiochem : a European journal of chemical biology*.
65. Berman HM*, et al.* (2002) The Protein Data Bank. *Acta crystallographica. Section D, Biological crystallography* 58(Pt 6 No 1):899-907.
66. Berman HM*, et al.* (2000) The Protein Data Bank. *Nucleic acids research* 28(1):235-242.
67. Pettersen EF*, et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* 25(13):1605-1612.

Figure Legends

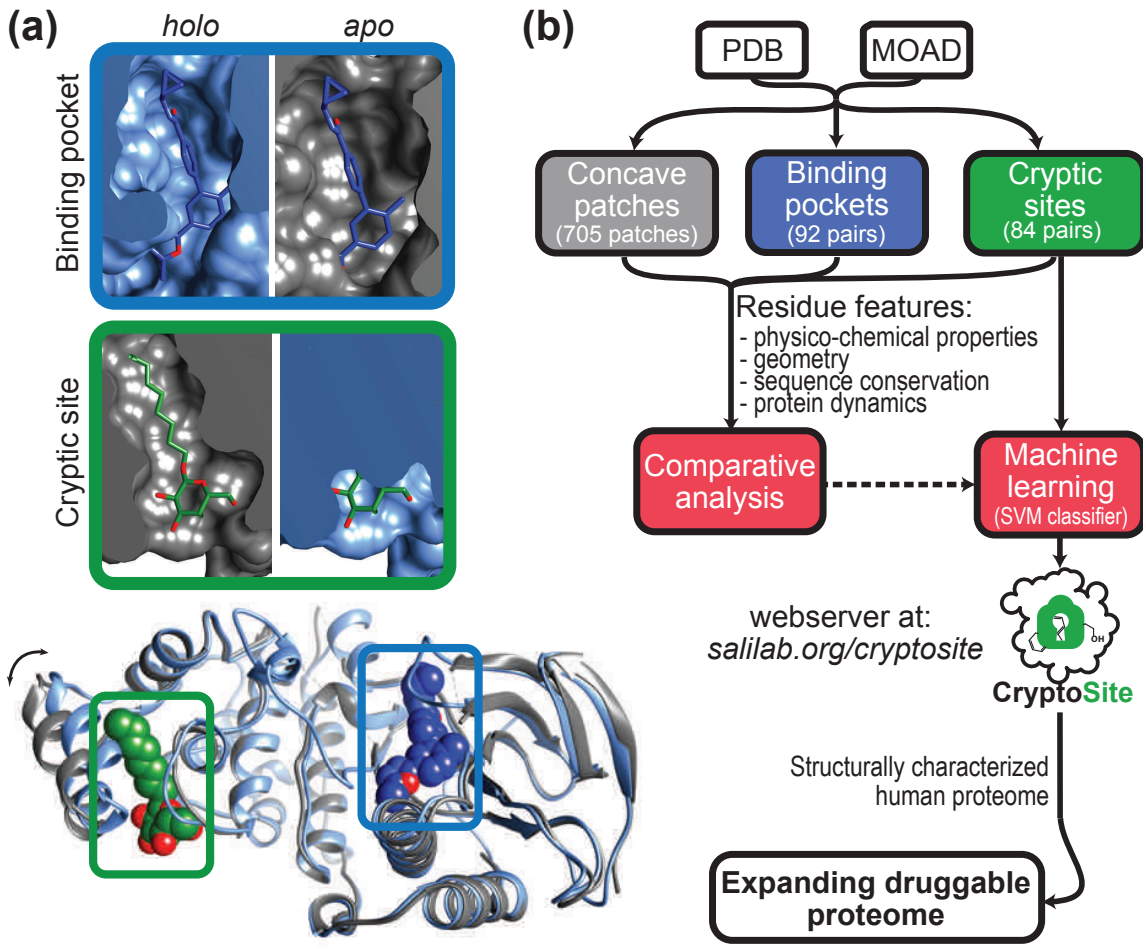

Figure 1: (*a*) Examples of a pocket and cryptic site in p38 MAP kinase. The nucleotide-binding site of the p38 MAP kinase is a pocket visible in both bound (*holo*; blue ribbon; PDB ID: 2ZB1) and unbound (*apo*; grey ribbon; PDB ID: 2NPQ) conformations. The ligand, biphenyl amide inhibitor, is depicted as blue spheres. On the other hand, the site in the C-lobe domain that binds octyglucoside lipid (green spheres) becomes a visible pocket only after the movement of the α-helix at the left of the structure (marked with the double-headed arrow). The small molecules are shown as they bind in the *holo* structures. UCSF Chimera software was used for the visualization (67). (*b*) Flowchart summarizing the analyses in this study. We started by creating a representative dataset of 84 known examples of cryptic binding sites, 92 binding pockets, and 705 concave surface patches from the Protein Data Bank (31) and the MOAD database (32). Next, we designed a set of 58 features that describe sequence, structure, and dynamics of individual residues and their neighbors. We then compared these attributes between the three types of a site to better understand the underlying characteristics of each site. Next, we used machine-learning algorithms to classify residues as belonging to a cryptic site or not. We then

predicted cryptic sites in the entire structurally characterized human proteome (**Materials and Methods, SI Text**).

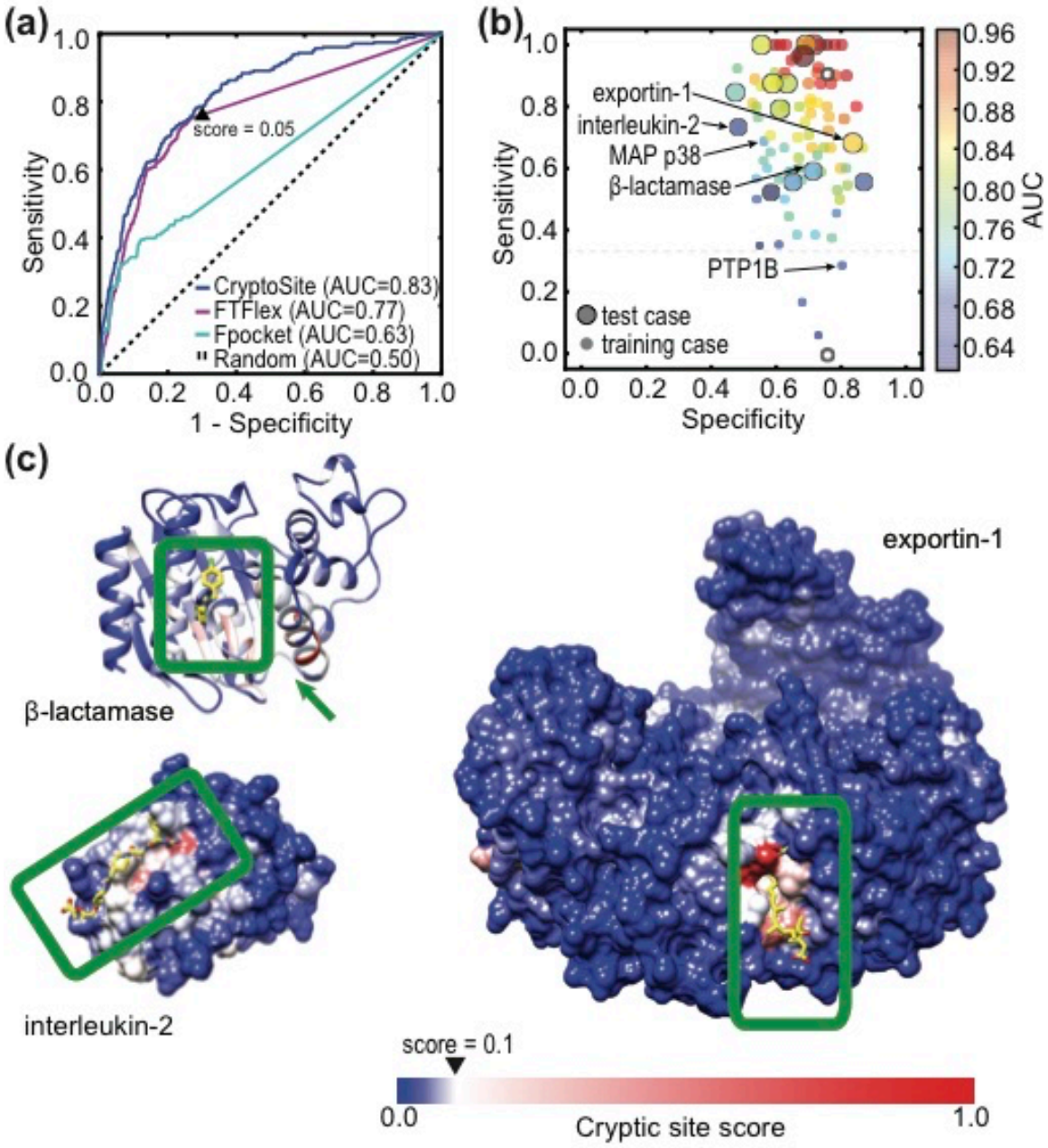

Figure 2: The accuracies of our predictive model, FTFlex, and Fpocket are measured as the area under the receiver-operating characteristic (ROC) curve based on predictions on all proteins in the test set (*a*), as well as based on sensitivity (true positive rate) and specificity (true negative rate) values from predictions on individual proteins (*b*). (*a*) Only ~45% and ~80% of cryptic site residues were detected by Fpocket and FTFlex, respectively; the area under the ROC curve was calculated by connecting the end of the ROC curve and the upper-right corner as a straight line. The accuracy of CryptoSite is comparable to that of FTFlex when small pockets that could fit

small-molecule fragments are already present in the *apo* state of a cryptic site (this is the case in 10 out of 14 testing examples). However, CryptoSite is more accurate than FTFlex when a cryptic site is buried or resides in a large protein (**Fig. S8*A***). (*b*) Sensitivities and specificities were determined for each protein in our test set (larger data points with black circle) and training set (smaller data points) based on leave-one-out cross-validation. The classification of the residues is based on the score threshold of 0.1. The two empty circles denote two predictions (one failed) of cryptic sites in proteins with more than one cryptic site. (*c*) The cryptic sites from our dataset are marked by green rectangles, and the computed scores that a residue is in a cryptic site are shown on the blue-to-red color scale. The small molecules that bind into the known cryptic sites are superposed from the alignment to the bound conformations and represented as yellow sticks.

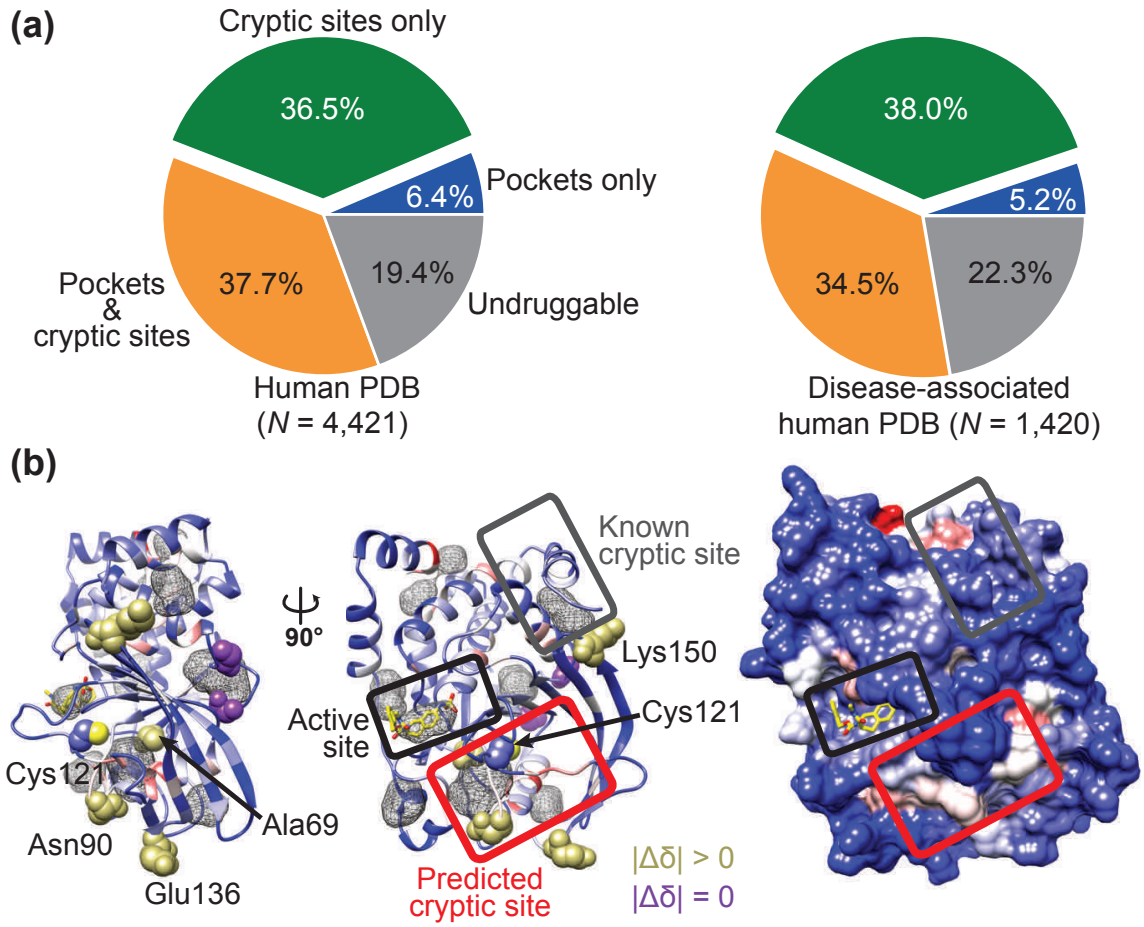

Figure 3: Cryptic binding sites are predicted to expand the size of the druggable proteome. (*a*) The percentage of proteins for which no binding sites (grey), only cryptic sites (green), only binding pockets (blue), and both cryptic sites and binding pockets (orange) were predicted for all human proteins with known structure (left pie chart) and for a subset of disease-associated proteins (right pie chart). Shown are the results of the fast version of our predictive model that does not take into account features based on molecular dynamics simulations. **(*b*)** Cryptic binding sites in PTP1B. Ribbon (left and center) and surface (right) representations of the PTP1B structure (PDB ID: 2F6V) are colored based on the cryptic site score as in **Fig. 2*C***. Residues with definitive chemical shift changes (|Δδ|) upon ABDF labeling (khaki) cluster around the cryptic and ABDF binding sites, whereas residues whose chemical shifts definitively do not change (purple) are more distal. The panel also shows positions and average volumes of the pockets (grey mesh) that are at least partially open more than 50% of the time, as observed in the molecular dynamics simulation at 300 K.

Supporting information

SI Text

The data set generation. We started by collecting all crystal structure PDB IDs of protein-ligand complexes from Binding MOAD (1) (downloaded on 2-27-2012); we only considered as ligands organic small molecules of biological relevance, excluding water and other solvent molecules, counterions, buffer components, metal ions, and crystallographic additives. We defined a binding site by selecting residues with at least one atom less than 5 Å away from any of the ligand atoms. Next, we searched for the structures of the same protein without any ligands at a given binding site, following these steps and criteria:

(i) we aligned all protein chain sequences from the Binding MOAD database to all protein chain sequences from PDB that are longer than 50 residues using the *blastp* algorithm (2), and then selected pairs with 100% sequence identity as *apo-holo* pair candidates (504,647 pairs);

(ii) we removed pairs for which either of the two structures was determined at worse than 2.5 Å resolution;

(iii) we removed pairs with ligands in *apo* structures that have at least one atom closer than 10 Å to any atom in the *holo* binding site;

(iv) we grouped *apo-holo* pairs with identical sequences into clusters and for each cluster selected a single pair with the lowest all-atom binding site RMSD as the cluster representative (this resulted in 46,436 pairs);

(v) we further removed *apo* structures that contain other proteins, peptides, or nucleic acids bound within 10 Å from the ligand of interest, superimposed from the *holo* structure;

(vi) we removed *apo-holo* pairs that contained multiple copies of a ligand at the *holo* binding site, that contained amino acid ligands, or pairs whose *holo* binding sites contained less than 5 residues (21,928 pairs remained);

(vii) we removed *apo-holo* pairs with sequence gaps in *apo* structures longer than 3 residues or less than 5 Å away from the binding site;

(viii) we grouped protein sequences into clusters of 40% protein sequence identity, and then further split these clusters into groups of proteins that bind similar ligands (we defined ligand similarity by the Tanimoto distance using linear path fingerprints (FP2) from Open Babel (3), followed by selecting the pair with the lowest all-atom RMSD from each group as the cluster representative;

(ix) and finally, we removed all *apo-holo* pairs with Cα-RMSD > 10 Å. This filtering resulted in a set of 4,766 *apo-holo* structure pairs.

We next utilized two pocket detection algorithms, ConCavity (4) and Fpocket (5), to evaluate the "goodness" of pockets in the *apo* and *holo* structures. The output of the Fpocket algorithm is a list of pockets with corresponding druggability scores, with each pocket defined as a set of coordinates depicting centers of fitting (alpha) spheres. We define the Fpocket residue pocket score as the maximum druggability score among the alpha spheres within 5 Å of the residue, or 0 if there are no alpha spheres (and hence pockets) in its neighbourhood. In contrast, ConCavity already provides a score on a per-residue basis, which we define as the ConCavity residue pocket score without additional processing. We use both Fpocket and ConCavity residue pocket scores to define cryptic sites and binding pockets. Cryptic sites are defined as sites with an average residue pocket score of less than 0.1 in the *apo* form and more than 0.4 in the *holo* form. Similarly, we defined binding pockets as binding sites with an average residue pocket score of more than 0.4 for the *apo* and *holo* forms, and Qi (6) between the *apo* and *holo* forms larger than 0.95. Such filtering resulted in a dataset of 468 *apo-holo* pairs with cryptic sites (190 unique *apo* structures), and 839 *apo-holo* pairs with binding pockets (191 unique *apo* structures).

We had to manually inspect both datasets of binding sites because of the high false-positive rate of pocket detection algorithms (the state-of-the-art algorithms are only ~70% sensitive (7, 8) when applied to the unbound conformation of a protein), which resulted in the final datasets of 89 cryptic sites and 92 binding pocket *apo-holo* pairs. 10 randomly chosen cryptic *apo-holo* pairs were put aside for testing purposes. Also for testing purposes, we additionally selected 4 proteins with known cryptic sites from the literature (exportin-1, TEM1 β-lactamse, IL-2, and Bcl-X) (**Tables S1** and **S5**).

In summary, the sequence similarity between a pair of two *apo* structures never exceeds 40%, except for 7 proteins that contain 2 different cryptic sites each, and a protein that contained 3 different cryptic sites. Moreover, out of 79 proteins in total, we obtained 59 groups of proteins with putative unique folds based on protein structure alignment (TM-align and TM-score thresholds of more than 0.7) (9). Similarly, we retrieved a non-redundant dataset of 92 protein structures with binding pockets; none of the protein sequences is more than 40% identical to any other sequence, and protein structure alignment suggests 69 putative folds.

Pre-processing PDB files. Many PDB files contain more than 1 macromolecule (*ie*, a biologically relevant assembly of multiple macromolecules or an assembly of macromolecules interacting through crystallographic contacts), non-specific solvent molecules, regions of missing density, and modified protein sequences (*eg,* truncated loops or termini). To more accurately assess structural properties (for example, an estimate of surface area would be inaccurate for the residues next to an interacting molecule or a region with a missing density), we deleted from the PDB file all macromolecules except the macromolecule (*ie,* chain) of interest. Furthermore, we filled the gaps in the crystal structures by aligning a PDB structure to the corresponding SEQRES

sequence, and then used the loop-modeling routine in Modeller (10) to build a loop conformation while keeping the rest of the protein structure rigid. We built 20 models per chain, and kept the one with the lowest DOPE score (11) for further analyses.

Molecular dynamics simulations. Standard molecular dynamics simulations are computationally expensive, which makes them impractical for studying the dynamics of the large number of proteins in our dataset. In contrast, AllosMod simulates dynamics more efficiently, by relying on a simplified energy landscape whose minimum is defined by the input native structure (6). We initialized 50 simulations from the randomized *apo* crystal structure coordinates, each 6 ns long. The 50 simulations include 10 repeats at 5 different temperatures (300 K, 350 K, 400 K, 450 K, and 500 K), with 3 ps time steps – resulting in a total of 100,000 snapshot conformations. All conformations were assessed using our statistical potential SOAP (12), and only those with SOAP scores lower than 160% of the score of the native protein structure were retained for further analysis.

Feature design. In total, we curated a set of 58 residue-based features that can be grouped into 3 categories: (*i*) features that describe protein sequence conservation, protein shape, and energetics, (*ii*) features that describe sequence conservation, shape, and energetics of neighborhood residues, and (*iii*) features derived from molecular dynamics simulations describing flexibility and dynamics of residues (**Table S2**). Protein shape calculations include *protrusion, compactness, convexity, rigidity, hydrophobicity* (using Wimley-White solvent model)*,* and *charge density*, as described previously (13). *Residue surface area* is defined as a sum of surface areas of individual atoms, which was determined by the CHASA algorithm (default probe radius) (14) and Modeller (probe radius of 1.4 Å and 3.0 Å). We define *residue packing* of a given residue as the number of atoms of other residues within 4 Å from any atom in the residue, divided by the number of atoms in the residue. The *number of neighbors* is defined as the number of different residues within the same distance. *Distance to the surface* is defined as the smallest distance between any atom of a given residue and the closest atom with surface area > 2 $Å^2$. *Pocket score* is derived from pocket prediction by Fpocket as explained above (*Data set generation* section). *Number of atoms and residues in the neighborhood*, *number (weighted or not) of side-chain rotatable bonds in the neighborhood*, and *local structural entropy* were calculated as described previously (15-17).

Sequence conservation of a given sequence position is defined as the Shannon's entropy of reweighted amino-acid frequency counts in a multiple sequence alignment (18). Multiple sequence alignments were obtained by aligning an individual *apo* sequence against the entire Uniprot (19, 20) database using the *blastp* algorithm. Clusters of homologous sequences above the 80% sequence identity threshold (used to reweight the amino-acid frequency counts) were calculated using the *usearch* algorithm (21).

The *fragment docking* feature was calculated as follows. We started by docking 16 small-molecule probes (22-24), using PatchDock (25), resulting in a number of different poses for each ligand. Next, we scored each ligand pose using the RankScore statistical potential (26), and filtered out poses with RankScore larger than 0. Finally, a fragment docking score was assigned to each residue in a protein, corresponding to the number of contacts between the residue and ligands in any calculated pose (a residue and a ligand are in contact when the minimum distance between any residue-ligand atom pair is less than 3.5 Å).

Features derived from molecular dynamics simulations include the mean and standard deviation of the following residue features: pairwise distance similarity metric (*Qi*), surface exposed area (with probe radius of 1.4 Å and 3.0 Å), protrusion, convexity, and pocket score. Additionally, we also calculated the percentage of snapshots with a given residue pocket score higher than 0.4, as well as the mean and standard deviation of the residue pocket scores above the 95th percentile.

Machine learning. To predict whether a given residue belongs to a cryptic site, we utilized Scikit-Learn and PyBrain implementations (27, 28) of several different supervised machine-learning algorithms. We varied many parameters associated with a given algorithm (*eg*, different kernel functions, a range of different values for penalty parameters, different penalty functions, *etc*.). Furthermore, we mapped the accuracy as a function of scaling the dataset or changing class weights to take into account the unbalanced dataset (only ~5% of residues in our dataset are in cryptic sites). The residue classification accuracy of each combination of scaling, algorithm, and the corresponding set of parameters was evaluated using the confusion matrix and leave-one-out cross-validation (**Fig. S4*A***), with *n - 1* proteins used for training and *1* for validation, repeated over all cases in the training set. The SVM algorithm with quadratic kernel function, scaling, and penalty parameter *C*, kernel coefficient *gamma*, and independent term in kernel function *coef0* of 0.158, 0.333, 2.154, respectively, was found to perform most accurately. Furthermore, using a greedy-forward approach, evaluating area under the ROC curve and leave-one-out cross-validation (**Fig. S4*B***), we selected a subset of 3 features (*the average pocket score in MD simulations, sequence conservation, and fragment docking*). The web server for predicting cryptic binding sites is available at http://salilab.org/cryptosite. On average, it takes less than 2 days on our web server to predict cryptic sites in a protein of ~300 residues (most of this time is spent on molecular dynamics simulations by AllosMod).

Estimating the size of the druggable proteome. To estimate the size of the druggable proteome, we first retrieved a subset of 11,201 human protein structures from the PDB longer than 50 residues and with X-ray resolution better than 3.5 Å. For each one of these structures, we predicted cryptic sites by using our algorithm without residue-based features that require time-consuming AllosMod simulations (**Table S2**). A cryptic site is predicted when at least 5 adjacent residues have the cryptic site score larger than 0.056; two residues are adjacent when any of

their atoms are within 3.5 Å of each other. A binding pocket is predicted equivalently, but using the Fpocket-based pocket score with a threshold of 0.5. The two thresholds were chosen to approximately match the sensitivity and specificity of cryptic site and binding pocket prediction (true positive rates of 0.51 and 0.57, and false positive rates of 0.22 and 0.21 for cryptic sites and binding pockets, respectively (7)). To estimate the number of druggable disease-associated proteins, we first retrieved a dataset of disease-associated genes from OMIM *morbidmap* (3,329 genes) (29). Druggable disease-associated proteins are defined as proteins of known structure that are encoded by these genes and have at least one predicted cryptic site or binding pocket; for proteins with more than one determined structure, we only include into our analysis the structure with the highest number of predicted cryptic sites or pockets.

Protein expression and purification. The short form of the catalytic domain (residues 1-298) of wild-type human PTP1B was cloned into pET24b. BL21 *E. coli* cells were transformed with this construct. 5 mL overnight cultures of the transformed cells were diluted into 1 L of M9 minimal medium with 1 g/L $^{15}NH_4Cl$ and 35 µg/mL kanamycin, and grown at 37ºC until absorbance at 600 nm reached 0.95 (about 7 hours). PTP1B expression was induced by adding isopropyl-β-D-thiogalactoside (IPTG) to a concentration of 0.5 mM and incubating for 16 hours at 18ºC. Cell pellets were harvested by centrifugation and stored at -80ºC.

For purification, cell pellets were resuspended in lysis buffer (100 mM MES pH 6.5, 1 mM EDTA, 1 mM DTT) (30) and lysed by homogenization with an Emulsiflex C3 machine. After centrifugation of the lysate, the supernatant was filtered and loaded onto a Sepharase (SP) cation exchange column equilibrated in lysis buffer. The column was run over a gradient from 0-1 M NaCl; PTP1B eluted around 200 mM NaCl. Those fractions were pooled, concentrated by centrifugation, and loaded onto a Superdex 200 (S200) size-exclusion column equilibrated in 100 mM MES pH 6.5, 1 mM EDTA, 1 mM DTT, 200 mM NaCl. PTP1B-containing fractions were pooled, filtered, and dialysed at 4ºC for 1-2 hours into NMR buffer (20 mM Bis-Tris propane, 25 mM NaCl, 3mM DTT, 0.2 mM EDTA, pH 6.5) (31). The protein sample was then concentrated via centrifugation to 230 µM.

Covalent labeling of PTP1B with ABDF. The protein sample was diluted to 25 µM in NMR buffer without DTT. We then added 500 µM ABDF for 1 hour at room temperature. Next, the unreacted ABDF was removed and the protein was exchanged back into NMR buffer with DTT using a PD10 desalting column. Finally, the protein was concentrated via centrifugation to 110 µM.

TROSY NMR data acquisition. We prepared NMR samples with 7% D_2O and 200 and 110 µM of the apo and ABDF-labeled protein species, respectively. ^{1}H, ^{15}N TROSY HSQC spectra were collected with a Bruker 800 MHz magnet at 293 K for >5 hours and >7 hours, respectively.

Although many resonances were too broadened to confidently match with published assignments (32) because we used undeuterated protein in contrast to previous work (31-33), we were able to confidently monitor the resonances of several residues between the two spectra (**Fig. 3*B*** and **S10**).

References

1. Benson ML*, et al.* (2008) Binding MOAD, a high-quality protein-ligand database. *Nucleic acids research* 36(Database issue):D674-678.
2. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3):403-410.
3. O'Boyle NM*, et al.* (2011) Open Babel: An open chemical toolbox. *Journal of cheminformatics* 3:33.
4. Capra JA, Laskowski RA, Thornton JM, Singh M, & Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* 5(12):e1000585.
5. Le Guilloux V, Schmidtke P, & Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 10:168.
6. Weinkam P, Chen YC, Pons J, & Sali A (2013) Impact of mutations on the allosteric conformational equilibrium. *Journal of molecular biology* 425(3):647-661.
7. Schmidtke P & Barril X (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of medicinal chemistry* 53(15):5858-5867.
8. Schmidtke P, Le Guilloux V, Maupetit J, & Tuffery P (2010) fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic acids research* 38(Web Server issue):W582-589.
9. Xu J & Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 26(7):889-895.
10. Sali A & Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779-815.
11. Shen MY & Sali A (2006) Statistical potential for assessment and prediction of protein structures. *Protein science : a publication of the Protein Society* 15(11):2507-2524.
12. Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, & Sali A (2013) Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics* 29(24):3158-3166.
13. Rossi A, Marti-Renom MA, & Sali A (2006) Localization of binding sites in protein structures by optimization of a composite scoring function. *Protein Sci* 15(10):2366-2380.
14. Fleming PJ, Fitzkee NC, Mezei M, Srinivasan R, & Rose GD (2005) A novel method reveals that solvent water favors polyproline II over beta-strand conformation in peptides and unfolded proteins: conditional hydrophobic accessible surface area (CHASA). *Protein science : a publication of the Protein Society* 14(1):111-118.
15. Chan CH*, et al.* (2004) Relationship between local structural entropy and protein thermostability. *Proteins* 57(4):684-691.
16. Zhu X & Mitchell JC (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins* 79(9):2671-2683.
17. Demerdash ON, Daily MD, & Mitchell JC (2009) Structure-based predictive models for allosteric hot spots. *PLoS computational biology* 5(10):e1000531.
18. Morcos F*, et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* 108(49):E1293-1301.
19. UniProt C (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research* 41(Database issue):D43-47.
20. UniProt C (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* 42(Database issue):D191-198.
21. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.
22. Brenke R*, et al.* (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* 25(5):621-627.
23. Kozakov D*, et al.* (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature protocols* 10(5):733-755.

24. Kozakov D*, et al.* (2011) Structural conservation of druggable hot spots in protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America* 108(33):13528-13533.
25. Schneidman-Duhovny D, Inbar Y, Nussinov R, & Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research* 33(Web Server issue):W363-367.
26. Fan H*, et al.* (2011) Statistical potential for modeling and ranking of protein-ligand interactions. *J Chem Inf Model* 51(12):3078-3092.
27. Pedregosa F*, et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825−2830.
28. Schaul T*, et al.* (2010) PyBrain. *Journal of Machine Learning Research* 11:743-746.
29. Hamosh A, Scott AF, Amberger JS, Bocchini CA, & McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* 33(Database issue):D514-517.
30. Puius YA*, et al.* (1997) Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design. *Proceedings of the National Academy of Sciences of the United States of America* 94(25):13420-13425.
31. Whittier SK, Hengge AC, & Loria JP (2013) Conformational motions regulate phosphoryl transfer in related protein tyrosine phosphatases. *Science* 341(6148):899-903.
32. Meier S*, et al.* (2002) Backbone resonance assignment of the 298 amino acid catalytic domain of protein tyrosine phosphatase 1B (PTP1B). *Journal of biomolecular NMR* 24(2):165-166.
33. Krishnan N*, et al.* (2014) Targeting the disordered C terminus of PTP1B with an allosteric inhibitor. *Nature chemical biology*.
34. Mitternacht S & Berezovsky IN (2011) A geometry-based generic predictor for catalytic and allosteric sites. *Protein engineering, design & selection : PEDS* 24(4):405-409.
35. Grove LE, Hall DR, Beglov D, Vajda S, & Kozakov D (2013) FTFlex: accounting for binding site flexibility to improve fragment-based identification of druggable hot spots. *Bioinformatics* 29(9):1218-1219.
36. Kadono S*, et al.* (2005) Structure-based design of P3 moieties in the peptide mimetic factor VIIa inhibitor. *Biochemical and biophysical research communications* 327(2):589-596.

SI Figure and Table Legends

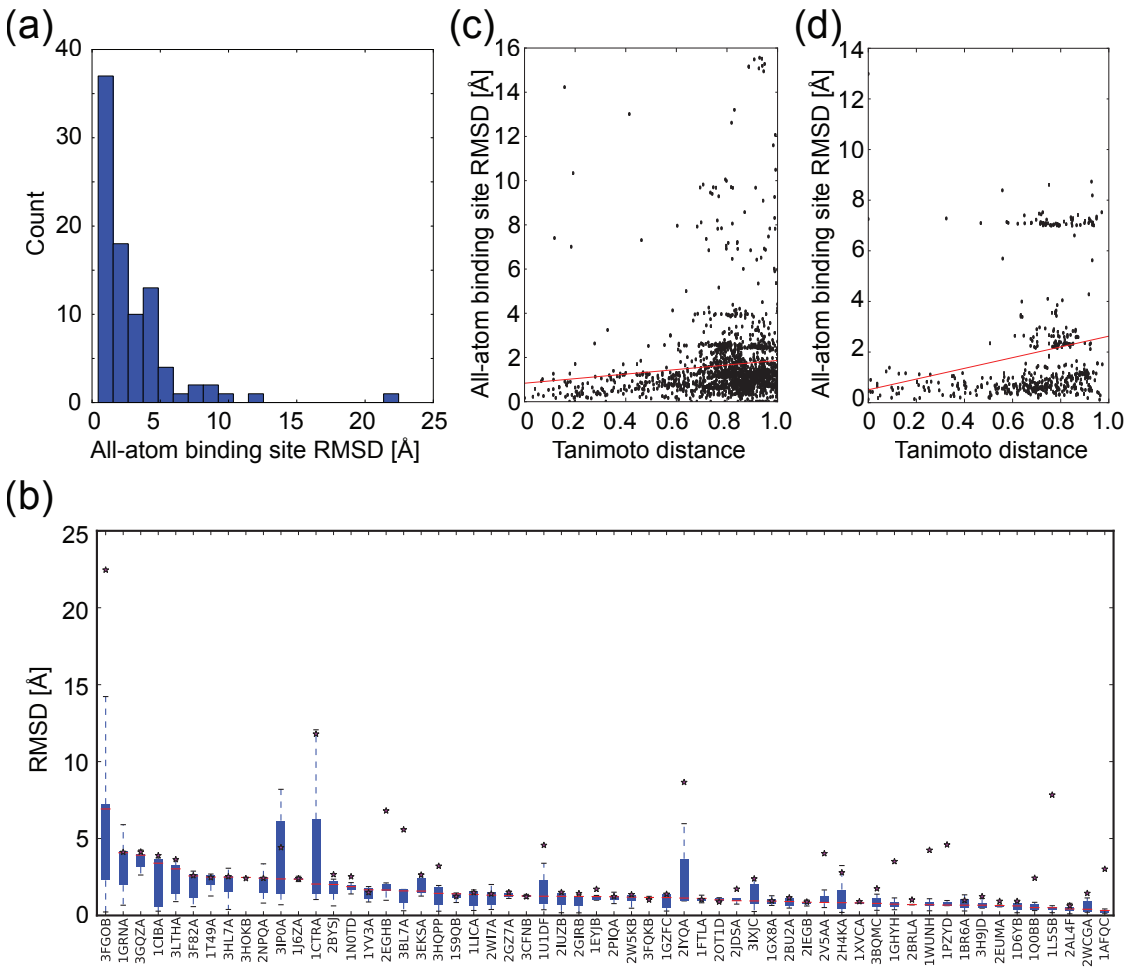

Figure S1: (*a*) Histogram of all-atom binding site RMSDs between *apo* and *holo* conformations. (*b*) Structural similarity (all-atom binding site RMSD) between cryptic site structures bound to at least 5 different ligands. Boxes, whiskers, and red lines denote 10th and 90th percentile, 5th and 95th percentile, and the median of the distribution. The similarities between unbound and bound conformations from our dataset are denoted by star symbols. The degree of structural similarity between bound cryptic sites (*c*) or binding pockets (*d*) is independent of the 2D structural similarity between the bound ligands. Linear path fingerprints (FP2) and Open Babel package were used to calculate the Tanimoto distances. The red line denotes linear fit, with a slope parameter that is not significantly different (R-value < 0.01) from the horizontal regression.

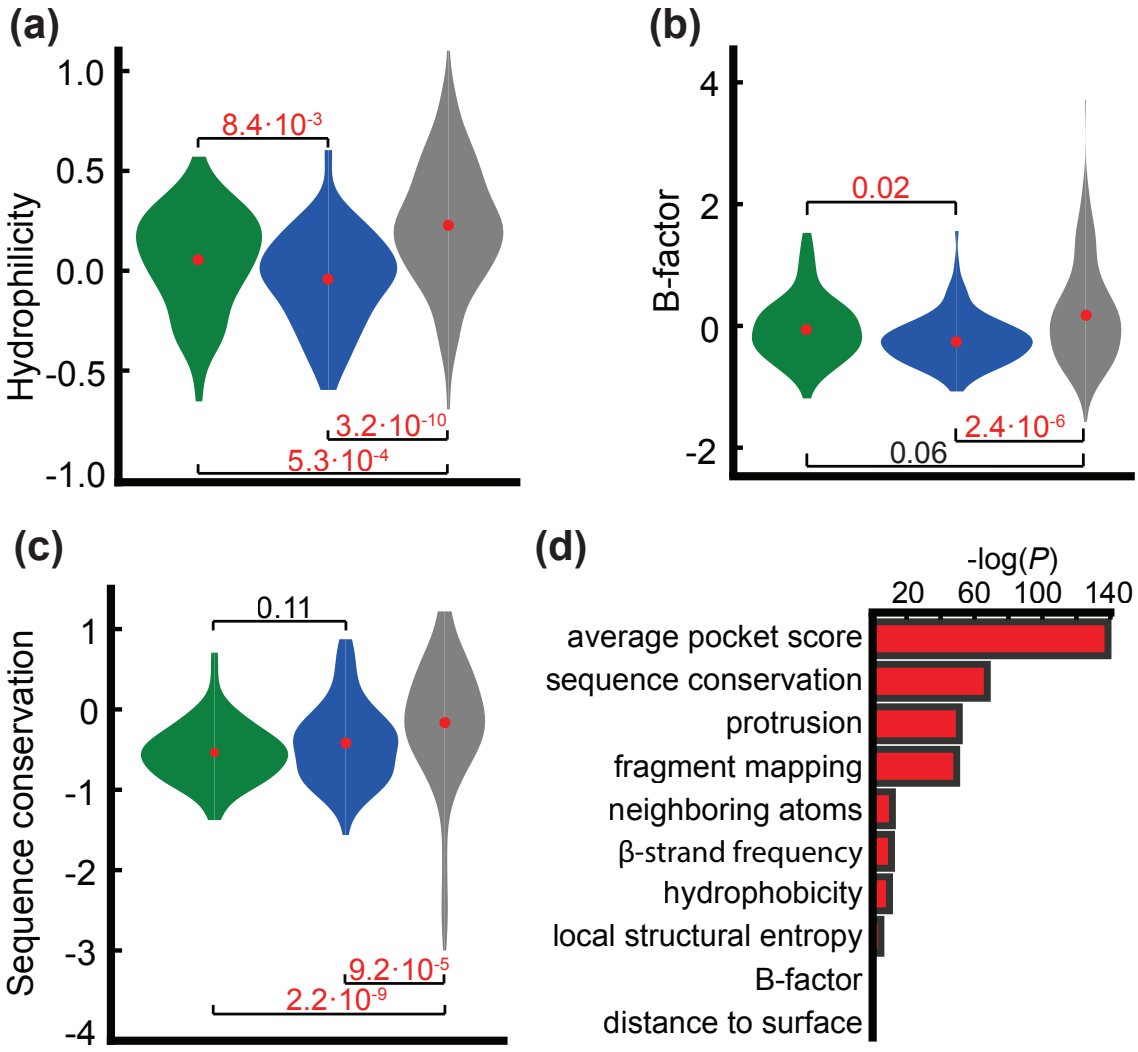

Figure S2: Comparison of cryptic sites, binding pockets, and random concave surface patches. (*a-c*) In each panel, the distribution of the feature values of binding site residues are shown as violin plots for cryptic sites (green), binchding pockets (blue), and random concave surface patches (grey). The edges between distributions denote P-values based on Kolmogorov-Smirnov two-sample statistics; numbers/letters in red are statistically significant ($P < 0.05$). (*d*) For a few selected residue-based features, the distributions of their values for the cryptic sites and the rest of residues in our dataset are compared. The bars denote statistical significance (P-value) from the two-sample Kologorov-Smirnov non-equality test (**Table S2** for the P-values of other features).

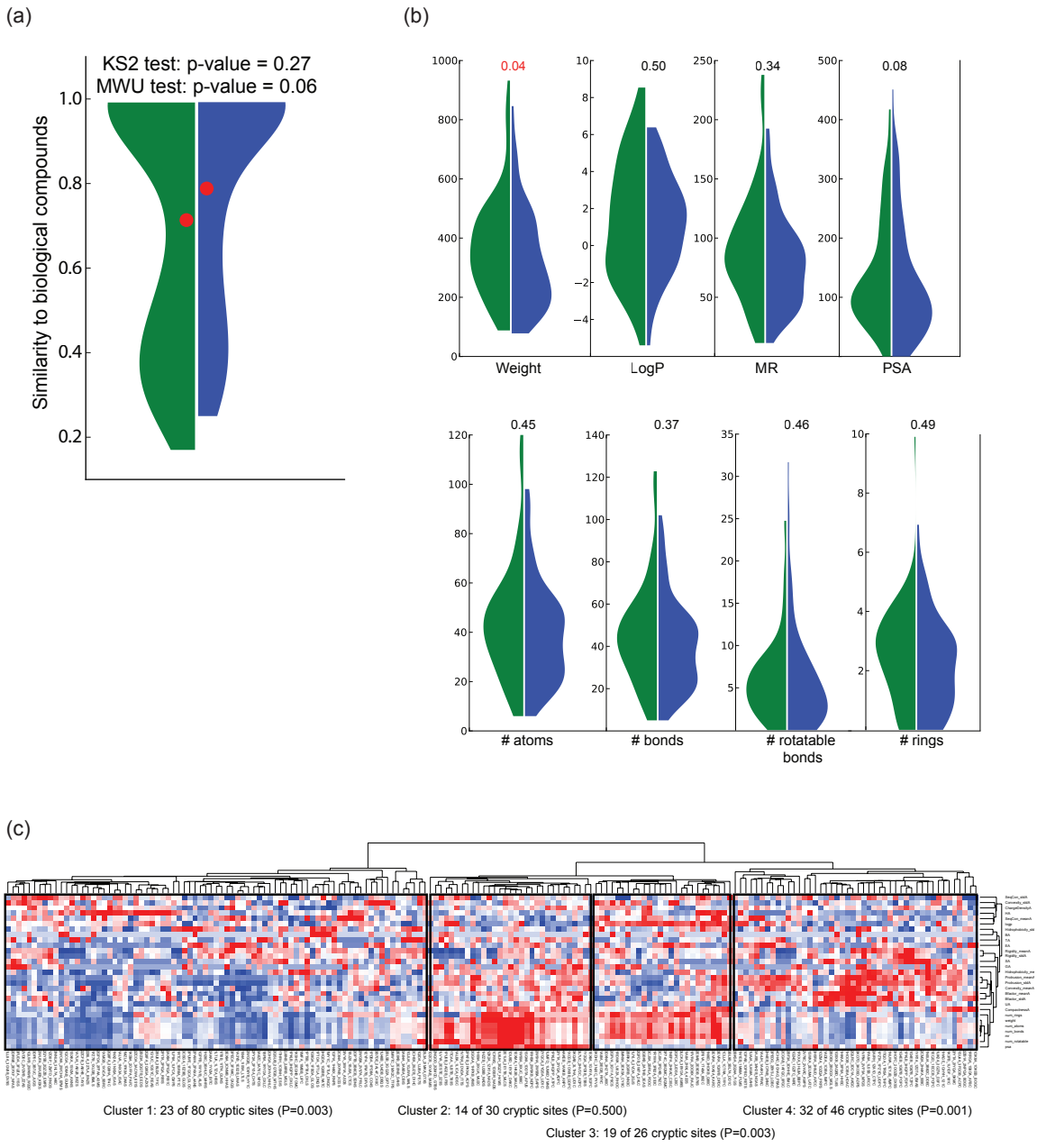

Figure S3: Comparison of small molecule-based features between ligands in cryptic sites (green half-violin plots), and ligands in pockets (blue-half violin plots). (*a*) The distributions of ligand similarities to biological compounds collected from the KEGG database of biological processes. (*b*) Distributions of several ligand descriptors, as determined by Open Babel. (*c*) 2-dimensional clustering of ligand and binding site features as well as binding sites identifies 4 clusters. Two of the clusters are significantly enriched with cryptic sites. One cluster includes convex sites with evolutionarily conserved residues and small hydrophilic ligands (cluster 4), and another one includes less convex and less conserved sites that bind larger hydrophobic ligands (cluster 3). The third cluster contains an equal number of cryptic sites and binding pockets that are

evolutionarily conserved and bind large hydrophilic ligands (cluster 2). The final cluster contains mostly binding pockets that are concave and evolutionarily conserved, and bind small and hydrophobic ligands (cluster 1).

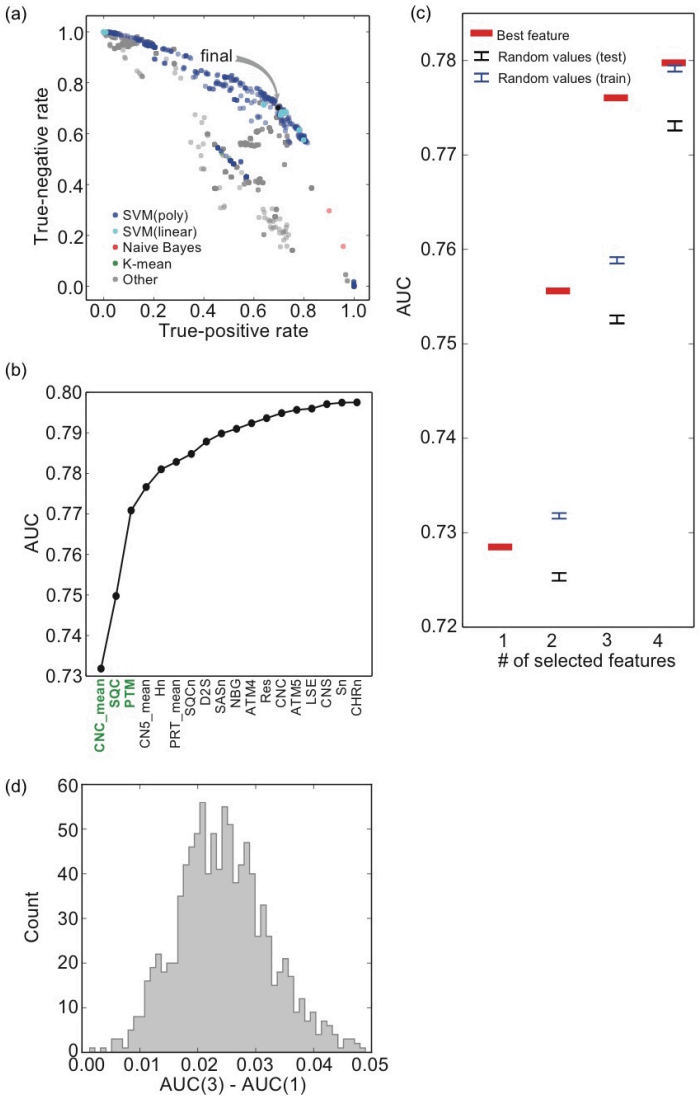

Figure S4: (*a*) Search for the most accurate machine-learning algorithm, data pre-processing method, and the corresponding set of parameters. The most accurate predictive model and its parameter values were selected by maximizing the sensitivity (true-positive rate) and the specificity (true-negative rate) of cryptic site residue classification, using leave-one-out cross validation on the training set of proteins with 84 cryptic binding sites. The arrow points to the most accurate algorithm. (*b*) Feature selection using greedy-forward approach. See **SI Table 3** for a description of feature labels. (*c*) To avoid the data overfitting during the feature selection protocol, we tested the statistical significance of the predictive model improvement by comparing the impact of each additional feature to that of a random value. Adding the best 3 features (red bars) always outperformed the models with the added random value feature (blue and black error bars), showing that the improvement based on adding the second and third features is statistically

significant (*P*-value < 0.001). The model with the 4 best-performing features was statistically no different from the predictive model with the 3 best-performing features and the random value feature (*P*-value > 0.05), leading to the final selection of only 3 features. The predictive models with the random value feature were evaluated using leave-one-out cross-validation, with the AUC values determined for both the data points left in (blue error bars) and those left out (black error bars); the difference in accuracy between the left-in and left-out samples suggests that our training strategy limits overfitting. The error bars denote standard deviations of the AUC values, based on 1,000 replicates. The small differences between the AUC values of models with the best set of features (red bars) in this plot and those in (*b*) are due to the numeric variability in the cryptic site prediction (**Fig. S5*A***). (*d*) To quantify the difference in accuracy between two models, we tested a null hypothesis that the difference between the AUC values from the two models is 0, when measured on exactly the same predictions. In 1000 bootstrapped samples, the AUC value from the model with 1 feature never exceeded that from the model with 3 features, rejecting the null hypothesis.

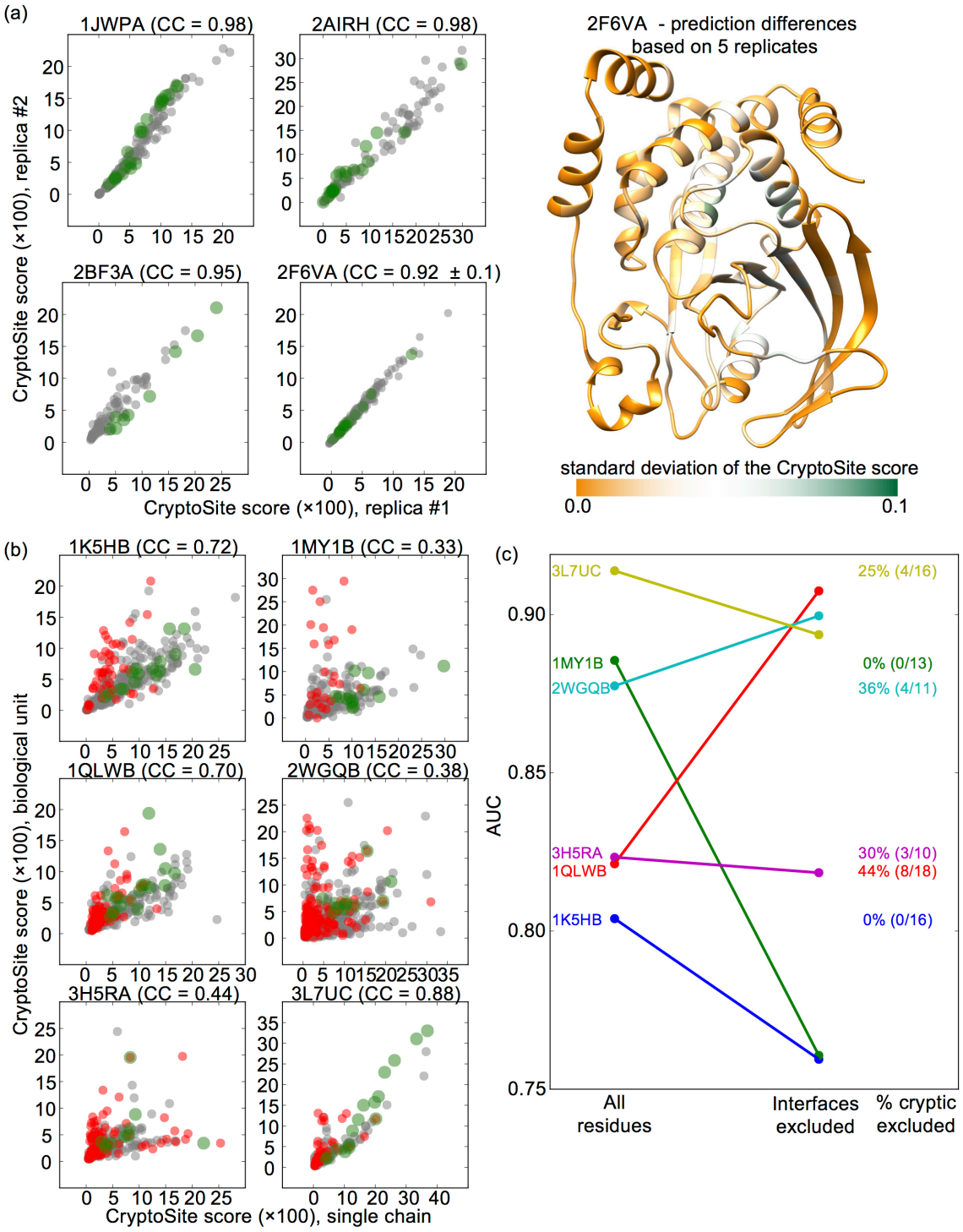

Figure S5: (*a*) Independent predictions based on different molecular dynamics trajectories are highly similar (left). Cryptic site residues and all other residues are shown in green and grey, respectively. The differences in the predicted score are the largest for residues that reside on α-

helices or β-sheets and are adjacent to flexible parts of a protein, but are too small in scale to affect the cryptic site predictions (the average residue score difference of the most variable decile is less than 0.04), as estimated from 5 runs on PTP1B (PDB ID: 2F6V:A) (right). (*b*) Predictions for a subunit on its own or in the context of a biological assembly are also highly similar, except for the subunit-subunit interface residues. The two types of a run were on average significantly correlated for the known cryptic sites (the mean cross-correlation coefficient of 0.60; green data points), but not for the interface residues, as expected (the mean cross-correlation coefficient of 0.41; red data points). In principle, prediction of cryptic sites based on an entire biological unit should be more accurate than that based on an isolated subunit. However, in practice, the actual accuracy may be smaller because of the increased inaccuracy of energy functions and less thorough sampling of larger systems compared to those for smaller systems (34). (*c*) Excluding protein-protein interface residues from the prediction of cryptic sites rarely improves the performance of CryptoSite.

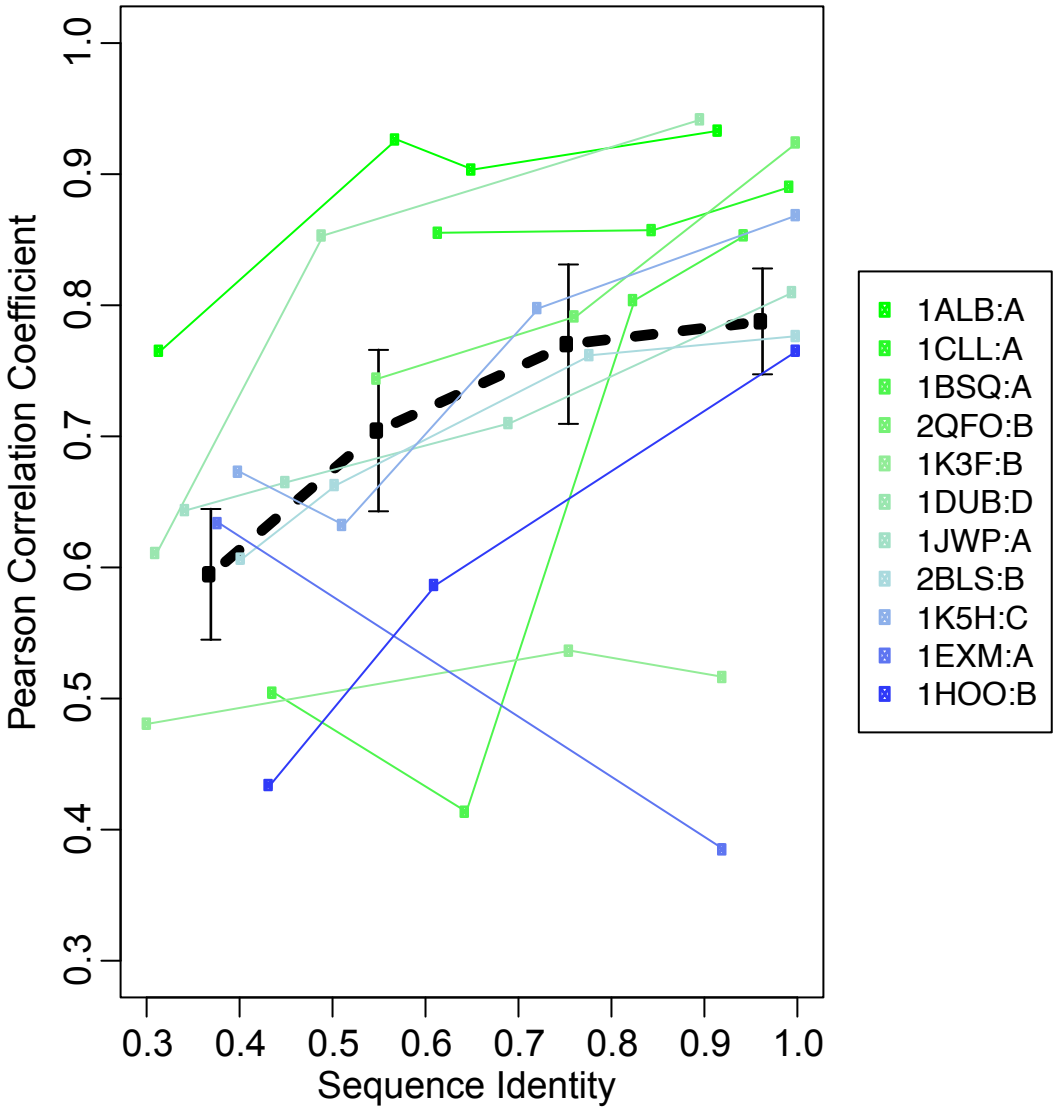

Figure S6: Comparative models based on templates with sequence identity larger than 50% result in cryptic site predictions similar to those based on original high-resolution X-ray structures (**Table S4**). We obtained multiple templates with varying sequence identities to the original protein sequence for a subset of proteins with cryptic sites in our dataset, using *pblast* (2). For each template, a comparative model was built using the default automodel class in Modeller (10), followed by prediction of cryptic site locations using CryptoSite. Trend line (dashed line) and error bars denote mean cross-correlation coefficients and their standard deviations, respectively. The outlying comparative model for elongation factor Tu (PDB ID: 1EXM) is due to a template (PDB

ID: 1MJ1) with a significantly different conformation of the C-terminal domain (backbone RMSD of 4.6 Å).

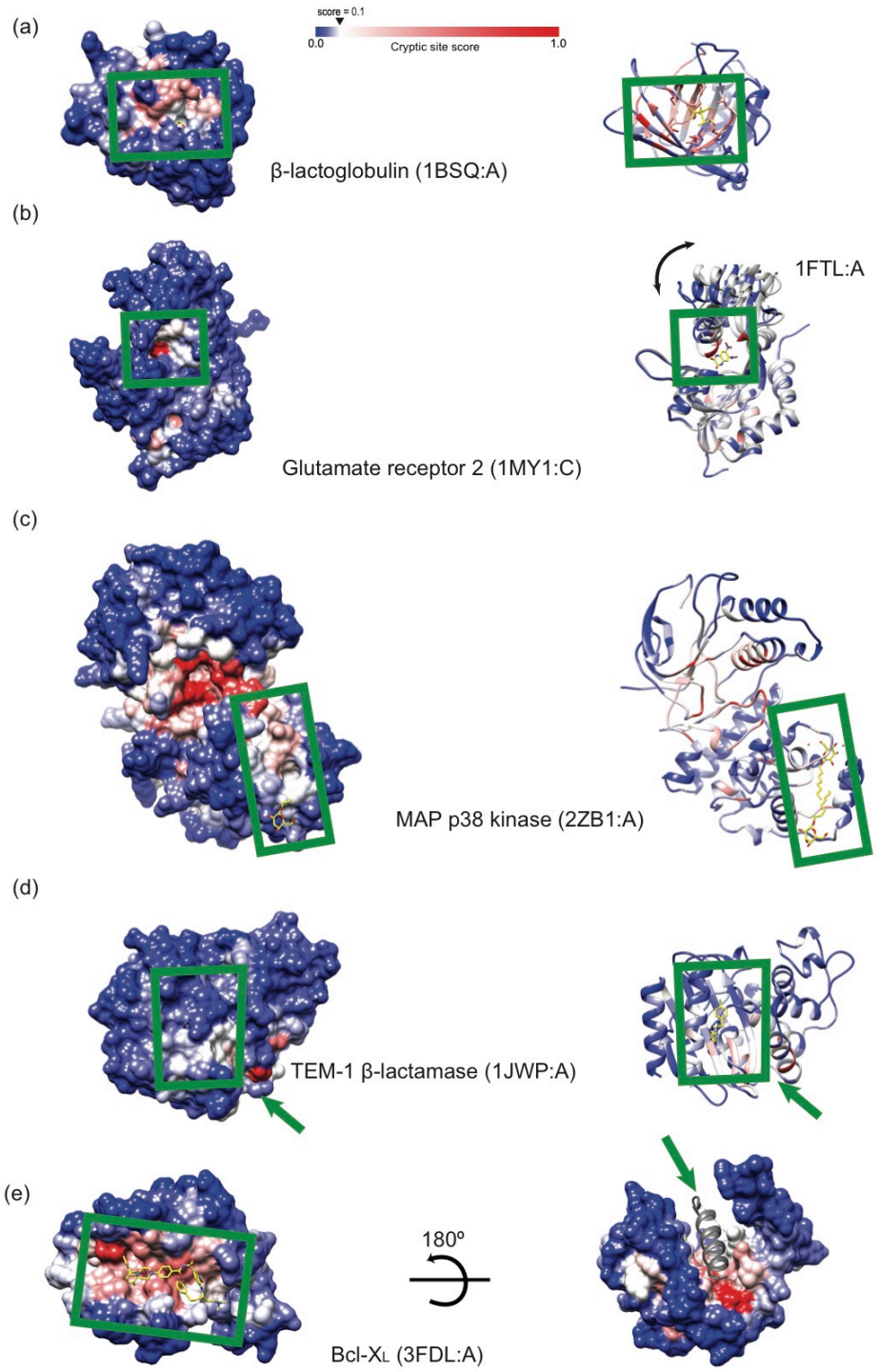

Figure S7: Examples of accurate predictions, shown in surface and ribbon representation of *apo* conformations. Ligands (yellow sticks) are superposed from the alignments with the *holo* conformations. (*a*) 94% of cryptic site residues are predicted accurately in the β-lactoglobulin (PDB ID: 1BSQ). To demonstrate the ability of our method to correctly identify the cryptic binding site residues, a few residues on β-strands are shown as sticks. These residues are predicted as a cryptic site with high scores and correctly point towards the binding site, whereas the neighboring residues on the β-strands that point in the other direction have low scores (the same pattern is observed in other proteins where a cryptic binding site includes β-strands). (*b*) Binding to the cryptic site of glutamate receptor 2 requires domain opening (indicated by a black double-headed arrow). The ribbon representation shows both the *apo* (PDB ID: 1MY1) and *holo* conformations (in grey; PDB ID: 1FTL). (*c*) Binding into the cryptic site of MAP p38 kinase requires α-helix translocation (**Fig. 1**). (*d*) Cryptic site residues that are not solvent accessible in the *apo* conformation of TEM-1 β-lactamase are correctly predicted (red patches on β-strands). (*e*) Cryptic site in Bcl-X_L is located at the protein-protein interface. The predictive model predicts another cryptic site at the interface of the Bcl-X_L core and its terminal α-helix (denoted by green arrow). Proteins are shown in scale.

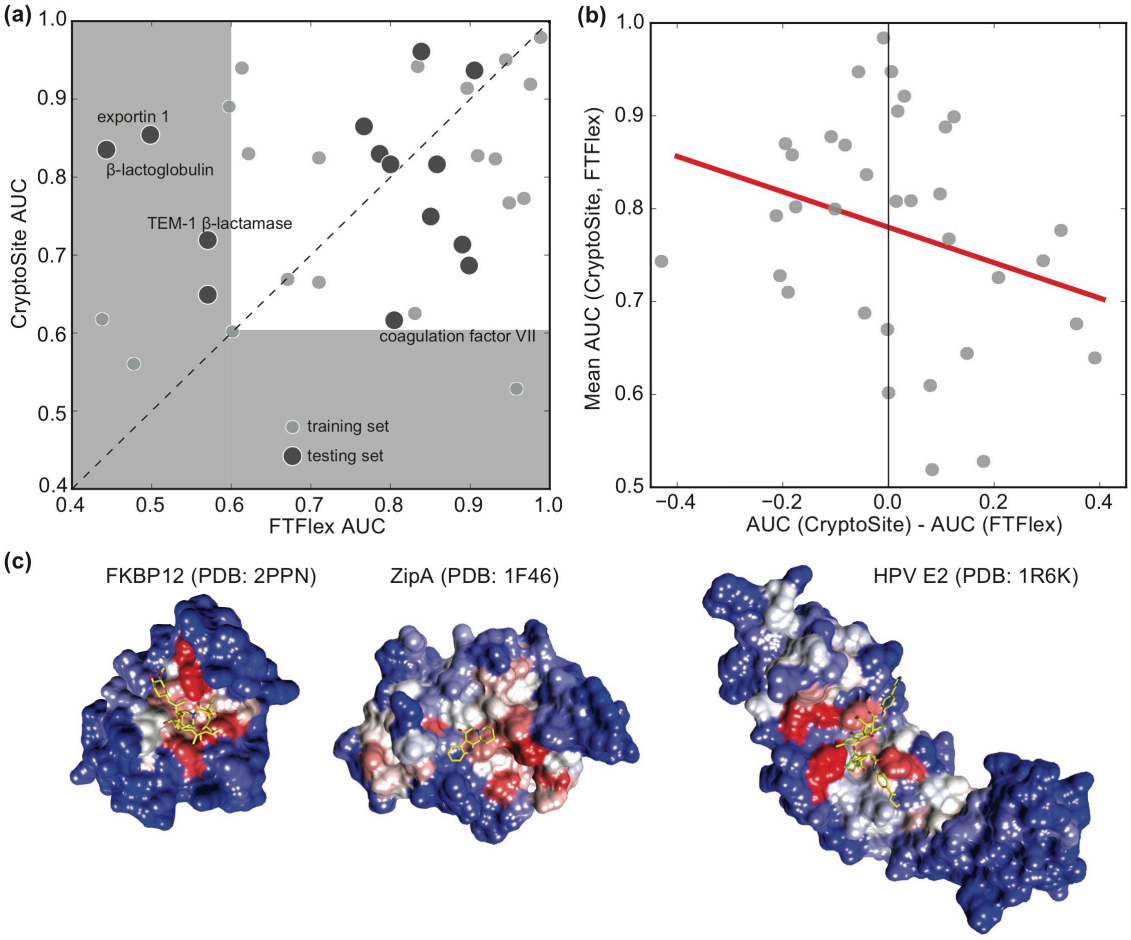

Figure S8: (*a*) Evaluation of the CryptoSite and FTFlex (webserver version) (35). Areas under the ROC curve (AUCs) demonstrate higher accuracy of CryptoSite, especially when a cryptic site is buried (β-lactoglobulin and TEM-1 β-lactamase) or when it resides in a large protein (exportin 1). While more than half of residues in the cryptic site for peptide mimetic inhibitor P5B (36) were predicted correctly (52%), a poor CryptoSite prediction of the cryptic sites in coagulation factor VII and exoenzyme C3 are due to a high false positive rate (46% and 41%). All proteins from the test set and 20 randomly chosen proteins from the training set were included in this analysis. (*b*) CrpyoSite tends to perform better than FTFlex, especially when a cryptic site is difficult to predict (*ie*, when the average accuracy of both algorithms is low). (*c*) Sample cryptic site predictions at druggable protein-protein interaction interfaces.

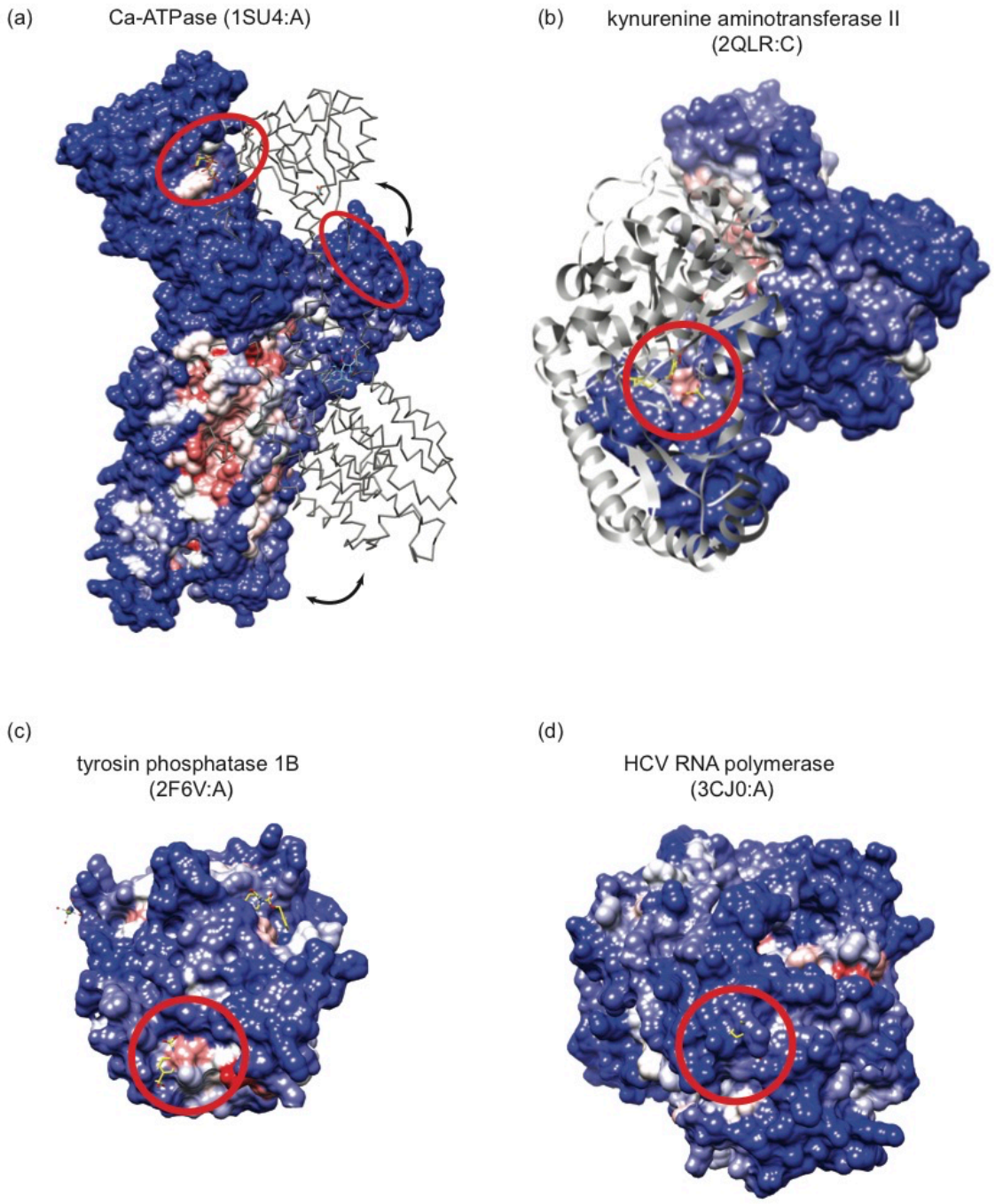

Figure S9: Four inaccurately predicted cryptic sites (marked by red ovals). (*a*) The cryptic site in Ca-dependent ATPase requires large conformational changes (denoted by black arrows and the *holo* conformation represented by grey trace), not sampled by our molecular dynamics simulations (PDB ID: 1SU4). (*b*) Cryptic site scores for the binding site residues in kynurenine aminotransferase II are higher then in the rest of the protein, but below our threshold (PDB ID: 2QLR:C), mainly because the binding site resides at an interface between two chains, only one of

which was used for the prediction (the second chain is shown in grey ribbon representation). (*c*) Similarly as in B, the cryptic binding site residues in tyrosin phosphatase 1B were predicted with scores higher than those for most of the protein, but are below our threshold for most of the binding site residues (PDB ID: 2F6V). The predictive model identifies two additional cryptic sites, one that is a site in proximity of Cys 121 and one that is unannotated site at the N-terminus. (*d*) The panel shows the structure of HCV RNA polymerase (PDB ID: 2BRK), with the incorrectly predicted cryptic site indicated. The red patch to the right of the cryptic binding site is a known cryptic site, and was predicted correctly.

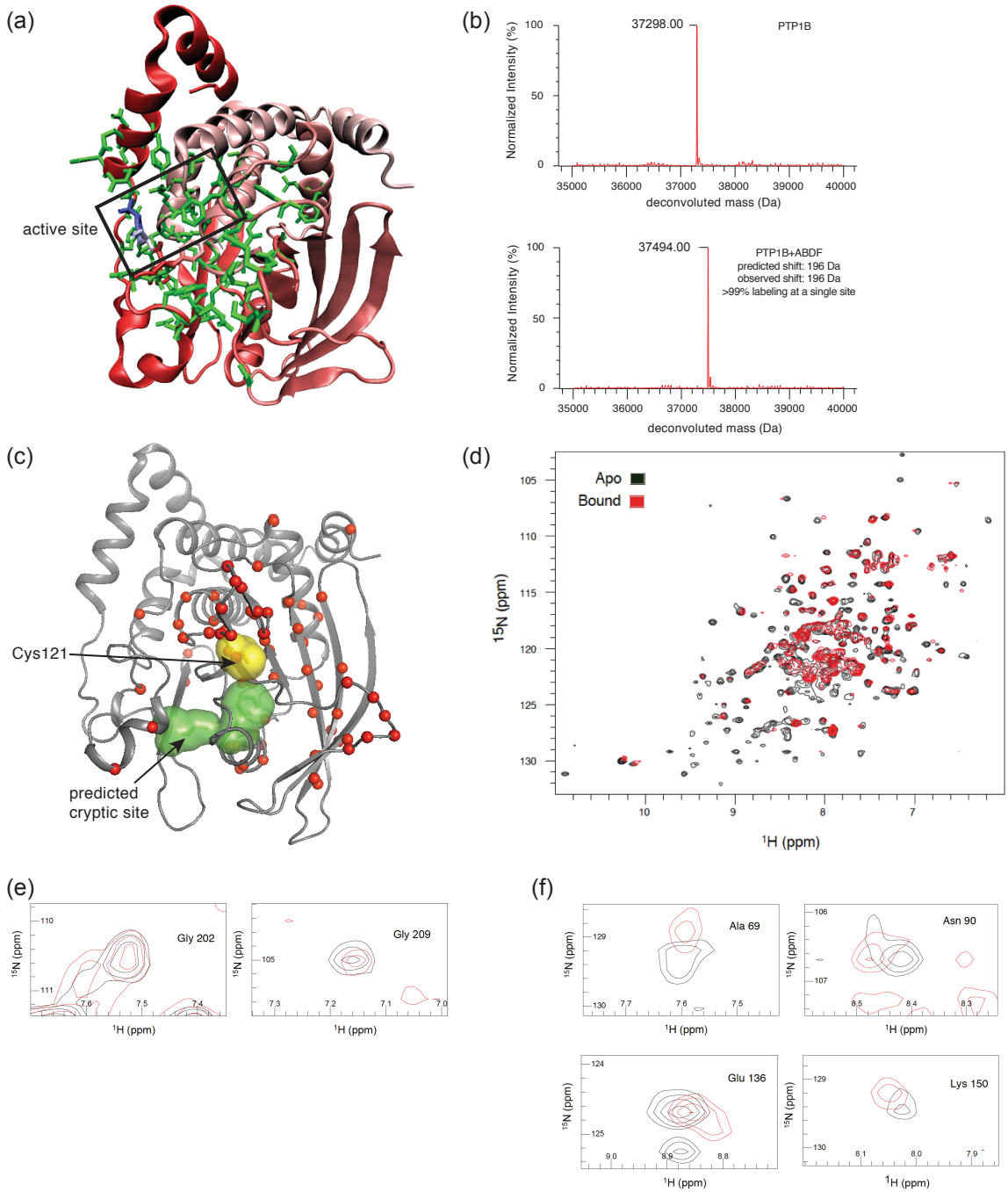

Figure S10: (*a*) Residues coupled with the active site of PTP1B are shown as green sticks (6). (*b*) Mass spectra of non-modified (top) and ABDF-modified PTP1B (bottom). The difference in mass (196) corresponds to the mass of the ABDF modification (197). (*c*) Many residues in PTP1B surrounding the predicted cryptic site (green surface) and the ABDF labeling site, Cys 121 (yellow surface), are unassigned due to broadened resonances (red spheres) (32). (*d*) Overlay of ^{1}H, ^{15}N TROSY HSQC spectra of PTP1B with (black) and without (red) labeling by ABDF. PTP1B

residues with no significant (*e*) or significant (*f*) chemical shift perturbations upon ABDF binding. Resonances are colored using the same color scheme as in *d*.

Apo	Holo	Ligand	Protein size	Site s	FPR(0.05)	TPR(0.05)	AUC(0.05)	MCC(0.05)	FPR(0.1)	TPR(0.1)	AUC(0.1)	MCC(0.1)	FPR(0.15)	TPR(0.15)	AUC(0.15)	MCC(0.15)
3CHEA	2IUZB	D1H	433	15	0.34	1.00	0.98	0.25	0.15	1.00	0.98	0.41	0.06	0.87	0.98	0.51
2AKAA	1YV3A	BIT	776	19	0.28	1.00	0.97	0.24	0.11	1.00	0.97	0.41	0.04	0.68	0.97	0.43
2GFCA	2JDSA	L20	350	26	0.27	1.00	0.97	0.41	0.11	0.96	0.97	0.59	0.06	0.85	0.97	0.62
1ALBA	1LICA	HDS	131	20	0.43	1.00	0.95	0.41	0.24	0.95	0.95	0.53	0.16	0.90	0.95	0.59
1NEPA	2HKAC	C3S	130	24	0.22	1.00	0.95	0.63	0.09	0.83	0.95	0.68	0.04	0.54	0.95	0.58
3MN9A	3EKSA	CY9	374	20	0.22	0.95	0.94	0.37	0.04	0.80	0.94	0.62	0.01	0.50	0.94	0.58
1ALVA	1NX3A	ISA	173	11	0.39	1.00	0.94	0.30	0.22	1.00	0.94	0.43	0.09	0.82	0.94	0.51
2YQCA	2YQSA	UD1	486	31	0.25	1.00	0.94	0.40	0.07	0.71	0.94	0.51	0.03	0.29	0.94	0.31
2QFOB	2WI7A	2KL	207	20	0.38	0.95	0.94	0.34	0.12	0.95	0.94	0.61	0.07	0.80	0.94	0.62
1RTCA	1BR6A	PT1	268	15	0.41	1.00	0.93	0.27	0.19	0.93	0.93	0.41	0.08	0.60	0.93	0.38
1RDWX	1J6ZA	RHO	375	10	0.18	0.80	0.93	0.25	0.06	0.70	0.93	0.39	0.01	0.50	0.93	0.49
1TQOA	1TR5A	THP	138	16	0.27	0.94	0.92	0.45	0.05	0.63	0.92	0.58	0.00	0.25	0.92	0.48
3PUWE	1FQCA	GLO	378	11	0.39	0.91	0.92	0.18	0.19	0.82	0.92	0.26	0.10	0.82	0.92	0.36
3L7UC	2HVDC	ADP	172	16	0.26	0.88	0.91	0.39	0.12	0.81	0.91	0.52	0.07	0.50	0.91	0.40
1R1WA	3F82A	353	312	26	0.37	0.96	0.91	0.33	0.13	0.77	0.91	0.46	0.06	0.58	0.91	0.46
1G4EB	1G67B	POP/TZP	227	27	0.29	0.93	0.91	0.43	0.08	0.59	0.91	0.48	0.04	0.44	0.91	0.46
3F74C	3BQMC	BQM	181	20	0.26	0.90	0.89	0.43	0.07	0.60	0.89	0.49	0.02	0.40	0.89	0.50
1MY1C	1FTLA	DNQ	263	13	0.32	1.00	0.89	0.31	0.12	0.62	0.89	0.30	0.04	0.15	0.89	0.12
2WGQB	1D6YB	HY1	727	11	0.30	0.91	0.88	0.16	0.13	0.64	0.88	0.18	0.05	0.27	0.88	0.13
1DUBD	1EY3F	DAK	261	25	0.21	0.76	0.86	0.37	0.07	0.68	0.86	0.54	0.05	0.44	0.86	0.42
1PZTA	1PZYD	UDP	286	16	0.33	0.88	0.86	0.26	0.13	0.69	0.86	0.35	0.08	0.38	0.86	0.23
1XMGB	1XVCA	5BR	527	11	0.33	1.00	0.84	0.20	0.18	0.73	0.84	0.20	0.09	0.00	0.84	-0.04
1IMFA	1IMBB	LIP	277	18	0.28	0.78	0.84	0.27	0.11	0.44	0.84	0.24	0.07	0.39	0.84	0.27
2AX9A	2PIQA	RB1	256	9	0.49	0.89	0.83	0.15	0.24	0.78	0.83	0.22	0.17	0.78	0.83	0.28
1EXMA	1HA3B	MAU	405	30	0.27	0.80	0.83	0.30	0.10	0.43	0.83	0.26	0.05	0.23	0.83	0.20
3KQAB	3LTHA	UD1	419	25	0.27	0.80	0.83	0.28	0.08	0.56	0.83	0.36	0.02	0.44	0.83	0.46
2BF3A	3DHHE	BML	92	9	0.35	0.89	0.82	0.33	0.11	0.44	0.82	0.29	0.02	0.33	0.82	0.41
1CLLA	1CTRA	TFP	148	15	0.25	0.73	0.82	0.32	0.08	0.60	0.82	0.46	0.07	0.40	0.82	0.33
3H5RA	3H9JD	APC	353	10	0.22	0.60	0.82	0.15	0.08	0.10	0.82	0.01	0.03	0.10	0.82	0.06
1NI6D	3HOKB	Q80	224	21	0.50	0.86	0.82	0.21	0.28	0.76	0.82	0.30	0.17	0.71	0.82	0.39
1QLWB	2WKWB	W22	328	18	0.32	0.83	0.82	0.25	0.13	0.50	0.82	0.23	0.06	0.06	0.82	-0.01
1HAGE	1GHYH	121	295	21	0.34	0.90	0.82	0.30	0.05	0.33	0.82	0.27	0.03	0.19	0.82	0.22
1OK8A	1OKEB	BOG	394	17	0.31	0.88	0.81	0.24	0.16	0.53	0.81	0.20	0.07	0.29	0.81	0.17
3DXNA	3HZTA	J60	287	16	0.31	0.81	0.81	0.24	0.17	0.69	0.81	0.30	0.07	0.50	0.81	0.33
1K5HC	2EGHB	FOM	398	16	0.44	0.88	0.80	0.17	0.18	0.50	0.80	0.16	0.07	0.25	0.80	0.13
1HKAA	3IP0A	HHS	158	17	0.45	0.71	0.80	0.16	0.23	0.71	0.80	0.32	0.08	0.65	0.80	0.51
2BU8A	2BU2A	TF1	394	15	0.48	0.93	0.79	0.17	0.22	0.73	0.79	0.23	0.14	0.40	0.79	0.14
2OHGA	2OHVA	NHL	264	20	0.33	0.75	0.79	0.23	0.12	0.50	0.79	0.28	0.05	0.35	0.79	0.30
2BLSB	3GQZA	GF7	358	10	0.36	0.70	0.79	0.11	0.15	0.60	0.79	0.20	0.06	0.40	0.79	0.21
1RHBA	2W5KB	NDP	124	12	0.13	0.67	0.79	0.41	0.08	0.33	0.79	0.24	0.04	0.08	0.79	0.05
1ADEA	1CIBA	IMP	431	24	0.36	0.71	0.77	0.17	0.13	0.25	0.77	0.08	0.04	0.13	0.77	0.10
1KS9A	2OFPA	PAF	291	10	0.41	1.00	0.77	0.22	0.19	0.30	0.77	0.05	0.10	0.30	0.77	0.12
1H09A	2IXUA	MU2	338	10	0.23	0.50	0.77	0.10	0.09	0.10	0.77	0.01	0.03	0.10	0.77	0.06
3BL9B	3BL7A	DD1	301	15	0.34	0.80	0.76	0.21	0.15	0.27	0.76	0.07	0.07	0.27	0.76	0.15
1BP5A	1RYOA	OXL	337	12	0.30	0.67	0.76	0.15	0.11	0.33	0.76	0.13	0.05	0.33	0.76	0.23
2Q8FA	2Q8HA	TF4	407	11	0.42	0.91	0.76	0.16	0.20	0.36	0.76	0.07	0.11	0.27	0.76	0.08
2IYTA	2IYQA	ADP/SKM	184	30	0.34	0.70	0.75	0.27	0.19	0.47	0.75	0.24	0.10	0.37	0.75	0.27
1EX6A	1GKYA	5GP	186	17	0.36	0.71	0.75	0.21	0.17	0.53	0.75	0.26	0.07	0.35	0.75	0.27
1RRGA	1S9DA	AFB	181	10	0.53	0.90	0.75	0.17	0.27	0.70	0.75	0.21	0.13	0.20	0.75	0.04
1ECJD	1ECCB	PCP	504	22	0.20	0.64	0.74	0.21	0.06	0.36	0.74	0.25	0.01	0.23	0.74	0.30
2CM2A	2H4KA	509	304	17	0.24	0.53	0.73	0.15	0.08	0.18	0.73	0.07	0.01	0.12	0.73	0.19
1NUWA	1EYJB	AMP	337	17	0.28	0.65	0.73	0.18	0.17	0.24	0.73	0.04	0.09	0.06	0.73	-0.02
3CJ0A	2BRLA/3FQKB	POO/79Z	576	39	0.28	0.62	0.73	0.18	0.08	0.21	0.73	0.11	0.03	0.08	0.73	0.07
1UK2A	2GZ7A	D3F	306	17	0.48	0.82	0.70	0.16	0.17	0.47	0.70	0.18	0.06	0.18	0.70	0.10
2WGBA	2V57A	PRL	190	13	0.55	0.92	0.70	0.19	0.33	0.62	0.70	0.15	0.18	0.23	0.70	0.03
1HOOB	1CIBA	HDA	431	14	0.32	0.64	0.69	0.12	0.11	0.36	0.69	0.13	0.03	0.00	0.69	-0.03
1FA9A	1L5SB	URC	846	10	0.23	0.60	0.69	0.10	0.07	0.20	0.69	0.06	0.04	0.10	0.69	0.03
1W50A	3IXJC	586	411	35	0.35	0.60	0.69	0.15	0.16	0.34	0.69	0.14	0.08	0.17	0.69	0.09
3B7DE	2AL4F	CX6	261	9	0.38	0.67	0.68	0.11	0.17	0.22	0.68	0.02	0.06	0.00	0.68	-0.05
1PKLB	3HQPP	ATP/FDP/OXL	499	48	0.25	0.38	0.68	0.08	0.09	0.23	0.68	0.13	0.05	0.13	0.68	0.10
1K3FB	1U1DF	181	253	16	0.47	0.69	0.68	0.10	0.13	0.31	0.68	0.13	0.04	0.13	0.68	0.10
1FVRA	2OO8X	RAJ	327	23	0.43	0.65	0.68	0.11	0.19	0.35	0.68	0.10	0.06	0.17	0.68	0.11
3HQDA	1Q0BB	NAT	369	16	0.35	0.56	0.67	0.09	0.16	0.31	0.67	0.08	0.10	0.13	0.67	0.02
2ZB1A	2NPQA	BOG	360	23	0.44	0.78	0.67	0.17	0.23	0.30	0.67	0.04	0.14	0.17	0.67	0.03
3PEOG	2BYSJ	LOB	228	15	0.39	0.80	0.67	0.21	0.27	0.27	0.67	0.00	0.16	0.13	0.67	-0.02
2BRKA	2GIRB	NN3	536	13	0.35	0.38	0.65	0.01	0.15	0.15	0.65	0.00	0.07	0.00	0.65	-0.04
2H4EB	3CFNB	2AN	127	10	0.44	0.60	0.64	0.09	0.25	0.60	0.64	0.21	0.16	0.20	0.64	0.03
2CGAB	1AFQC	0FG	245	13	0.33	0.46	0.64	0.06	0.14	0.31	0.64	0.10	0.07	0.08	0.64	0.01
1XCGB	1OW3B	GDP	178	22	0.37	0.59	0.64	0.15	0.22	0.09	0.64	-0.10	0.14	0.05	0.64	-0.09
1FXXA	3HL8A	BBP	482	12	0.42	0.58	0.64	0.05	0.17	0.25	0.64	0.04	0.09	0.00	0.64	-0.05
4AKEB	1ANKB	ANP	214	30	0.48	0.60	0.63	0.08	0.27	0.40	0.63	0.10	0.15	0.27	0.63	0.11
3NNUA	3HL7A	I46	354	16	0.47	0.75	0.62	0.12	0.25	0.31	0.62	0.03	0.16	0.13	0.62	-0.02
1A8IA	2IEGB	FRY	842	11	0.25	0.45	0.60	0.05	0.06	0.18	0.60	0.06	0.01	0.09	0.60	0.07
1SWXA	2EUMA	LAT	209	22	0.48	0.64	0.58	0.10	0.28	0.36	0.58	0.05	0.13	0.14	0.58	0.01
2QLRC	3DC1A	AKG	425	6	0.34	0.17	0.56	-0.04	0.16	0.17	0.56	0.00	0.07	0.17	0.56	0.04
2F6VA	1T49A	892	298	14	0.24	0.43	0.56	0.09	0.05	0.07	0.56	0.02	0.01	0.00	0.56	-0.02
1G24D	1GZFC	NIR	211	17	0.41	0.35	0.53	-0.03	0.12	0.24	0.53	0.10	0.02	0.12	0.53	0.16
1SU4A	3FGOB	ACP	994	17	0.29	0.06	0.49	-0.07	0.16	0.06	0.49	-0.04	0.09	0.00	0.49	-0.04
2AIRH	1ZA1D	CTP	153	20	0.47	0.35	0.47	-0.08	0.38	0.15	0.47	-0.16	0.30	0.10	0.47	-0.15
		Average:	335.13	17.7	0.34	0.75	0.77	0.21	0.15	0.49	0.77	0.22	0.07	0.30	0.77	0.20

Table S1: Training set. The table lists the *apo* and *holo* PDB identifiers, ligands that bind the cryptic sites, protein lengths, the number of residues in cryptic sites, the false positive rates (FPR), true positive rates (TPR), as well as the Matthews correlation coefficient (MCC) and the area under the ROC curve (AUC) from the leave-one-out cross-validation for 3 different CryptoSite score thresholds (0.05, 0.1, and 0.15).

Number	Feature	Description	-log10(P-value)
1	CNC_mean	the average pocket score in the MD snaphots	138.628175
2	CNS	the percentage of sMD snapshots with pocket score > 0.4	134.401307
3	CNC_std	standard deviation of the pocket scores in the DM snapshots	120.958142
4	CN5_std	standard deviation of the top 5% pocket scores in the DM snapshots	94.412299
5	CN5_mean	the average of the top 5% pocket scores in the MD snaphots	91.872607
6	SQC***	sequence conservation	67.529399
7	SQCn***	sequence conservation of neighbors	55.087905
8	PRT_mean	the average protrusion in the MD snaphots	53.482597
9	PRT***	protrusion	51.04202
10	PatchMap (PTM)	fragment docking	49.510374
11	CVX_mean	the average convexity in the MD snaphots	49.174305
12	SAS30_mean	the average surface accessiblity area in the MD snaphots (sphere radius - 3 Å)	47.572171
13	SAS30_std	standard deviation of the srufface accessilbity areas in the DM snapshots (sphere radius - 3 Å)	39.999809
14	CVXn***	convexity of neighbors	36.427921
15	SAS14_mean	the average surface accessiblity area in the MD snaphots (sphere radius - 1.4 Å)	35.392587
16	CNCn***	pocket score of neighbors	33.397185
17	PRTn***	protrusion of neighbors	31.561099
18	CNC***	pocket score	28.865335
19	SASn***	surface accessibility area of neighbors	26.864333
20	CVX_std	standard convexity of the protrusion in the DM snapshots	25.542429
21	CVX***	convexity	25.151335
22	SAS14_std	standard deviation of the srufface accessilbity areas in the DM snapshots (sphere radius - 1.4 Å)	25.112301
23	SAS***	surface accessibility area	17.392102
24	QI_std	standard deviation of the Qi in the DM snapshots	12.709686
25	RESN5	number of neighbor residues within 5 Å	11.652929
26	En***	percentage of strand residues in neighborhood	11.114616
27	HYDn***	hydrophobicity of neighbors	10.171353
28	HYD***	hydrophobicity	9.935696
29	Hn***	percentage of alpha-helix residues in neighborhood	9.931429
30	CHRn***	charge density of neighbors	9.160735
31	QI_mean	the average Qi in the MD snaphots	7.743785
32	CHR***	charge density	6.78546
33	LSE1	local structural entropy (sliding window over 5 residues)	4.879514
34	ATM5	number of atoms in neighbor residues within 5 Å	4.391877
35	Gn***	percentage of 3-10 helix residues in neighborhood	3.873649
36	Un***	percentage of disordered residues in neighborhood	3.051307
37	WT_ROT5	weighted number of side chain rotatable bonds in neighbor residues within 5 Å	2.887659
38	TDSN5	changes in side-chain conformational entropy in neighbor residues within 5 Å	2.771082
39	Sn***	percentage of bend residues in neighborhood	2.720204
40	PCKn***	packing of neighbors	2.582132
41	LSE2	local structural entropy (no sliding window)	2.568867
42	ATM4	number of atoms in neighbor residues within 4 Å	2.2857
43	NBG***	number of neighbor residues	2.172065
44	ROT5	total number of side chain rotatable bonds in neighbor residues within 5 Å	2.161973
45	BFC***	B-factor	2.133939
46	RESN4	number of neighbor residues within 4 Å	1.983735
47	SSE***	secondary structure element	1.737934
48	BFCn***	B-factor of neighbors	1.550362
49	Res***	amino acid	1.408591
50	PRT_std	standard deviation of the protrusion in the DM snapshots at	0.935858
51	WT_ROT4	weighted number of side chain rotatable bonds in neighbor residues within 4 Å	0.851296
52	Tn***	percentage of turn residues in neighborhood	0.618936
53	D2S***	distance to the surface	0.50163
54	PCK***	packing	0.443524
55	ROT4	total number of side chain rotatable bonds in neighbor residues within 4 Å	0.341533
56	TDSN4	changes in side-chain conformational entropy in neighbor residues within 4 Å	0.091504
57	Bn***	percentage of beta-bridges residues in neighborhood	0.028279
58	In***	percentage of pi-helix residues in neighborhood	0

Table S2: List of residue-based features. The last column lists P-values from Kolmogorov-Smirnov two-sample test or χ^2 test (for amino acid type and secondary structure element), used to compare the distributions of feature values based on cryptic site residues and the rest of a protein. *** denotes features used to estimate the size of the druggable proteome.

	Cryptic sites	Binding pockets	Random surface patches	Cryptic-Pocket	Cryptic-Random patch	Pocket-Random patch
Feature	**Mean value or Count**	**Mean value or Count**	**Mean value or Count**	**P-value**	**P-value**	**P-value**
SAS	2.886807822	2.731434862	5.22383845	0.187219991	1.42E-14	8.42E-19
PRT	170.0875544	182.8012748	121.3798615	0.007803677	1.31E-18	7.43E-26
CVX	2.394675569	1.902770509	7.120342898	0.820430696	2.53E-13	2.22E-17
CNC	0.071621963	0.420939626	0.004202988	1.67E-31	7.13E-25	1.10E-52
HYD	0.094858564	-0.001179616	0.267756205	0.008363064	0.000532374	3.22E-10
CHR	0.001105976	-0.004050565	-0.000750921	0.000304239	0.122304442	0.03658217
SQC	-0.42691919	-0.312434552	-0.062681622	0.111788161	2.19E-09	9.25E-05
PCK	4.177631399	4.246615688	3.992333988	0.063721088	0.004308327	1.87E-06
BFC	-0.032600774	-0.2210273	0.771607347	0.019658	0.060555402	2.44E-06
NBG	8.339503755	8.539359978	7.206137332	0.16872252	4.92E-07	3.28E-11
Residue						
LEU	128	139	110	0.835147141	0.271267171	0.165711239
GLY	121	132	109	0.812985378	0.480232422	0.308705803
VAL	100	111	75	0.739032583	0.067240382	0.02429225
ARG	94	47	70	1.84E-05	0.070481153	0.01666201
ILE	87	76	72	0.254184804	0.270741145	0.948245883
ALA	84	93	126	0.784861953	0.002766313	0.007087505
TYR	82	96	36	0.510847957	2.69E-05	8.44E-07
PHE	79	122	49	0.006905002	0.009679412	1.18E-07
ASP	75	61	82	0.148274468	0.592555688	0.038632091
GLU	73	58	90	0.121195589	0.181928016	0.002920559
SER	71	99	101	0.072561836	0.019930182	0.622033281
THR	70	82	77	0.551024726	0.582856767	0.964431832
LYS	65	48	89	0.070077904	0.051296781	0.000119178
MET	48	50	41	0.957579583	0.538391291	0.565000461
PRO	47	57	68	0.524704412	0.052279499	0.214124532
ASN	40	48	61	0.605314722	0.039313494	0.139626071
TRP	35	37	22	0.921612329	0.113822188	0.105092996
HIS	34	44	36	0.419836077	0.88179575	0.592848967
GLN	32	34	45	0.933271418	0.156768786	0.161684188
CYS	20	23	17	0.888678343	0.755752726	0.538989285
Secondary structure						
B	15	25	13	0.203306732	0.862956435	0.105287567
E	332	306	157	0.064145816	8.35E-18	7.37E-12
G	73	73	78	0.818560927	0.706917362	0.4864712
H	373	421	439	0.260888628	0.00472031	0.089119651
S	146	152	179	0.97314657	0.050900685	0.037981829
T	133	155	183	0.394166233	0.002782715	0.033510041
U	313	325	327	0.886877711	0.49626038	0.380454587

Table S3: Comparison of cryptic sites, binding pockets, and random protein surface patches. The distributions of residue-based feature values were compared using Kolmogorov-Smirnov test (P-values reported), except for amino-acid type and secondary structure element counts, which were compared using the χ^2 test.

	Sequence Identity			
Query	**30-40**	**50-60**	**70-80**	**90-100**
1ALBA	1GGLA	4A60A	3RSWA	3RZYA
2BLSB	3WS2A	2QZ6A	1FR1A	4OKPA
1BSQA	4R0BA	1EXSA	3KZAA	1YUPA
2QFOB		2O1WA	3K60B	2YEGA
1CLLA		4DS7A	1GGZA	3CLNA
1DUBD	3T8AA	3MOYA		2HW5F
1EXMA	1F60A			1MJ1A
1JWPA	1PIOA	1G6AA	2G2WA	4GKUA
1K3FB	3EMVA		4YJKD	4OF4A
1K5HC	2JCYA	1R0KA	3IIEA	1ONNA
1HOOB	1J4BA	3UE9A		1ADEA

Table S4: The table lists template structures used to assess the performance of CryptoSite on comparative models.

Apo	Holo	Ligand	Protein size	Site size	FPR(0.05)	TPR(0.05)	AUC(0.05)	MCC(0.05)	FPR(0.1)	TPR(0.1)	AUC(0.1)	MCC(0.1)	FPR(0.15)	TPR(0.15)	AUC(0.15)	MCC(0.15)
1E2XA	1H9GA	MYR	243	28	0.32	0.96	0.96	0.42	0.08	0.89	0.96	0.68	0.03	0.61	0.96	0.63
1MY0B	1N0TD	AT1	263	18	0.31	1.00	0.94	0.36	0.10	0.78	0.94	0.49	0.02	0.22	0.94	0.28
1ZAHB	2OT1D	N3P	363	9	0.32	1.00	0.86	0.22	0.12	0.56	0.86	0.20	0.03	0.11	0.86	0.08
4HB2C	4HATC	LMB	1023	22	0.19	0.68	0.85	0.18	0.04	0.32	0.85	0.19	0.01	0.09	0.85	0.12
1BSQA	1GX8A	RTL	162	16	0.44	0.88	0.83	0.26	0.21	0.75	0.83	0.37	0.12	0.63	0.83	0.41
2GPOA	1S9QB	CHD	230	28	0.46	1.00	0.83	0.36	0.26	0.89	0.83	0.44	0.14	0.32	0.83	0.16
3FDLA	2YXJA	N3C	158	24	0.43	0.83	0.82	0.29	0.22	0.63	0.82	0.33	0.10	0.50	0.82	0.40
1B6BA	1KUVA	CA5	174	32	0.39	0.91	0.82	0.40	0.23	0.72	0.82	0.40	0.11	0.47	0.82	0.36
1KZ7D	1GRNA	AF3	188	13	0.54	0.92	0.75	0.20	0.29	0.69	0.75	0.22	0.13	0.31	0.75	0.13
1JWPA	1PZOA	CBT	263	22	0.33	0.68	0.72	0.20	0.10	0.27	0.72	0.15	0.02	0.00	0.72	-0.04
3GXDB	2WCGA	MT5	497	18	0.40	0.67	0.71	0.10	0.12	0.39	0.71	0.15	0.02	0.28	0.71	0.27
1BNCB	2V5AA	LZL	449	18	0.17	0.61	0.69	0.22	0.03	0.17	0.69	0.14	0.01	0.00	0.69	-0.02
1Z92A	1PY2A	FRH	133	15	0.55	0.73	0.65	0.12	0.18	0.40	0.65	0.17	0.10	0.20	0.65	0.10
1JBUH	1WUNH	P5B	254	23	0.46	0.61	0.62	0.09	0.13	0.22	0.62	0.07	0.04	0.17	0.62	0.16
		Average:	314.29	20.43	0.38	0.82	0.79	0.24	0.15	0.55	0.79	0.29	0.06	0.28	0.79	0.22

Table S5: Test set. The table lists the *apo* and *holo* PDB identifiers, ligands that bind the cryptic sites, protein lengths, the number of residues in cryptic sites, the false positive rates (FPR), true positive rates (TPR), as well as the Matthews correlation coefficient (MCC) and the area under the ROC curve (AUC) from the leave-one-out cross-validation for 3 different CryptoSite score thresholds (0.05, 0.1, and 0.15).