


RESEARCH ARTICLE

Open Access



Mapping the 3D structures of small molecule binding sites

Joshua Meyers, Nathan Brown* and Julian Blagg* 

Abstract

Background: Analysis of the 3D structures of protein–ligand binding sites can provide valuable insights for drug discovery. Binding site comparison (BSC) studies can be employed to elucidate the function of orphan proteins or to predict the potential for polypharmacology. Many previous binding site analyses only consider binding sites surrounding an experimentally observed bound ligand.

Results: To encompass potential protein–ligand binding sites that do not have ligands known to bind, we have incorporated fpocket cavity detection software and assessed the impact of this inclusion on BSC performance. Using fpocket, we generated a database of ligand-independent potential binding sites and applied the BSC tool, SiteHopper, to analyze similarity relationships between protein binding sites. We developed a method for clustering potential binding sites using a curated dataset of structures for six therapeutically relevant proteins from diverse protein classes in the protein data bank. Two clustering methods were explored; hierarchical clustering and a density-based method adept at excluding noise and outliers from a dataset. We introduce circular plots to visualize binding site structure space. From the datasets analyzed in this study, we highlight a structural relationship between binding sites of cationic trypsin and prothrombin, protein targets known to bind structurally similar small molecules, exemplifying the potential utility of objectively and holistically mapping binding site space from the structural proteome.

Conclusions: We present a workflow for the objective mapping of potential protein–ligand binding sites derived from the currently available structural proteome. We show that ligand-independent binding site detection tools can be introduced without excessive penalty on BSC performance. Clustering combined with intuitive visualization tools can be applied to map relationships between the 3D structures of protein binding sites.

Keywords: Binding site prediction, Binding site comparison, Mapping binding site space, Protein structure

Background

Analysis of the three-dimensional structures of proteins is integral to our understanding of the molecular machinery involved in their biological function and is increasingly enabled by the wealth of structural data available in the Protein Data Bank (PDB) [1]. In particular, the examination of functional binding sites is of importance in biological chemistry and drug discovery by rational design [2]. Here, we present a method for generating a structural map of potential small molecule binding sites derived from the currently available structural proteome [3, 4].

Evidence for the existence and location of a binding site can be built through experimental observation of protein–ligand binding events—often facilitated by protein X-ray crystallography and/or Nuclear Magnetic Resonance (NMR) spectroscopy. However, prospective computational analysis to discover novel potential binding sites requires an objective and systematic cavity detection method, for which many tools exist [5–7]. For example, *fpocket* is a widely used and freely available software that employs geometric alpha shape theory to detect cavities in protein structure coordinates [8].

A number of *Binding Site Comparison* (BSC) tools exist to quantify the structural similarity between a pair of binding sites [9–12]. BSC has been applied to suggest the function of orphan proteins and to predict the

*Correspondence: nathan.brown@icr.ac.uk; julian.blagg@icr.ac.uk
Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London SM2 5NG, UK

potential for polypharmacology [4, 12]. *SiteHopper* [13] is a recently developed BSC tool in which binding sites are represented as three-dimensional *patches* encoded with spatial information concerning the local molecular surface (shape) and chemical properties (color) of residues lining the binding site. An example binding site patch is shown in Fig. 1 for the cofactor binding site of CDK2 [14]. Binding site patches can be aligned rapidly in a pairwise manner and assessed for structural similarity based on the maximal overlap of atom-centered Gaussian functions [15]. A more detailed description of the SiteHopper BSC tool has recently been published [16].

BSC tools commonly define binding sites as the protein environment surrounding an experimentally observed bound ligand. Importantly, this definition excludes potential binding sites that have not been demonstrated to bind ligands (so called *unliganded* binding sites), thereby creating a bias towards currently exemplified

protein–ligand complexes. To address this, tools such as *CavBase* [18], *RAPMAD* [19], *IsoMIF* [20] and *TrixP* [21], have integrated binding site detection algorithms with BSC. However, to the best of our knowledge, there has been no systematic analysis of the implications for BSC performance with unliganded cavities in the dataset. To mitigate this concern, we applied a modular approach and independently validated both the cavity detection and BSC components when applied to datasets comprising both liganded and unliganded protein binding sites.

A structural mapping of protein binding sites can provide a useful tool for probing the three-dimensional structural relationships between biological macromolecules [3, 10]. Tools that aim to provide an assessment of similarity between protein–ligand binding sites include *Relibase*, a database of known protein–ligand binding sites [22]; the *sc-PDB*, an annotated database of druggable binding sites from the PDB [23]; and the *Pocketome*,

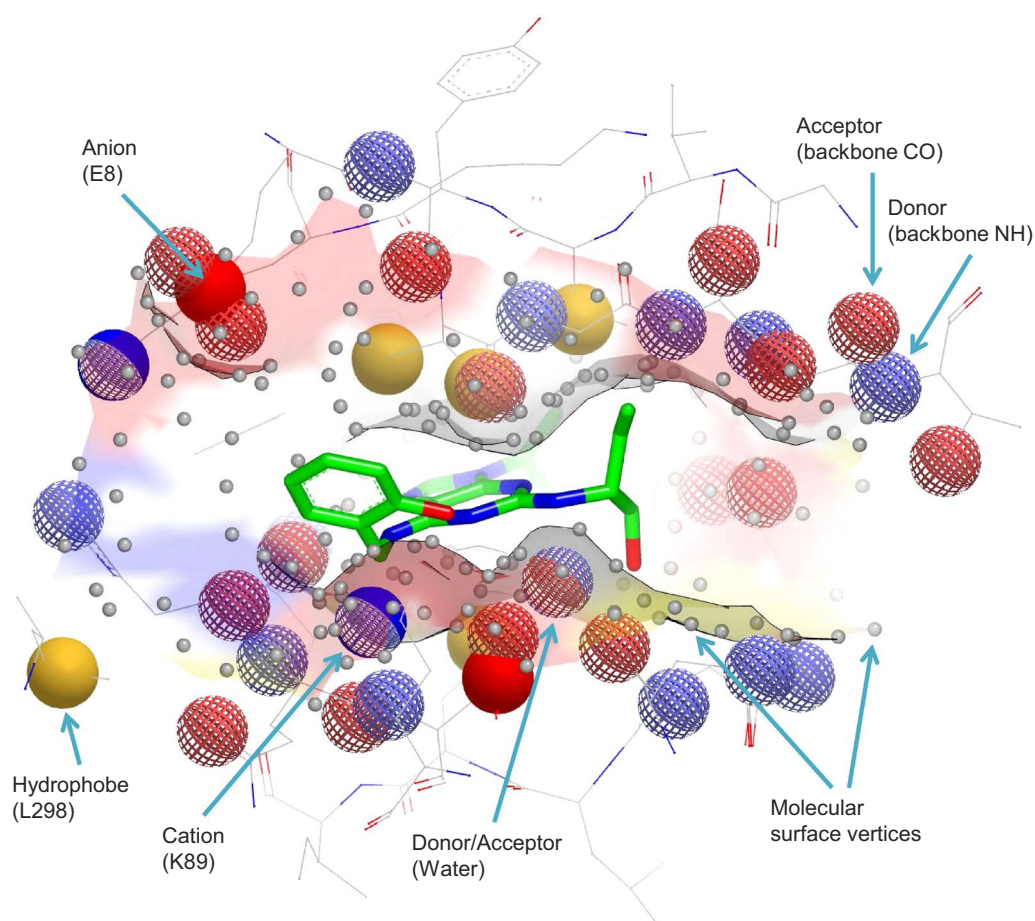


Fig. 1 SiteHopper patch exemplified for the cofactor binding site of CDK2 (PDB ID: 2A0C) defined by surface protein atoms within 4 Å of a bound ligand (shown in green). A pharmacophore model defines pseudocenters for five key interaction types: hydrogen bond donor (blue mesh), hydrogen bond acceptor (red mesh), anion (red solid), cation (blue solid) and hydrophobe (yellow). Surface vertices also encode the shape of the binding site (gray). Image produced using VIDA [17]

an encyclopedia of conformational ensembles of druggable binding sites [24]. While each of these tools provides an assessment of similarity between binding sites, the potential for identifying novel three-dimensional relationships involving currently unliganded binding sites is limited without the incorporation of objective methods for cavity detection.

The workflow presented here enables a structural mapping of potentially ligandable binding sites of the currently available structural proteome. We apply fpocket to objectively detect protein cavities and SiteHopper BSC to systematically generate pairwise structural similarities between all detected cavities. We also assess the performance of BSC incorporating all fpocket-detected cavities versus datasets only containing cavities surrounding an experimentally observed ligand. We then describe a number of clustering methods and visualization techniques for the mapping of potential binding site space. Altogether we present a validated workflow and describe challenges associated with the methodologies employed therein. In this work, we have adopted the following definitions throughout: a cavity is a surface depression identified by fpocket in static protein structure data. A potential protein–ligand binding site is one predicted by fpocket to bind small molecules, whereas a known protein–ligand binding site is one that has been experimentally shown to bind small molecules.

Methods

Protein structure datasets

Four datasets of protein structures were studied.

An ensemble of five cAMP-dependent protein kinase structures, all bound to the endogenous co-factor adenosine 5'-triphosphate (ATP) (Additional file), was retrieved from the PDB [1]. Crystallographic structure data was selected to satisfy ligand-centric quality criteria: resolution ≥ 2.7 ångströms (Å) [25], Real-Space Correlation Coefficient (RSCC) of ligand instance ≥ 0.9 , Real-Space R-factor (RSR) of ligand instance ≤ 0.15 and Occupancy-Weighted Average B-factor (OWAB) of ligand instance between 5 and 50 Å² [26]. All structures were aligned using the PDB entry 1ATP as the reference coordinates and Schrödinger's *Protein Preparation Wizard* [27] was applied to ensure consistent protonation, removal of waters and assignment of tautomers. All structures exhibit very similar conformations of the protein with a mean all-atom Root-Mean-Square Deviation (RMSD) for pairwise alignment of 1.0 Å, calculated in PyMOL using the *align* command with the *cycles* flag set to zero [28].

The *PDBBind*-refined set (2014) [29] is a curated set of 3446 high-quality, binary protein-small molecule complexes associated with measured binding affinity (K_i or K_d). This dataset was used to evaluate models for ranking

detected cavities and to determine a threshold above which detected cavities constitute potential binding sites.

The *sc-PDB* (2013) [23] is a curated database of 9283 proteins bound to drug-like ligands and was used to assess the performance of SiteHopper BSC [13].

A further dataset was manually curated from the PDB to guide the generation of a map of the structures of diverse and therapeutically relevant potential small molecule binding sites. This dataset contains 1085 crystallographically determined protein structures of the following protein targets: bromodomain-containing protein 4 (BRD4) (93), cyclin-dependent kinase 2 (CDK2) (148), estrogen receptor (52), human immunodeficiency virus-1 (HIV-1) protease (335), prothrombin (142) and cationic trypsin (315). Structures were retrieved by their respective *UniProt* [30] identifiers, except HIV-1 Protease for which structures were retrieved with 90% sequence identity (*Protein BLAST* [31], $E = 10^{-20}$) to a reference sequence [32]. Retrieved crystal structures were selected to satisfy protein-centric crystallographic quality criteria: resolution ≥ 2.5 Å, Free R-factor (Rfree) ≤ 0.3 and Diffraction Precision Index (DPI) [33] ≤ 0.5 Å [34]—calculated using DPI calculator [35]. This dataset is referred to as the *Pilot* dataset (Additional file).

All crystallographic quality descriptors were retrieved from either the PDB or *Electron Density Server* (EDS) [36] unless otherwise stated.

Binding site detection

fpocket (version 2.0) [8] was implemented for ligand-independent cavity detection using default settings with two parameter alterations; the *-r* flag was set to 3.0 (default 4.5) and the *-n* flag was set to 3 (default 2). fpocket ranks cavities according to a *Partial Least Squares* (PLS) model *Score* trained on five descriptors relating to hydrophobicity, polarity and the size of a detected binding site [8]. Cavities were detected for protein structures in the *PDBBind*-refined set (2014) [29] and an fpocket *Score* ≥ 16.8 was determined, above which cavities were considered as potential ligand-binding sites. This threshold corresponds to the *Score* above which 95% of known ligand binding sites from the *PDBBind*-refined set were identified.

Binding site comparison (BSC)

A dataset of ligand-dependent binding sites was generated using the SiteHopper *create* tool [13] where default parameters create a binding site patch within 4 Å of a specified bound ligand. This approach was followed to generate 9275 binding site patches for the *sc-PDB* (2013) database; this is referred to as the *ligCav* binding site dataset. Eight protein structures failed to yield binding site patches.

To generate ligand-independent binding site patches, surface protein atoms associated with fpocket cavities were utilized as a pseudo-ligand for input to the SiteHopper *create* tool. Binding site patches were defined as surface protein atoms lying within 0.3 Å of the fpocket surface atoms. This *site size* value was determined empirically through a number of retrieval experiments with a range of *site size* values (0.1–0.6 Å, increments of 0.1 Å). The ability of SiteHopper to identify similarity between a query estrogen receptor binding site patch and other members of the estrogen receptor in the sc-PDB (2013) was assessed using binding site patches created with varying *site sizes* (Additional file 1: Figure S1). Larger binding site patches incur a penalty in calculation time during BSC, and therefore the chosen *site size* represents a balance between computational expense and retrieval success.

The SiteHopper tool was utilized to generate binding site patches and to assess pairwise structure similarity between reference and query patches. The default SiteHopper *PatchScore* represents a summation of Tanimoto similarity coefficients [37] weighted 3:1 in favor of color similarity over shape similarity, yielding a continuous value between zero and four, conveying complete dissimilarity and perfect similarity respectively [13]. Utilization of the symmetric Tanimoto similarity coefficient causes an inherent size matching to exist between pairs of binding site patches that show high levels of structural similarity.

Retrieval analysis

The sc-PDB (2013) database [23] was utilized to assess BSC performance through a series of retrieval experiments evaluating the ability of SiteHopper to identify similar binding sites belonging to the same protein target. True positives were defined as binding site patches with the same UniProt identifier as the query patch, except for those belonging to HIV-1 protease, which were defined by sequence searching as previously described (“Methods” section). Due to the presence of multiple binding sites per protein structure, only the binding site with the highest SiteHopper PatchScore derived from a matching protein structure was considered a true positive. Reference binding site patches used as queries for retrieval experiments are shown in Additional file 1: Table S1.

Mapping binding sites

To guide mapping of the potentially ligandable binding sites of the structural proteome, an exhaustive all-against-all BSC was performed on the Pilot dataset containing 2708 binding sites generated by fpocket. A breakdown of the Pilot dataset, including the number of binding sites

detected for each protein target, is shown in Additional file 1: Table S2. The resulting matrix of (2708 × 2708) SiteHopper PatchScores was exploited to produce a clustered heat map of potential binding site space. To remove non-conserved binding sites from the dataset, patches with fewer than five pairwise SiteHopper PatchScores ≥ 2.0 were filtered out. Binding sites were first clustered within the protein targets from which they were derived using average-linkage agglomerative hierarchical clustering and the Euclidean distance measure. Subsequent clustering was performed in the same way across the global Pilot dataset. Plots were generated using *matplotlib* [38] and clustering was implemented in the *Python* programming language using the *SciPy* package [39].

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [40] was applied to cluster binding sites for each protein using $n \times n$ matrices of SiteHopper PatchScores. DBSCAN was implemented in the *Python* programming language using the *scikit-learn* machine learning toolkit [41] with range $\epsilon = 7$ and a minimum number of points per core cluster being ten. Circular plots were generated as an alternative visual tool for mapping potential binding site space using the *Circos* software package [42].

Evaluation techniques

Receiver Operating Characteristic (ROC) curves are a widely used tool employed to quantify the ability of a method to identify instances with similar characteristics to a reference (true positives). The *Area Under a Receiver Operating Characteristic* (AUROC) curve provides a measure of how well a method distinguishes between true positives and false positives in a dataset [43]. A perfect separation of all true positives from the data would result in an AUROC of 1, whilst a random classifier would be expected to distribute true positives throughout the whole dataset resulting in an AUROC of 0.5.

Often, it is the early recognition of true positives that is important [44], especially in cases where n is large and AUROC results are indistinguishable between methods. To this end, *Enrichment Factors* (EF) at 5% and the *Boltzmann-Enhanced Discrimination Receiver Operating Characteristic* (BEDROC) [45] were also calculated. An EF is the ratio of the percentage of true positives in an initial portion of a dataset, to the overall percentage of true positives in the entire dataset. Thus an EF = 1 implies no enrichment in the initial portion of the data (no early enrichment); EF < 1 implies the classifier performs worse than random at identifying true positives, and EF > 1 implies there is some quantifiable enrichment of true positives among the highest ranked data points [46]. BEDROC applies Boltzmann weighting to the AUROC calculation thereby emphasizing the initial

portion of the ROC curve—calculated using the *CROC* package [45] at $\alpha = 20$ [46, 47].

Throughout this study, the *OEChem* cheminformatics toolkit [48] was used as an interface between tools and data was handled using the *pandas* data analysis framework [49].

Results and discussion

Ligand-independent binding site detection

Typically, BSC studies make use of known binding sites characterized by surface protein atoms surrounding an experimentally observed bound ligand. To objectively consider currently unliganded binding sites, ligand-independent binding site detection tools were evaluated. Incorporation of binding site detection tools removes bias associated with utilizing currently known liganded binding sites; however, it may also introduce noise to the data through inclusion of cavities that are incapable of ligand binding. Therefore, an assessment of the noise introduced to the data by binding site detection was conducted and also of the subsequent implications for BSC performance.

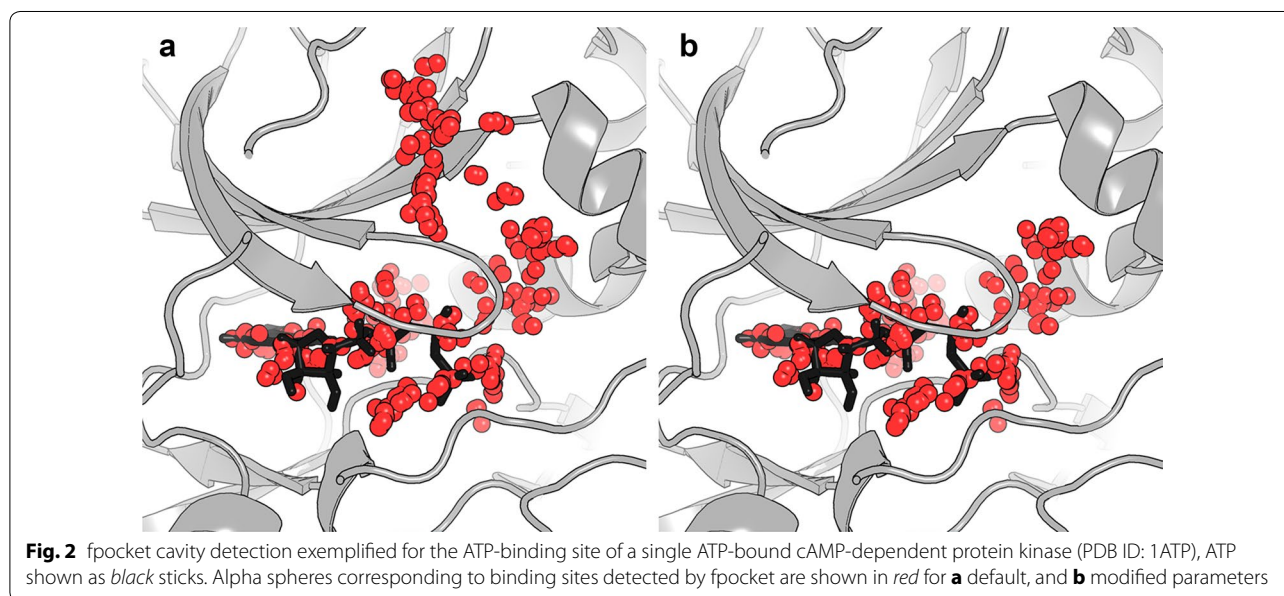
fpocket is a well-established and freely available binding site detection tool capable of operating in high-throughput and therefore applicable to large datasets of protein structure data [e.g. the sc-PDB (2013) contains 9283 structures]. *fpocket* was evaluated according to three criteria: Its ability to (1) detect cavities corresponding to functionally relevant binding sites starting from a global search of a protein structure; (2) detect similar cavities from an ensemble of structurally similar experimental structures of the same protein bound to the

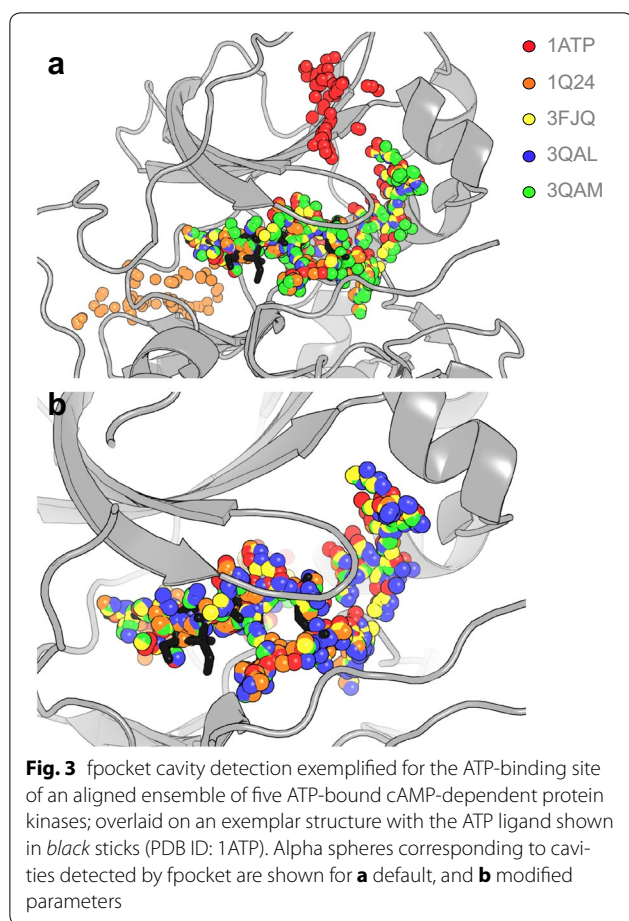
same ligand; and (3) rank and prioritize detected cavities according to their likelihood of binding small molecule ligands. Two datasets were utilized to assess these criteria: an ensemble of five ATP-bound cAMP-dependent protein kinases, and the PDBBind-refined set (2014) [29].

Initially, *fpocket* was evaluated qualitatively on a model ensemble of five structurally similar ATP-bound cAMP-dependent protein kinases for its ability to detect a cavity corresponding to the well-characterized ATP-binding site. *fpocket* implemented with default parameters tended to identify cavities extending beyond the ATP-binding domain. To attenuate this phenomenon, default parameters were modified to prevent the merging of distinct sub-pockets, yielding more concisely defined and consistent cavities amenable to BSC studies. The difference between default and modified *fpocket* parameters for an exemplar protein structure is shown in Fig. 2a, b, respectively.

Geometric binding site detection algorithms are inherently sensitive to slight variations in protein atomic coordinates. A second qualitative validation shows that *fpocket* is capable of identifying similar cavities for each member of the aligned cAMP-dependent protein kinase ensemble. This assessment is depicted for *fpocket* with both default and modified parameters in Fig. 3a, b respectively. *fpocket* cavities detected for PDB IDs 1ATP and 1Q24 show variation from the core ATP-binding cavity when implemented with default parameters; however, the modified parameters provided more consistent cavity representation.

fpocket binding site detection often identifies multiple cavities per protein structure. To reduce the complexity





associated with carrying forward multiple cavities per protein for BSC, the detected binding sites were ranked and subsequently filtered according to their potential to bind a small molecule ligand. Many other binding site detection tools rank cavities according to binding site volume as often the largest cavity corresponds to the observed ligand binding site [50]. The fpocket *Score* model aims to predict whether a cavity may contain a bound small molecule ligand and is distinct from a drug ability model since ligands are not necessarily drug-like [8].

The PDBBind-refined set (2014) [29], comprising 3446 protein structures, was processed by fpocket. The ability of the three descriptors, *Score*, *Volume* and *Druggability Score* [51], to prioritize experimentally validated ligand-binding sites over unliganded cavities was assessed. Figure 4a, b show ROC curves for fpocket with both default and modified parameters, respectively. In both cases, the PLS model *Score* was superior in discriminating between true known liganded binding sites, and those corresponding to cavities without ligands bound.

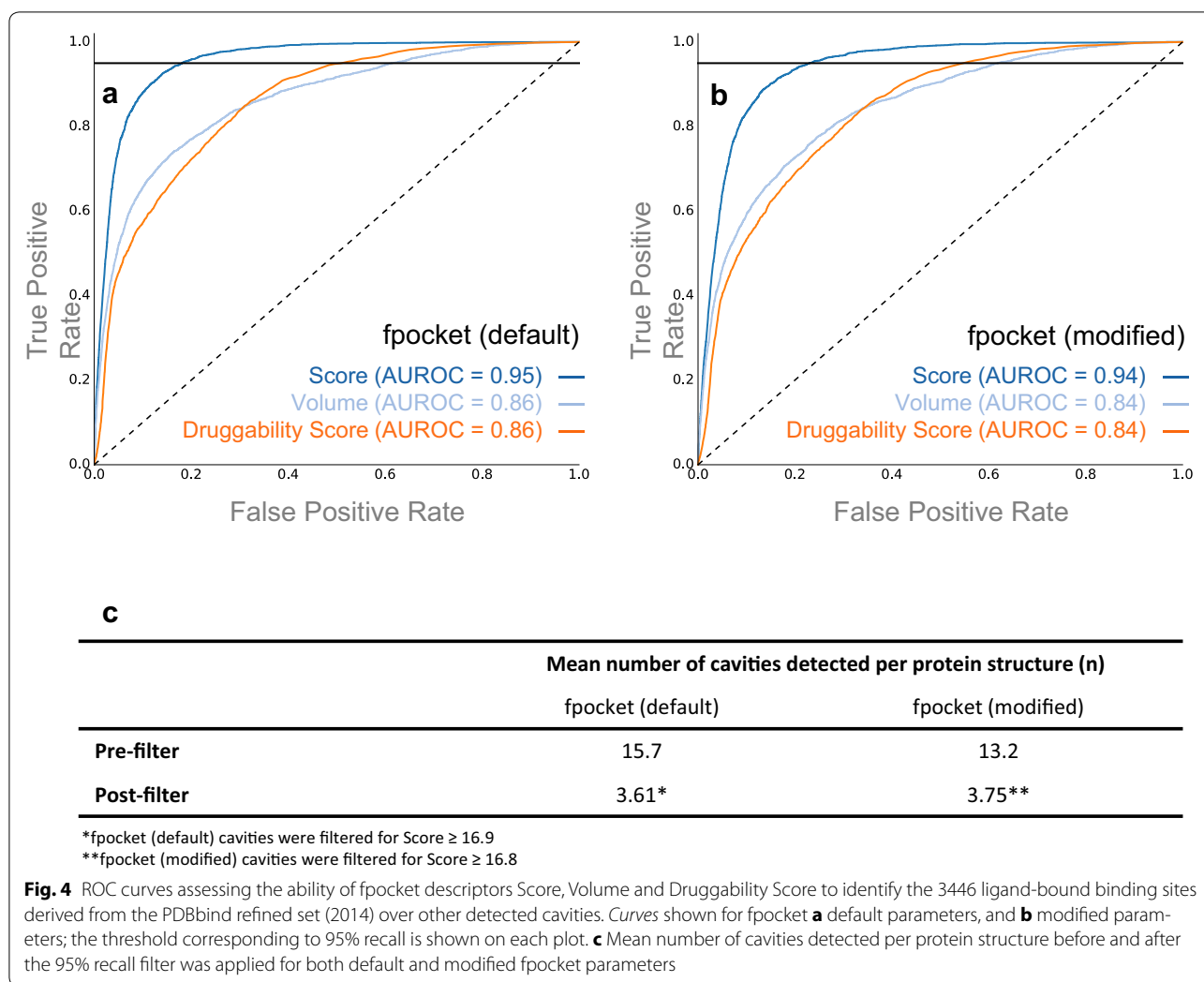
A comparison of binding site detection tools has previously found that 95% of ligand-bound sites are identifiable using geometric algorithms [7]. Accordingly, a

sensitivity threshold was determined above which 95% of observed ligand binding sites from the PDBBind-refined set were identified. This sensitivity threshold precedes a sharp increase in false positive rate and therefore excluding sites below the sensitivity threshold ensures that the number of cavities without ligand-binding potential introduced to the dataset is limited. Thus, cavities were taken forward to BSC if the fpocket *Score* is ≥ 16.8 , corresponding to the *Score* above which 95% of the ligand-bound cavities from the PDBBind-refined set are identified.

The average number of cavities identified *per* protein structure before and after applying the 95% recall filter is shown for both default and modified fpocket parameters (Fig. 4c). Although fpocket with modified parameters (yielding smaller, more concise cavities) performs slightly worse than default according to ROC analysis, the number of cavities detected per protein is comparable and therefore both parameter sets introduce similar levels of noise to the dataset. However, smaller and more consistent cavities are beneficial for BSC in terms of studying binding site similarity. Therefore, we elected to study BSC using cavities detected by fpocket with modified parameters; it can be assumed that further mentions of fpocket refer to this non-standard modified model.

Binding site comparison (BSC)

Binding site patches were generated using the SiteHopper create tool, utilizing fpocket surface protein atoms as a pseudo-ligand and isolating binding site patches from the original protein structure. The BSC performance of SiteHopper was assessed for its ability to find structural similarity between a query patch and analogous patches derived from the same protein. Six protein targets of relevance to small molecule therapy were selected from the sc-PDB (2013) database as exemplar queries (Additional file 1: Table S1). To assess the impact of incorporating binding site detection into BSC, the retrieval performance of SiteHopper was first evaluated utilizing only binding site patches defined surrounding an observed bound ligand. This ligand-dependent dataset contains 9275 known binding sites and is referred to as the ligCav dataset. The sc-PDB (2013) database was also processed by fpocket to produce a dataset of 24,345 potential binding sites (including known binding sites), for which SiteHopper patches were generated. Retrieval performance was evaluated through generation of ROC curves along with two early enrichment metrics: The EF at 5% and the BEDROC at $\alpha = 20$ [45]. Table 1 summarizes the ability of SiteHopper to detect structural similarity between analogous binding sites derived from structures of the same protein, for both the ligCav and fpocket-derived datasets.



As described above, the incorporation of fpocket cavity detection into BSC introduces the potential for noise in the binding site dataset compared to only defining binding sites surrounding observed bound ligands, and this may result in poorer retrieval performance metrics. However, in our retrieval analysis, we only observed a slight impact on BSC performance using early enrichment metrics; the AUROC enrichment remains high when compared to retrieval analyses performed using the ligCav dataset. Thus, the incorporation of fpocket objective cavity detection into BSC workflows is not associated with an unreasonable decrease in retrieval capability. In summary, we show that SiteHopper is able to identify structural similarity between potential binding sites that have been detected objectively from protein structure coordinates.

Interestingly, we observed a variation in retrieval rates across protein targets. Retrieval scores for the

acetyl-lysine binding site of BRD4 are high, likely due to the rigidity of the protein structure surrounding this site. On the contrary, EF at 5% and BEDROC for the protein kinase CDK2 are relatively poor, likely due to the flexibility and range of protein conformations exemplified by crystal structures of this protein. Upon inspection of instances where structural similarity was expected, but not assigned a high SiteHopper score, we found that, in many cases, analogous fpocket-detected potential binding sites showed structural variability. This observation highlights the importance of consistency in the binding site detection tool; for example, upon inspection of cavities detected for prothrombin, we found that overlapping but dissimilar fpocket-detected sites were extracted from very similar protein conformations. This exemplifies how the objective implementation of fpocket binding site detection can still introduce noise into the BSC workflow despite the modifications we describe.

Table 1 Evaluation of SiteHopper retrieval of binding site patches from the sc-PDB (2013) belonging to the same protein as a query patch

	N (sc-PDB)	AUROC	EF (5%)	BEDROC
ligCav				
BRD4 (<i>n</i> = 2)	15	1.00 ± 0.00	20.03 ± 0.00	1.00 ± 0.00
Carbonic anhydrase 2 (<i>n</i> = 3)	76	1.00 ± 0.00	19.77 ± 0.00	0.99 ± 0.00
CDK2 (<i>n</i> = 3)	180	1.00 ± 0.00	19.66 ± 0.07	0.97 ± 0.01
Estrogen receptor (<i>n</i> = 5)	58	1.00 ± 0.00	20.03 ± 0.00	0.96 ± 0.01
HIV-1 protease (<i>n</i> = 3)	219	0.99 ± 0.00	19.94 ± 0.00	0.99 ± 0.00
Prothrombin (<i>n</i> = 3)	126	1.00 ± 0.00	20.03 ± 0.00	1.00 ± 0.00
fpocket				
BRD4 (<i>n</i> = 2)	15	0.97 ± 0.03	18.67 ± 1.33	0.94 ± 0.06
Carbonic anhydrase 2 (<i>n</i> = 3)	76	0.99 ± 0.00	18.69 ± 0.00	0.93 ± 0.00
CDK2 (<i>n</i> = 3)	180	0.76 ± 0.03	8.83 ± 1.74	0.43 ± 0.10
Estrogen receptor (<i>n</i> = 5)	58	0.94 ± 0.01	16.07 ± 0.17	0.80 ± 0.01
HIV-1 protease (<i>n</i> = 3)	219	0.98 ± 0.00	19.52 ± 0.00	0.97 ± 0.00
Prothrombin (<i>n</i> = 3)	126	0.94 ± 0.01	17.76 ± 0.32	0.88 ± 0.01

Comparison between the ligCav dataset of ligand-dependent binding sites (*N* = 9275), and those identified through fpocket cavity detection (*N* = 24,345). Mean results and standard errors are shown. *n* = number of queries; for further detail of the query and true positive patches see Additional file 1: Table S1

Clustering and mapping of potential binding sites

We applied the Pilot dataset to develop and validate a method for clustering and mapping objectively detected potentially ligandable binding sites. An exhaustive all-against-all SiteHopper BSC was performed to generate a 2708 × 2708 matrix of SiteHopper PatchScores. Possible PatchScores range from zero to four, where zero indicates total pairwise dissimilarity and four indicates perfect similarity. To remove non-conserved, information-poor potential binding sites, those with fewer than five PatchScores ≥ 2.0 were removed. Starting from an initial pool of 2708 potential binding sites, this criterion reduced the data to a 1706 × 1706 matrix of SiteHopper PatchScores.

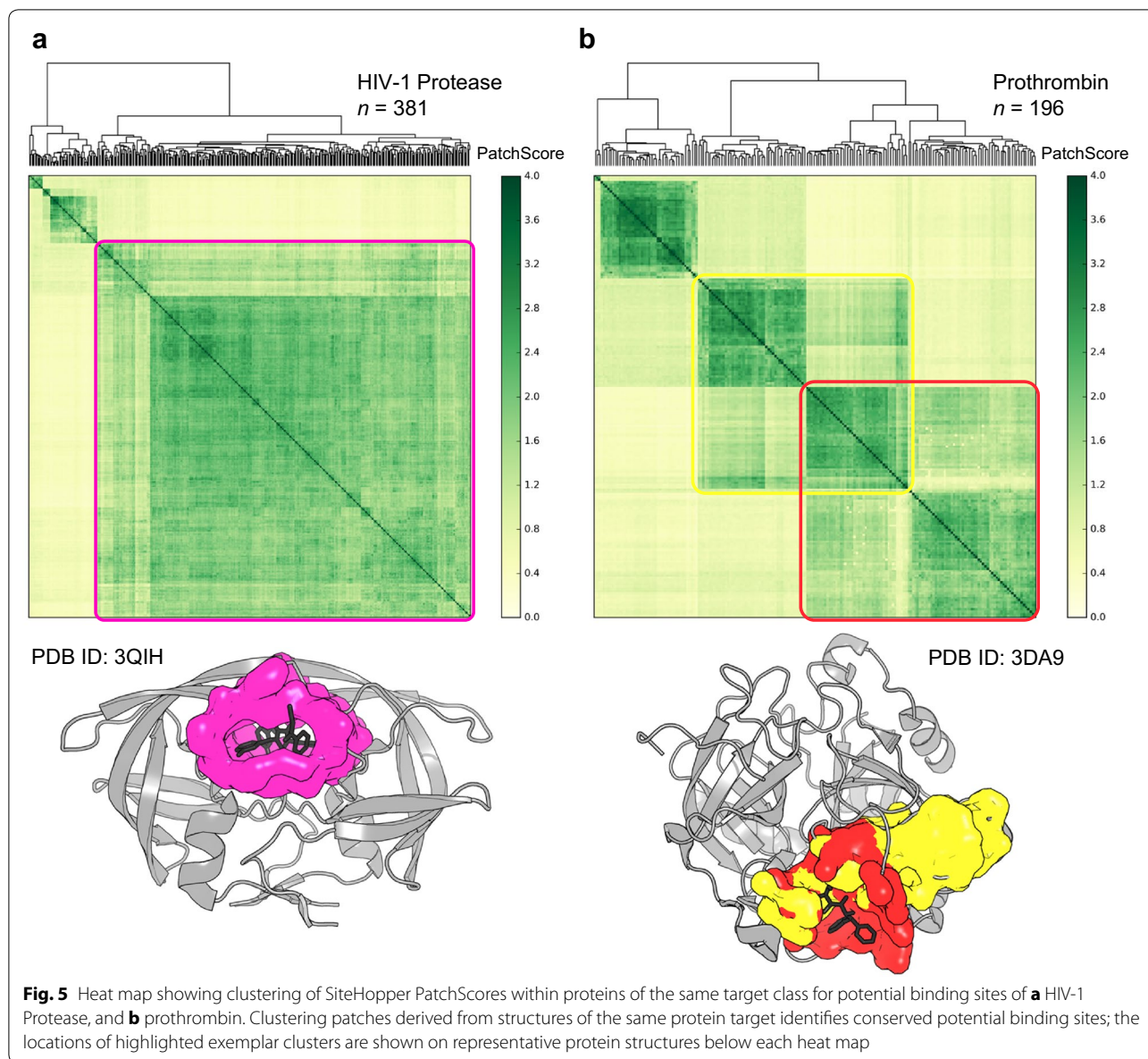
Hierarchical clustering provides an objective method for grouping structurally similar binding sites. Potential binding sites were first clustered locally according to their SiteHopper PatchScores across all available structures of the same protein target. Binding sites that share a pattern of similarity across the dataset are clustered together and represent groups of distinct potential binding sites within the protein family. Clusters vary in size, indicating that some cavities are more conserved than others. For all protein targets in the Pilot dataset, the largest cluster for

each protein corresponds to the conserved, orthosteric ligand-binding site. Figure 5 shows exemplar clustered heat maps of SiteHopper PatchScores for HIV-1 protease and prothrombin with the location of highlighted exemplar clusters shown on representative protein structures below each heat map. The most conserved cluster of potential binding sites for HIV-1 protease corresponds to the protease catalytic binding site to which many compounds are known to bind. Interestingly, binding site similarity is identified despite the prevalence of binding site mutations among structures of this protein. The large volume of the HIV-1 protease binding site, such that each pharmacophoric pseudocenter contributes less to the overall similarity, may contribute to the homogeneity of this conserved cluster. Figure 5b highlights two distinct prothrombin potential binding sites that show a degree of structural similarity to each other. This is a phenomenon arising from the integration of cavity detection into the workflow and represents a case where two overlapping binding sites are detected (depicted in red and yellow).

Next, we applied hierarchical clustering globally to the Pilot dataset, grouping clusters of similar binding sites regardless of their parent protein target. A combination of clustering both locally within protein targets and globally across the entire Pilot dataset generates a map of potential binding site space (Fig. 6).

The number of dominant clusters in the Pilot dataset represents the homogeneity of potential binding sites across the six proteins (Fig. 6). The scarcity of clusters of conserved binding sites among CDK2 structures is consistent with the flexibility of this kinase observed in protein crystal structures—notably the presence of diverse active and inactive protein conformations. Notably, SiteHopper identifies dominant clusters of substantially conserved binding sites for each of the five other protein targets. The most highly conserved binding site is that of the HIV-1 protease, likely due to the large volume and enclosed shape of the catalytic binding site that enables consistent identification by fpocket and robust detection of similarity by SiteHopper, respectively. Other factors that will likely affect the presence or absence of conserved binding sites within available structures of a particular protein include the presence of apo and holo bound structures, particularly for proteins containing multiple domains [52].

The global cluster analysis highlights a region of overlap between two clusters of binding sites belonging to prothrombin and cationic trypsin (Fig. 6). SiteHopper identifies structural similarity between the catalytic protease binding sites of these proteins (highlighted in red, Fig. 6). These two proteins are known to bind similar compounds and are annotated with a *selectivity group* of 254 compounds in the ChEMBL database (version 21)

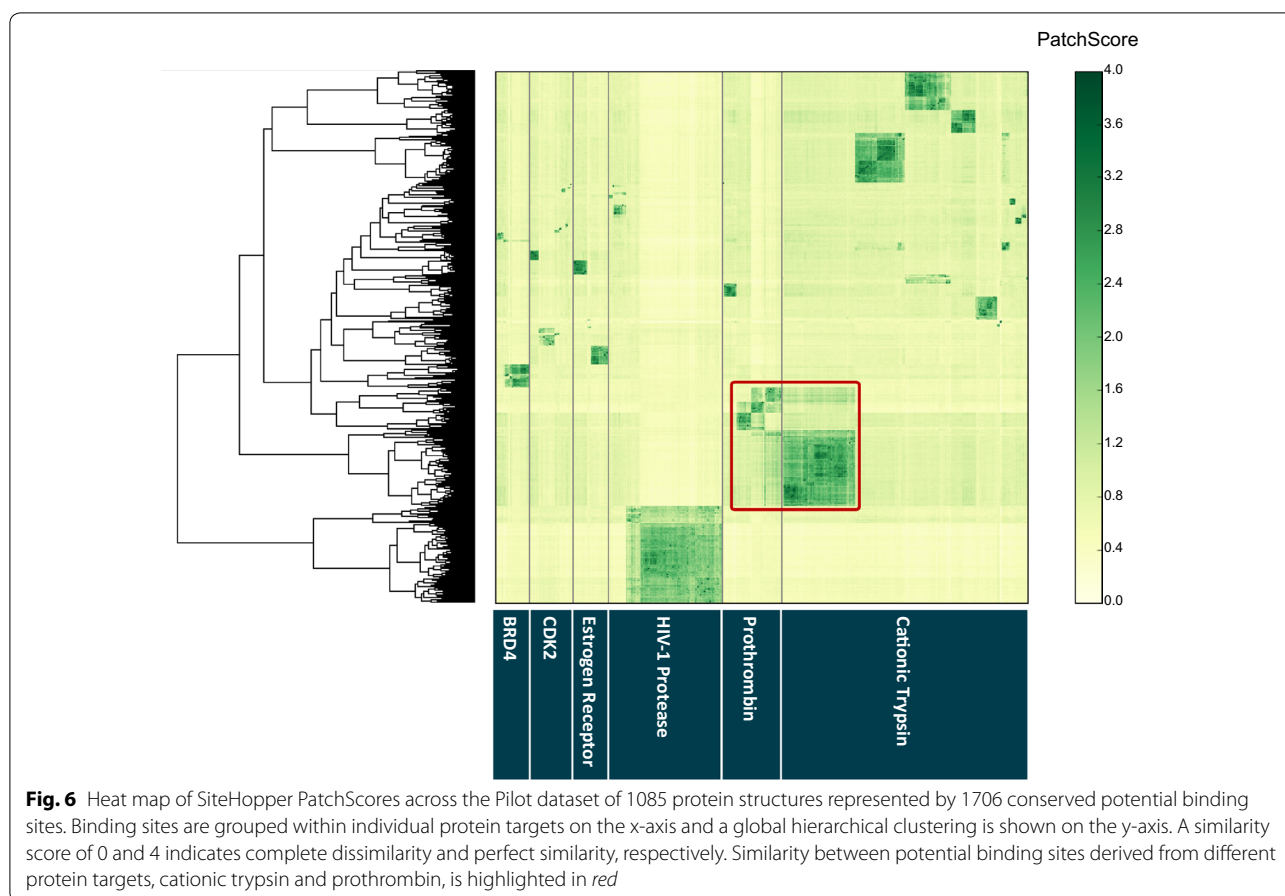


[53, 54] of bioactive molecules. Compounds in the selectivity group represent literature examples where the ratio of binding (selectivity coefficient) between prothrombin and trypsin has been measured; 99 of the 254 selectivity group examples exhibit a selectivity ratio of less than ten indicating that ligands commonly bind to both protein targets. The identification of similarity between these binding sites exemplifies the potential of BSC tools to rationalize and predict polypharmacology independent of ligand data.

Despite efforts to minimize the noise introduced by cavity detection, non-conserved potential binding sites inevitably affect the interpretability of clustered heat maps because much of the heat map conveys regions of

structural dissimilarity—which is less informative than similarity. Furthermore, non-conserved binding site patches that do not show SiteHopper similarity to other patches are grouped together by clustering methods, generating a group of information-poor binding sites. One method to reduce the presence of these information-poor binding sites is to apply stricter binding site conservation criteria. However, these would penalize potentially interesting novel proteins for which there are fewer instances exemplified in the PDB versus more extensively studied proteins.

DBSCAN [40] is a clustering algorithm widely used in data science that aims to group closely related points, and to label those points with few neighbors as noise.



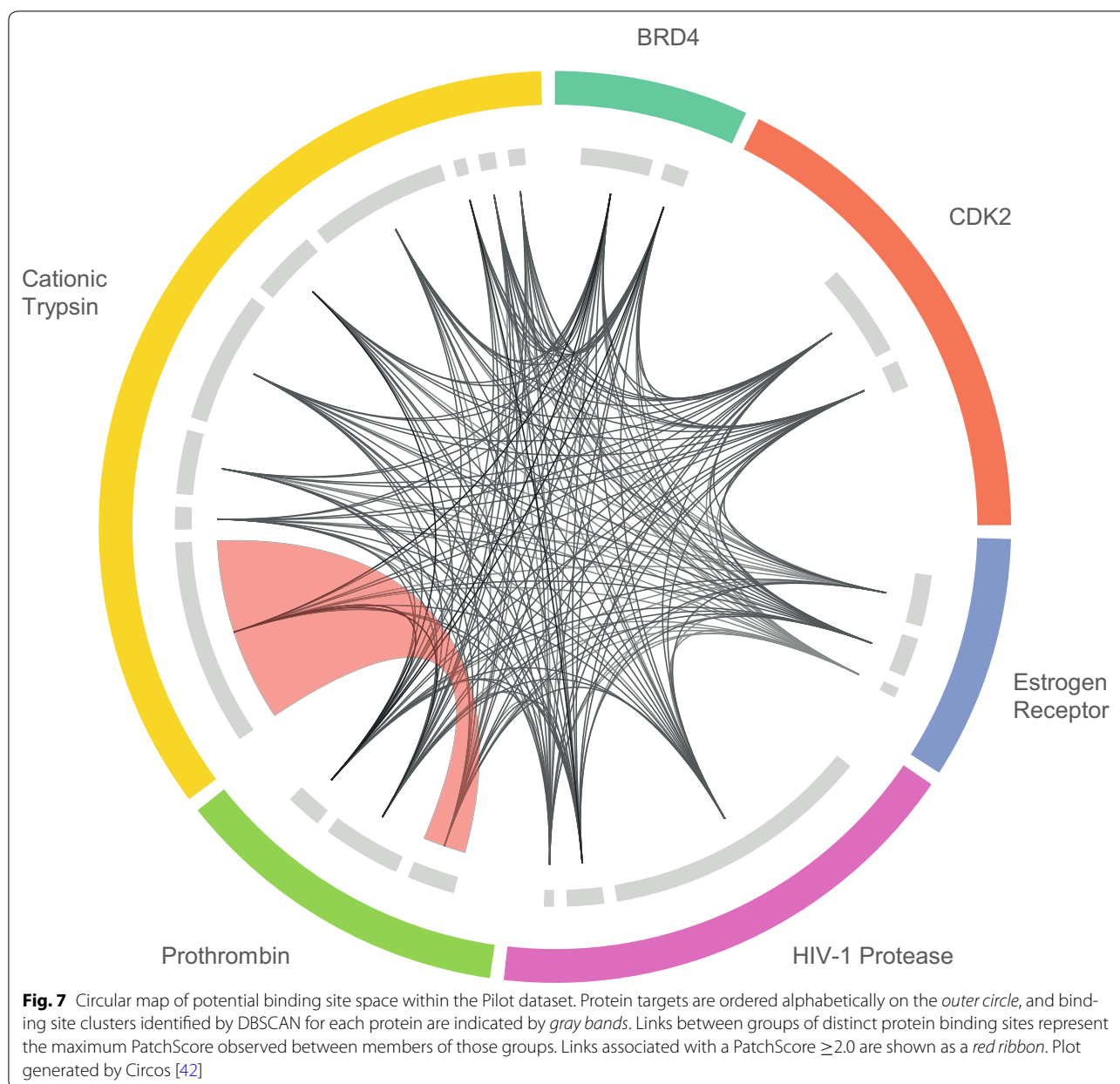
DBSCAN is suited to binding site analysis because it is not designed to create uniformly sized clusters [55]. DBSCAN was implemented to cluster potential binding sites locally within protein targets for the Pilot dataset. Due to the high dimensionality of the underlying data, visualization of DBSCAN clustering has proven challenging (the variance explained by the first two principal components is 0.42 and 0.18, respectively). Here we present a circular plot to map DBSCAN clustering of potential binding sites in the Pilot dataset, with protein targets arranged on the outer circle (Fig. 7). Conserved binding site clusters associated with each protein target are represented by gray bands on the inner circle; links describing the structural similarity between two binding sites are shown in the center. Each link is associated with the maximum SiteHopper PatchScore exhibited by members of two binding site clusters and those associated with a SiteHopper PatchScore ≥ 2.0 are highlighted as red ribbons. Consistent with the global cluster analysis (Fig. 6), this approach also highlights a structural similarity between binding sites belonging to prothrombin and cationic trypsin confirming the ability of BSC tools to identify the potential for ligand polypharmacology.

Conclusions

In constructing a workflow to map the binding sites of the currently characterized structural proteome, we adopted a modular approach that comprises objective binding site detection, binding site comparison (BSC), mapping of detected binding sites using unsupervised learning methods, and visualization of binding site maps. Although we outline a workflow for mapping potential small molecule binding sites in proteins, each of the components can be altered according to the tools available and specific hypothesis under test.

We applied fpocket as a geometric cavity detection tool to identify potentially novel unliganded binding sites, and modified fpocket parameters to yield concise cavities that are better suited to subsequent BSC studies. To filter out fpocket cavities that are unlikely to be ligandable, we determined a threshold fpocket Score by analyzing retrieval rates from the PDBbind-refined set; cavities were taken forward to BSC if the fpocket Score is ≥ 16.8 , corresponding to the Score above which 95% of the ligand-bound cavities from the PDBBind-refined set are identified.

Applying fpocket cavity detection to the sc-PDB dataset (2013) to assess the impact of incorporating objective



and unbiased cavity detection to BSC compared with only defining binding sites that surround exemplified bound ligands. Using SiteHopper for BSC, we show that the penalty associated with replacing ligand-dependent binding sites with objectively detected cavities is minimal and importantly also allows consideration of currently unliganded sites in BSC studies.

The workflow we describe applies the fpocket geometric detection algorithm to detect cavities in a protein structure. A limitation is that local chemical interaction hotspots and flat binding sites that are particularly relevant for the study of Protein–Protein Interactions (PPIs),

will not be identified. To map such binding sites, it may be possible to introduce an interaction hotspot prediction tool such as FTMap [56], GRID [57] or SuperStar [58] into the modular workflow; this will be the subject of future studies.

The Pilot dataset was processed by fpocket and an all-against-all SiteHopper BSC was performed to create a matrix of binding site similarities. Hierarchical clustering within protein structures derived from the same protein target reveals a large proportion of cavities that are not conserved across multiple structures of the same protein; we therefore introduced a conservation filter (removal of

cavities with fewer than five PatchScores ≥ 2.0) to minimize the number of information-poor cavities in the dataset. A combination of clustering both locally within protein targets and globally across the entire dataset, generates a map of potential binding site space. Furthermore, we show that density-based clustering by DBSCAN is an appropriate method for generating clusters of binding sites and mitigating the noise introduced to the dataset by objective fpocket cavity detection.

Although a powerful visualization, heat maps can be challenging to interpret, and therefore we introduce circular plots as an intuitive tool for visualizing and mapping structural binding site space. We show that such plots can highlight the similarity between binding sites derived from different proteins. Here, we exemplify an objectively identified similarity between binding sites of the serine proteases prothrombin and cationic trypsin that is consistent with literature reports that their catalytic sites bind similar ligands. We suggest that such protein binding site maps will be useful for building further understanding of the relationship between small molecules and complex biological systems; this approach is potentially applicable to the discovery of hit matter for novel biological targets, for predicting and rationalizing ligand polypharmacology and for predicting protein function [3, 4]. In addition, we suggest that such an objective binding site map, which encompasses unliganded cavities, will also be useful for optimizing compound screening collections towards a more complete chemical coverage of binding site space. We will present examples of such applications in due course.

Additional file

Additional file 1. A list of PDB IDs for two curated datasets utilised in this study: ATP-bound cAMP-dependent Kinase ensemble ($n = 5$) and the Pilot Dataset ($n = 1085$). **Figure S1.** Mean ROC curves for a series of experiments to determine the optimum site size to generate binding site patches surrounding fpocket surface atoms. **Table S1.** Summary of proteins from the sc-PDB (2013) that were considered in ROC retrieval studies. **Table S2.** Summary of the Pilot Dataset.

Abbreviations

AUROC: area under receiver operating characteristic; BSC: binding site comparison; BEDROC: Boltzmann-enhanced discrimination receiver operating characteristic; DBSCAN: density-based spatial clustering of applications with noise; EF: enrichment factor; PDB: protein data bank; BRD4: bromodomain-containing protein 4; CDK2: cyclin-dependent kinase 2; ER: estrogen receptor; HIV-1: human immunodeficiency virus-1.

Authors' contributions

JB, NB and JM conceived and designed the project. JM performed the experiments; JM analyzed the data with input from NB and JB. JM drafted the manuscript; JB and NB revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Yi Mok for his helpful guidance and comments on the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

JM is supported by Wellcome Trust Grant 102361/Z/13/Z. NB and JB are supported by Cancer Research UK Grant C309/A11566.

Received: 15 August 2016 Accepted: 16 November 2016

Published online: 06 December 2016

References

- Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. *Nucleic Acids Res* 28:235–242
- Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* 7:1884–1897. doi:10.1002/pro.5560070905
- Holm L, Sander C (1996) Mapping the protein universe. *Science* 273:595–602
- Ehrt C, Brinkjost T, Koch O (2016) Impact of binding site comparisons on medicinal chemistry and rational molecular design. *J Med Chem* 59:4121–4151. doi:10.1021/acs.jmedchem.6b00078
- Henrich S, Salo-Ahen OMH, Huang B et al (2010) Computational approaches to identifying and characterizing protein binding sites for ligand design. *J Mol Recognit* 23:209–219. doi:10.1002/jmr.984
- Dukka BK (2013) Structure-based methods for computational protein functional site prediction. *Comput Struct Biotechnol J* 8:e201308005–e201308008. doi:10.5936/csbj.201308005
- Schmidtke P, Souaille C, Estienne F et al (2010) Large-scale comparison of four binding site detection algorithms. *J Chem Inf Model* 50:2191–2200. doi:10.1021/ci1000289
- Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinform* 10:168–179. doi:10.1186/1471-2105-10-168
- Kellenberger E, Schalon C, Rognan D (2008) How to measure the similarity between protein ligand-binding sites? *Curr Comput Aided Drug Des* 4:209–220. doi:10.2174/157340908785747401
- Pérot S, Sperandio O, Miteva MA et al (2010) Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov Today* 15:656–667. doi:10.1016/j.drudis.2010.05.015
- Nisius B, Sha F, Gohlke H (2012) Structure-based computational analysis of protein binding sites for function and druggability prediction. *J Biotechnol* 159:123–134. doi:10.1016/j.jbiotec.2011.12.005
- Jalencas X, Mestres J (2013) Identification of similar binding sites to detect distant polypharmacology. *Mol Inf* 32:976–990. doi:10.1002/minf.201300082
- Batista J, Hawkins PC, Tolbert R, Geballe MT (2014) SiteHopper: a unique tool for binding site comparison. *J Cheminform* 6:P57. doi:10.1186/1758-2946-6-S1-P57
- Kryštof V, McNaie IW, Walkinshaw MD et al (2005) Antiproliferative activity of olomoucine II, a novel 2,6,9-trisubstituted purine cyclin-dependent kinase inhibitor. *Cell Mol Life Sci* 62:1763–1771. doi:10.1007/s00018-005-5185-1
- Grant JA, Pickup BT (1995) A gaussian description of molecular shape. *J Phys Chem* 99:3503–3510. doi:10.1021/j100011a016
- Chen Y-C, Tolbert R, Aronov AM et al (2016) Prediction of protein pairs sharing common active ligands using protein sequence, structure, and ligand similarity. *J Chem Inf Model* 56:1734–1745. doi:10.1021/acs.jcim.6b00118
- VIDA 4.3.0.4 (2015) OpenEye Scientific Software, Santa Fe, NM, USA
- Kuhn D, Weskamp N, Schmitt S et al (2006) From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J Mol Biol* 359:1023–1044. doi:10.1016/j.jmb.2006.04.024

19. Krotzky T, Grunwald C, Egerland U, Klebe G (2015) Large-scale mining for similar protein binding pockets: with RAPMAD retrieval on the fly becomes real. *J Chem Inf Model* 55:165–179. doi:10.1021/ci5005898
20. Chartier M, Najmanovich R (2015) Detection of binding site molecular interaction field similarities. *J Chem Inf Model* 55:1600–1615. doi:10.1021/acs.jcim.5b00333
21. von Behren MM, Volkamer A, Henzler AM et al (2013) Fast protein binding site comparison via an index-based screening technology. *J Chem Inf Model* 53:411–422. doi:10.1021/ci300469h
22. Hendlich M, Bergner A, Gunther J, Klebe G (2003) Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol* 326:607–620. doi:10.1016/S0022-2836(02)01408-0
23. Kellenberger E, Muller P, Schalon C et al (2006) sc-PDB: an annotated database of druggable binding sites from the protein data bank. *J Chem Inf Model* 46:717–727. doi:10.1021/ci050372x
24. Kufareva I, Ilatovskiy AV, Abagyan R (2012) Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucl Acids Res* 40:D535–D540. doi:10.1093/nar/gkr825
25. Hawkins PCD, Skillman AG, Warren GL et al (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and Cambridge structural database. *J Chem Inf Model* 50:572–584. doi:10.1021/ci100031x
26. Sitzmann M, Weidlich IE, Filippov IV et al (2012) PDB ligand conformational energies calculated quantum-mechanically. *J Chem Inf Model* 52:739–756. doi:10.1021/ci200595n
27. Sastry GM, Adzhigirey M, Day T et al (2013) Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* 27:221–234. doi:10.1007/s10822-013-9644-8
28. The PyMOL molecular graphics system (2016) Version 1.7.2.3. Schrödinger, LLC. New York
29. Liu Z, Li Y, Han L et al (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 31:405–412. doi:10.1093/bioinformatics/btu626
30. The UniProt Consortium (2014) UniProt: a hub for protein information. *Nucl Acids Res* 43:D204–D212. doi:10.1093/nar/gku989
31. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 25:3389–3402. doi:10.1093/nar/25.17.3389
32. HIV-1 Protease Consensus B Reference Sequence, HIV drug resistance database. <http://hivdb.stanford.edu/pages/3DStructures/pr.fasta>. Accessed 24 May 2016
33. Blow D (2002) Outline of crystallography for biologists. Oxford University Press, Oxford
34. Wirth M, Volkamer A, Zoete V et al (2013) Protein pocket and ligand shape comparison and its application in virtual screening. *J Comput Aided Mol Des* 27:511–524. doi:10.1007/s10822-013-9659-1
35. Vainio MJ, Puranen JS, Johnson MS (2009) ShaEP: molecular overlay based on shape and electrostatic potential. *J Chem Inf Model* 49:492–502. doi:10.1021/ci800315d
36. Kleywegt GJ, Harris MR, Zou JY et al (2004) The uppsala electron-density server. *Acta Crystallogr D* 60:2240–2249. doi:10.1107/S0907444904013253
37. Tanimoto TT (1957) IBM internal report 17th (November 1957)
38. Hunter JD (2007) Matplotlib: a 2D graphics environment. *Comput Sci Eng* 9:90–95. doi:10.1109/MCSE.2007.55
39. Jones E, Oliphant T, Peterson P (2014) SciPy: open source scientific tools for Python 2001–2012. *Mach Learn Optim Big Data* 9432:145–157. doi:10.1007/978-3-319-27926-8_13
40. Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* 96(34):226–231
41. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
42. Krzywinski M, Schein J, Birol I et al (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19:1639–1645. doi:10.1101/gr.092759.109
43. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143:29–36
44. Ballester PJ, Finn PW, Richards WG (2009) Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. *J Mol Graph Model* 27:836–845. doi:10.1016/j.jmgs.2009.01.001
45. Swamidass SJ, Azencott C-A, Daily K, Baldi P (2010) A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* 26:1348–1356. doi:10.1093/bioinformatics/btq140
46. Truchon J-F, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *J Chem Inf Model* 47:488–508. doi:10.1021/ci600426e
47. Yabuuchi H, Nijijima S, Takematsu H et al (2011) Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol Syst Biol* 7:472. doi:10.1038/msb.2011.5
48. OEChem Toolkit (2015) version 2015.02.09; OpenEye Scientific Software: Santa Fe, NM, USA
49. McKinney W (2010) Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, pp 51–56
50. Fauman EB, Rai BK, Huang ES (2011) Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. *Curr Opin Chem Biol* 15:463–468. doi:10.1016/j.cbpa.2011.05.020
51. Schmidtke P, Le Guilloux V, Maupetit J, Tuffery P (2010) fpocket: online tools for protein ensemble pocket detection and tracking. *Nucl Acids Res* 38:W582–W589. doi:10.1093/nar/gkq383
52. Brylinski M, Skolnick J (2007) What is the relationship between the global structures of apo and holo proteins? *Proteins* 70:363–377. doi:10.1002/prot.21510
53. ChEMBL v21. <https://www.ebi.ac.uk/chembl/downloads>. Accessed 8 March 2016
54. Gaulton A, Bellis LJ, Bento AP et al (2011) ChEMBL: a large-scale bio-activity database for drug discovery. *Nucl Acids Res* 40:D1100–D1107. doi:10.1093/nar/gkr777
55. Ivan G, Szabadka Z, Grolmusz V (2010) A hybrid clustering of protein binding sites. *FEBS J* 277:1494–1502. doi:10.1111/j.1742-4658.2010.07578.x
56. Kozakov D, Grove LE, Hall DR et al (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc* 10:733–755. doi:10.1038/nprot.2015.043
57. Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28:849–857
58. Verdonk ML, Cole JC, Watson P et al (2001) Superstar: improved knowledge-based interaction fields for protein binding sites. *J Mol Biol* 307:841–859. doi:10.1006/jmbi.2001.4452

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com
