mkgridXf: Consistent Identification of Plausible Binding Sites Despite the Elusive Nature of Cavities and Grooves in Protein Dynamics

Damien Monet, Nathan Desdouits, Michael Nilges, Arnaud Blondel

mkgridXf: Consistent Identification of Plausible Binding Sites Despite the Elusive Nature of Cavities and Grooves in Protein Dynamics

Damien Monet,[†,‡] Nathan Desdouits,[†,‡] Michael Nilges,[†] and Arnaud Blondel[*,†]

†*Unité de Bioinformatique Structurale, Département de Biologie Structurale et Chimie, CNRS-UMR 3528, CNRS-USR 3756, Institut Pasteur, 28 rue du Dr. Roux, 75015, Paris, FRANCE*
‡*Sorbonne Université, ED515 - Complexité du Vivant, 75005 Paris, France*

E-mail: ablondel@pasteur.fr

Abstract

We describe here a method to identify potential binding sites in ensembles of protein structures as obtained by molecular dynamics simulations. This is a highly important task in the context of structure based drug discovery, and many methods exist for the much simpler case of static structures. However, during molecular dynamics, the cavities and grooves that are used to define binding sites merge, split, appear and disappear, and cover a large volume. Combined with the large number of sites ($\sim 10^5$ and more) these characteristics hamper a consistent and comprehensive definition of binding sites. Our method is based on the calculation of instantaneous cavities and of the pockets delineating them. Classification of the pockets over the structure ensemble generates consensus pockets, which define sites. Sites are reported as lists of atoms or residues. This avoids the pitfalls of the classification of cavities by spatial overlap, used in most existing methods, which is bound to fail on non-ordered or unaligned ensembles, or as soon as significant molecular motions are involved. To achieve a robust and consistent classification we thoroughly optimized and benchmarked the method. For this we assembled from the literature a set of reference sites on systems involving significant functional molecular motions. We tested different descriptors, metrics and clustering methods. The resulting method is able to perform a global analysis of potential sites efficiently. Tests on examples show that our approach can make predictions of potential sites on the whole surface of a protein, and identify novel sites absent from static structures.

1 Introduction

1.1 Role of Cavities and their Dynamics in Protein Function

The function of a protein directly relies on its interactions with its substrates and/or partner.[1] Its conformation, the spatial arrangement of its atoms, is essential to establish these interactions. Noticeably, to ensure good affinity and specificity, a protein has to form a sufficient number of interactions with its ligand(s).[2] Gathering interacting residues on the boundary of a concave hole is a favorable way to reach a relevant number. Hence, substrates are mostly found in cavities, inside or at the surface of the protein.[3] Therefore, the ligand binding "site" can be advantageously identified by the "cavity" it forms and the amino-acids delineating it, the "pocket".

Identification of functionally relevant sites can be important to understand the protein mechanism of action, to evaluate its spectrum of substrates or to identify means to affect its function through, for instance, drug design.[4,5] Such sites are classically ascribed in two categories: orthosteric or primary binding site, often the active site of an enzyme; and allosteric sites, usually binding effectors inducing a change in the protein conformation and thus modulating the function.[6–8]

In either case, due to thermal motions and/or various activation processes, the protein conformation fluctuates and evolves in time. These evolutions can be essential for the function, either because only a subset of conformations is relevant for ligand binding or because the function of the protein indeed involves essential conformational changes that can be quite radical.[9] Many examples of holo and apo structures show such subtle or radical conformations changes depending on the trapped functional state.[10]

1.2 Ambiguous Nature of Site Definition: the Static Case

There is an inherent ambiguity in site definition. Sites are classically defined by ligands, but different ligands may lead to significantly different definitions.[11] The difficulty is even more fundamental when scouting for previously unknown sites for which there is no reference ligand.[12–14]

In the latter case, we have to use geometry and often composition criteria to identify a site forming favorable interactions. There might be different definitions, in particular for the boundary between the bulk solvent and the volume engulfed in grooves. They can also be somewhat subjective, which adds up to the intricacy.

1.3 Ambiguous Nature of Site Definition: the Dynamic Case

In addition, sites, in particular allosteric or cryptic ones can have an intrinsically dynamic nature. In highly dynamical systems, the relevant conformations may vary widely, and identification of binding site(s) becomes even more intricate and ambiguous due to appearance – disappearance (Figure 1.a), and fusion – splitting events (Figure 1.b). This may lead to drastic changes with wide merging or complete site dismantling.

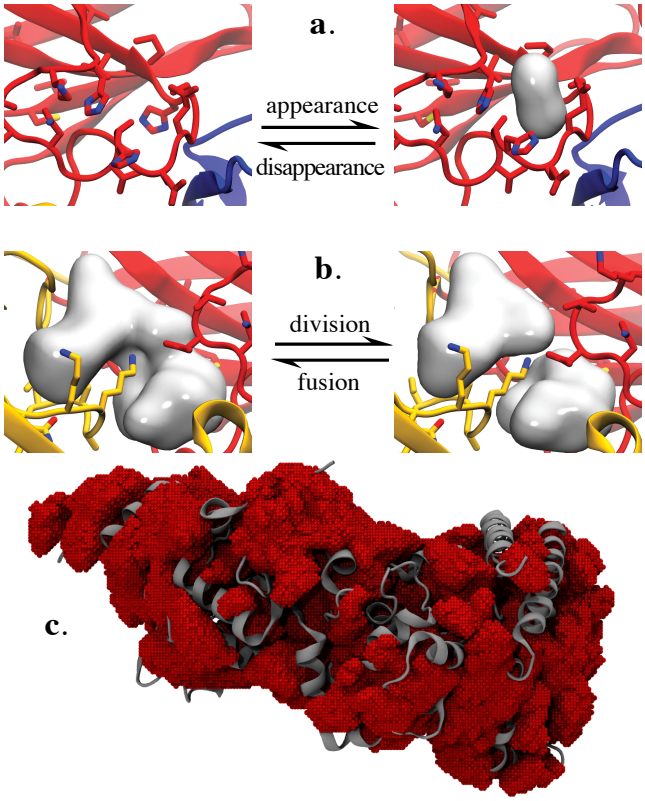


Figure 1: Examples of Events Hampering Consistent Site Definition. **a.** Appearance/disappearance of a cavity when its geometry reaches a threshold, for example, defining minimum volume or bulk solvent. **b.** Division/fusion of cavities triggered by thresholding on the median part. **c.** Cavities appear mobile and ubiquitous during molecular dynamics evolution (cavities domain of definition displayed by red spheres on a grid).

The wide mobility and ubiquity of cavities creates further difficulties. They can appear almost anywhere in the protein volume during molecular dynamics evolution (Figure 1.c). With protein motions, cavity identification is likely to face irreconcilable situations if it relies exclusively on spatial overlaps from one conformation to another. [15]

Hence, large relative domain motions appear an obvious source of failure, irrespective of the geometrical cavity descriptors, cubes, spheres, cylinders or any other. As a result, no global referential can be relevant in general for systems involving large relative domain motions.

1.4 Other Factors Hampering Site Identification

The total number of cavities in long trajectories can be up to millions, way beyond what is manually tractable. This obscures the process of consistent cavity or site identification, and following a particular cavity across multiple different conformations becomes a real challenge.

With a large number of cavity, and the fusion – splitting events between two or more cavities (see Figure 1.b), making consistent delineation of individual sites is difficult. Although, the user could decide the delineation of a site for a limited number of intermediates, when a large number of conformations are analyzed, it is likely to become rapidly impracticable. It could also be useful to limit the user's bias. Automation requires the identification of relevant criteria/thresholds to group or divide cavities from each intermediate conformation into sites, but proper benchmark is necessary to establish them.

1.5 Importance and Application of Dynamic Cavity Analysis

Dynamic analysis of cavities can unveil transient cavities, absent from the crystallographic structures, but which can be important for the protein function. [16] An archetypical example for functional cavity dynamics is given in oxygen release from myoglobin. [17–19] It supports the need for such an analysis to explain how "breathing" is associated with the kinetics of internal ligand diffusion. [17,20–26]

Interestingly, cavity geometry has a crucial impact on ligand binding [3,27,28] and virtual screening on multiple conformations is also developing to improve results. [29–31] But this approach has not been applied systematically so far.

Another application is in the identification of functionally relevant sites for drug design such as allosteric sites. [32] This requires a consistent site delineation to reliably correlate the cavity geometry and the different functional states of the protein, and estimate the impact that a ligand could have on the function.

1.6 Consistent Site Identification in Conformation Ensembles: Approach and Objectives

Identification of relevant sites for ensemble of conformations is important, but relevant tools are needed. In effect, currently, cavity analyses are performed mostly on single or few structures, or focus on a predefined locus (see [33,34] for recent reviews).

In this article, we propose an approach to identify relevant sites in conformation ensembles. For this, cavities are detected for each conformation, the list of residues or atoms that engulf each of them, the "pockets" is determined and those pockets are classified consistently to define potential binding site. This overcomes the weakness of a spatial/geometrical approaches: spatial alignment of proteint conformations is not needed and large relative motions can be handled.

Cavity pockets vary greatly in the course of a protein molecular dynamics, and making self-consistent and relevant clusters required thorough optimization. Hence, we systematically tested different internal definitions, methods, and parameters. To overcome the ambiguities and subjective perception, we selected a set of 15 "reference" sites previously described in the literature to challenge the approach.

A consensus set of parameters proved applicable to most systems. Other combinations with similar performance are also documented and listed. Comparison with existing methods shows only one comparable method. [16] However it appeared far less efficient and accurate, supporting the novelty and genuine contribution of the approach presented here.

2 Materials and Methods

2.1 Cavity Tracking Methods and Definitions

Instantaneous Cavity Detection Any cavity detection tool providing a spatial descriptor of the cavities allowing the identification of the pockets can be used. We used an in-house program, *mkgridXf*, for performance and better integration with the clustering. It is based on the principle of molecular surface [35,36] with a grid based implementation. The details of the method and its implementation are given in Supporting Information).

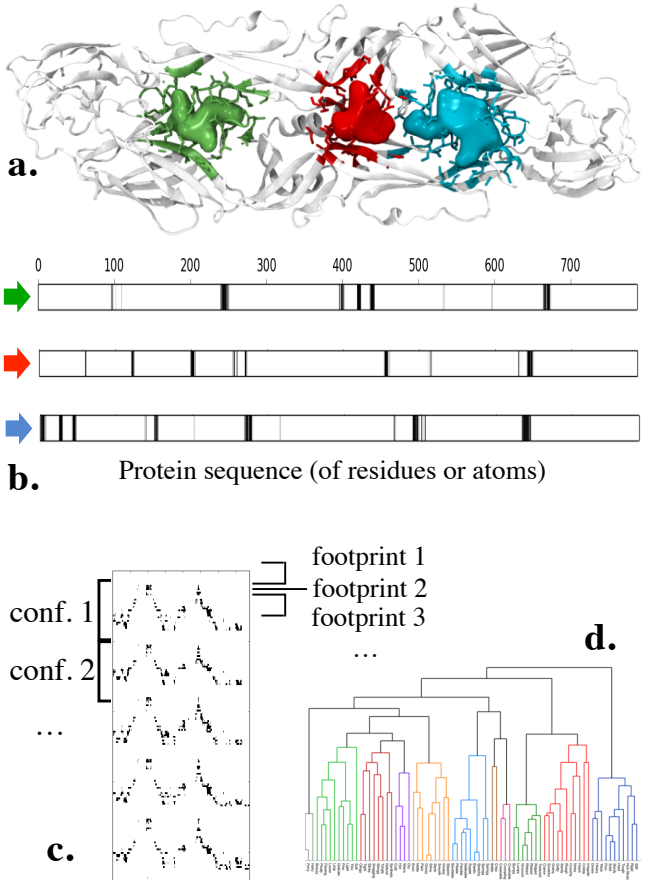


Figure 2: Schematic View of Cavity Descriptor Generation and Clustering. **a.** Three cavities, marked by green, read and cyan volumes respectively, are surrounded by protein atoms/residues shown in sticks and ribbons with the corresponding color for one of the conformation of the protein sampling ensemble. The rest of the protein (Dengue E protein), is shown with white ribbons. **b.** Fingerprints for each of the cavities (arrow of same color as in top panel) report the delineating atoms/residues in the protein sequence. **c.** View of the ensemble of footprints calculated for each cavity (footprint 1, 2, 3, ...) of each conformation (conf. 1, conf. 2, ...). **d.** Schematic view of the clustering of the footprint, grouping cavities having the same or similar protein environment.

Pocket Descriptor: Footprint The "footprint" of an instantaneous cavity is a descriptor recording the atoms or groups of atoms of the protein delineating the cavity: the pocket (Figure 2). Implementation allows the user to make this delineation at different levels of details with groups either composed of single atoms, *atoms*, residues, *residues*, or 2 groups per residue: its backbone and its sidechain parts, *B.S.*. Then, the footprints are defined as vectors indexed over all the groups composing the protein. Hence all the instantaneous cavities have a footprint of the same length, either the number of atoms, the number of residues or twice the latter.

Let c be a cavity composed of voxels, which centers are called v; g a group of the protein composed of atoms a, the distance (not in the mathematical sense) between c and g is given by:

$$\delta(c,g) = \min_{v \in c,\ a \in g} (d(v,a) - rad(a)), \tag{1}$$

where $d(v,a)$ is the Euclidean distance between v and a, and $rad(a)$ the van der Waals radius of atom a.

Three types of footprints, fp, are defined depending on how the delineating groups are encoded:

- *Real*,

$$fp_g^{real}(c) = \begin{cases} \sigma - \delta(c,g) & \text{if } \delta(c,g) < \sigma \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where the cutoff, σ, is 5 Å.

- *Boolean*,

$$fp_g^{bool}(c) = \begin{cases} 1 & \text{if } \delta(c,g) < \sigma \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

- *Local Boolean*,

$$fp_g^{lc-bool}(c) = \begin{cases} 1 & \text{if } \exists\ v \in c\ /\ a_v \in g \text{ and} \\ & \quad d(v,a_v) - rad(a_v) < \sigma \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where a_v is the atom at the smallest Euclidean distance from v. In the unlikely event where there is more than one atom at the same closest distance, the first one is taken.

Footprints are independent from the protein orientation within round-off variations. Real footprints were intended to keep some depth information in the pocket definition. Boolean definition appeared simple and consensual. Local Boolean was designed to define the minimal and essential pocket delineating the cavity and are used for splitting.

Distance between Footprints Distances (not in the mathematical sense) between footprints a and b are calculated with different dissimilarity measures depending on the footprint nature.

- *Euclidean Distance, for real footprints:*

$$d(a,b) = \parallel a - b \parallel = \sqrt{(a-b)^2} \tag{5}$$

- *Cosine dissimilarity, for real footprints:*

$$d^{cos}(a,b) = 1 - \frac{a \cdot b}{\|a\|\|b\|} \tag{6}$$

- *Jaccard dissimilarity, for Boolean footprints*

$$d^{jac}(a,b) = 1 - \frac{|a \cap b|}{|a \cup b|}, \tag{7}$$

it is noted $d^{jac-loc}$ when applied on Local-Boolean footprints.

Footprint Clustering Different methods were used, including hierarchical clustering (*UPGMA*, *Complete*), *Spectral* clustering, *DBSCAN*, and *MeanShift*.

Jaccard and Cosine distances are between 0 and 1 and the distance clustering threshold can either be specified as a fix value (from 0.05 to 0.95 by 0.05 increments) or by a ratio comparing distances between intra and inter conformation footprint distances histograms (ratio: .01, .05, 0.1, 0.3, 0.5, 0.7, 0.9, .95, .99). Real distances do not have an upper limit, and only threshold by ratio was considered. See in Supporting Information for details and implementation. When the number of footprint is too large, clustering can be performed on a subset, followed by reassignment.

Mean Footprint / Consensus Pocket For a cluster of footprints (i.e. transverse pocket) the mean-footprint is calculated to define a consensus pocket. Hence, the pocket is composed of groups with weighted occupancy. The user

can apply a truncation with an appropriate cutoff value to make a Boolean selection when required for a practical usage. Value of $\sigma/2$ for real footprint, or $1/2$ for Boolean ones, appeared relevant when a representative pocket is sought (see SI, Table S4).

Reassignment and Splitting When unassigned instantaneous cavities exist, reassignment consists in assigning instantaneous pockets/cavity to the transverse pockets having either the closest consensus pocket ("mean" assignment) or to the transverse pocket containing the closest instantaneous pocket ("min" assignment) (see SI-M&M for details).

When cavity splitting is requested, cavity voxels are reassigned to the most appropriate transverse pocket by best matching their atomic environment. This process is performed in two passes and has been tuned to avoid inconsistent repartition of the cavities voxels between two or more transverse pockets (see SI-M&M for details).

2.2 Implementation

Two implementations were made.
- The first in Python, *PyCAV*, incorporates all the options, methods, complementary analyses, as well as trajectory manipulation and geometry analysis of cavities. [20,37]
- The second, in C, *mkgridXf*, implements the options that proved practical as well as a fast cavity detection module and the ability to perform cavity splitting at the voxel level.

Output: All the data structures can be exported from *PyCAV* in binary format with a consistent indexation scheme allowing reimportation for further analysis. *mkgridXf* can export a concise data structure composed of the cavity grid points for each instantaneous cavity for each frame of the trajectory, or if the site identification is performed, for each transverse cavity. In the latter case, the mean-footprint for each identified site is also given. It can be read for further analysis by a small companion software, *mkread*.

2.3 Method Assessment

To overcome the fundamental ambiguities in site definition, especially in a dynamical context, we selected "Reference Cavities" to guide the algorithmic choices and to calibrate the method. The assessment was performed by systematic testing of the combinations of options and thresholds implemented in *PyCAV*.

2.3.1 Reference Sites

We chose protein systems with different sizes and involving different types of functional motions and for which main and accessory binding sites had been described.

Among the 12 sites (15 with symmetry), selected on 4 proteins, 10 were defined by crystallographic co-complex structures: 1J52 for Myoglobin; 2HZI and 3K5V for Abl1 kinase; 1OKE for Dengue E protein (short name DENV) and 1K90 for EF anthrax toxin. The other sites were defined from residues listed in the literature (see Results).

Molecular dynamics trajectories for myoglobin [20] (120ns), Abl1 (200ns), Dengue E protein (10ns), and EF anthrax Toxin [32,38] (10ns in inactive form and 10ns in activated form) were used to analyze cavity evolution. For the first three, 1000 frames evenly distributed were extracted. For EF, the two trajectories (inactive/active) were concatenated, and 2000 frames were extracted (see SI-M&M for details).

Then, in all cases, the instantaneous cavity corresponding to the site was identified manually in the crystallographic structure, and its *Local_Boolean - residue* footprint was used to define the reference pocket.

Reference assignment and Reference cavity trajectory definitions are given in the M&M of SI and Figure S1.

2.3.2 Score of Clustering Assignment

To score assignment, reference site assignment and clustering results are encoded in vectors, one vector element per instantaneous cavity c. Cavities of each step are aggregated to the previous ones, irrespective of their number per step (identifier and step origin is kept for later analysis). Hence, the reference assignment for reference site, *site*, is given by the Boolean vector, P^{site}, which is 1 if c is in *site* and 0 otherwise.

For a given combination of options, the result of the footprint clustering is encoded in a vector P, defined by $P(c) = k$, where k is the cluster number assigned to each cavity c (the total number of clusters is K).

For each cluster k, we defined the Boolean vector P_k by $P_k(c) = 1$ when $P(c) = k$, and 0 otherwise. Since those vectors are sparse (mostly zeros) and we focus on the actual detection of *site* cavities, we used the $F1$-score to compare P_k and P^{site}:

$$F1(P^{site}, P_k) = \frac{2TP}{2TP + FP + FN} \quad \text{with,} \tag{8}$$

$TP = P^{site}.P_k$, $FP = |P_k| - TP$, and $FN = |P^{site}| - TP$.

The clustering score for one site *site* is given by the score of the best cluster based on the overall-$F1$_score: [39]

$$F1_{site}(P^{site}, P) = max_{k\in[1;K]} F1(P^{site}, P_k) \tag{9}$$

Finally, we compute an $F1$ score averaging the scores of the different $sites \in Sites(prot)$ studied in the current protein:

$$F1_{prot} = \frac{\sum_{site\ \in\ Sites(prot)} |P^{site}| \times F1_{site}(P^{site}, P)}{\sum_{site\ \in\ Sites(prot)} |P^{site}|} \tag{10}$$

2.3.3 Geometrical Assessment of the Transverse Cavities

Beyond the assignment in *transverse cavities*, the method returns the cavity grid points labeled accordingly for each frame of the trajectory. Transverse cavities and reference cavities can be compared through their volumes (summing the respective voxels volumes). The geometry can be assessed more precisely with a distance d_{Geo} (not in a mathematical sense) inspired by Hausdorff. [40] For two non-empty cavities C_1 and C_2 made of voxels of center, v with zero radius ($rad(v) = 0$), it is given by:

$$d_{Geo}(C_1, C_2) = \sqrt{\frac{\sum_{v\in C_1} \delta^2(v, C_2) + \sum_{v\in C_2} \delta^2(v, C_1)}{|C_1 \cup C_2|}}, \tag{11}$$

$$d_{Geo}(C_1, \emptyset) = d_{Geo}(\emptyset, C_1) = 2 \cdot d_{Geo}(C_1, g_1)\ , \tag{12}$$

where $\emptyset$ is the empty cavity, and g_1 is the geometric center of C_1 with $rad(g_1) = 0$, and finally,

$$d_{Geo}(\emptyset, \emptyset) = 0. \tag{13}$$

Table 1: Cavities Analysis in Molecular Dynamics Trajectories. Volumes, and domains of definition are calculated as explained in M&M. *EF-Cam trajectory includes 1000 steps of inactive form (2.07 Å RMSD), 1000 steps of active form (2.3 Å RMSD), which when assembled yield 6.17 Å global RMSD. Additional data on the fluctuations along the trajectory are given in SI Results, Figure S2&S3. [a]Protein Envelope, [b]volume of cavity domain of definition as a percentage of the mean volume of the protein envelope

System	Traj. Steps	mean RMSD (Å)	Mean Cavities/fr. (std/min/max)	Mean Volume per cavity, Å^3	Mean volume per frame (Å^3)			Domain of definition (Å^3)			
					Cavities	Protein Env.[a]	(%)	Cavities	Protein Env.[a]	(%)	(% cd/mpe[b])
Myoglobin	1000	1.29	11.9 (2.4/6/20)	37	438	21,576	(2.0)	7,803	43,589	(17.9)	(36.2)
Abl1	1000	3.32	27.3 (4.3/12/40)	86	2,354	37,725	(6.2)	33,462	98,981	(33.8)	(88.7)
EF-CaM	2000	*	36.3 (4.8/22/54)	104	3,788	63,903	(5.9)	56,872	188,303	(30.2)	(89.0)
Dengue E p.	1000	2.63	58.4 (5.3/42/83)	75	4,365	107,622	(4.1)	51,630	186,858	(27.6)	(48.0)

3 Results

Site identification performed on sets of structures, e.g. protein molecular dynamics, appeared ambiguous and computationally demanding. We present here a method that efficiently and consistently performs this task.

3.1 Instantaneous Cavities

Analysis of the trajectories of our model systems revealed a total number of instantaneous cavities that varied from 11863 for Myoglobin to 72507 in the EF system (Table 1). In agreement with previous studies,[41–43] the number of cavities by frame, as well as their total volume, is roughly proportional to the average volume of the protein envelope (see Figure 3). The mean volume ratio between the cavity and the protein was low, between 2% and 6%. By contrast, the volume covered by cavities during the whole dynamics (domain of definition) was more than 10 times the mean cavity volume, representing up to 34% of the protein envelope domain of definition volume (Abl1), and up to 89% of the average protein volume (Abl1). Thus, cavities appeared numerous and they covered a large portion of the protein volume. This illustrates the difficulty of the identification in the dynamics context.

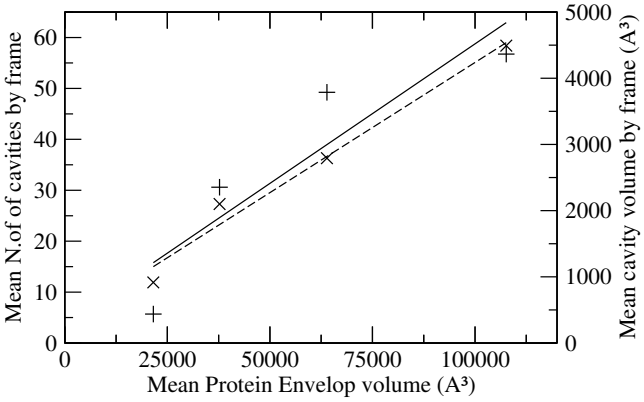


Figure 3: Number of Cavity by Frame (×) and Mean Cavity Volume (+) as a Function of Protein Volume. Linear regressions are respectively shown with dashed and straight lines.

This difficulty is confirmed by the lack of success in the attempts to perform cavity identification by spatial overlapping (see SI Results, Table S1 and Figure S4).

3.2 Architecture and Definitions of the Method

The method we found to effectively perform consistent site identification is based on the classification of instantaneous pockets. Its general organization and definitions are the following.

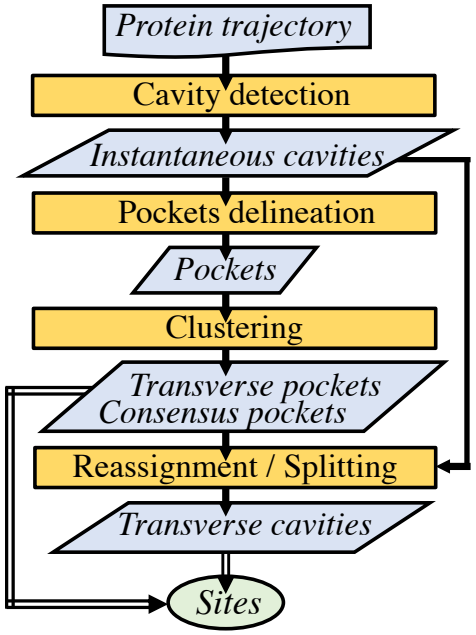


Figure 4: Workflow of the Site Identification/Cavity Tracking Algorithm. Routines are represented by orange boxes, data by blue parallelograms. Plain arrows represent the flow of data. Double line arrows convey Site definition (green ellipse).

Instantaneous Cavity Connected piece of volume between protein atoms, that is accessible to solvent, but not to the bulk solvent. An instantaneous cavity can be buried or at the protein surface (groove).

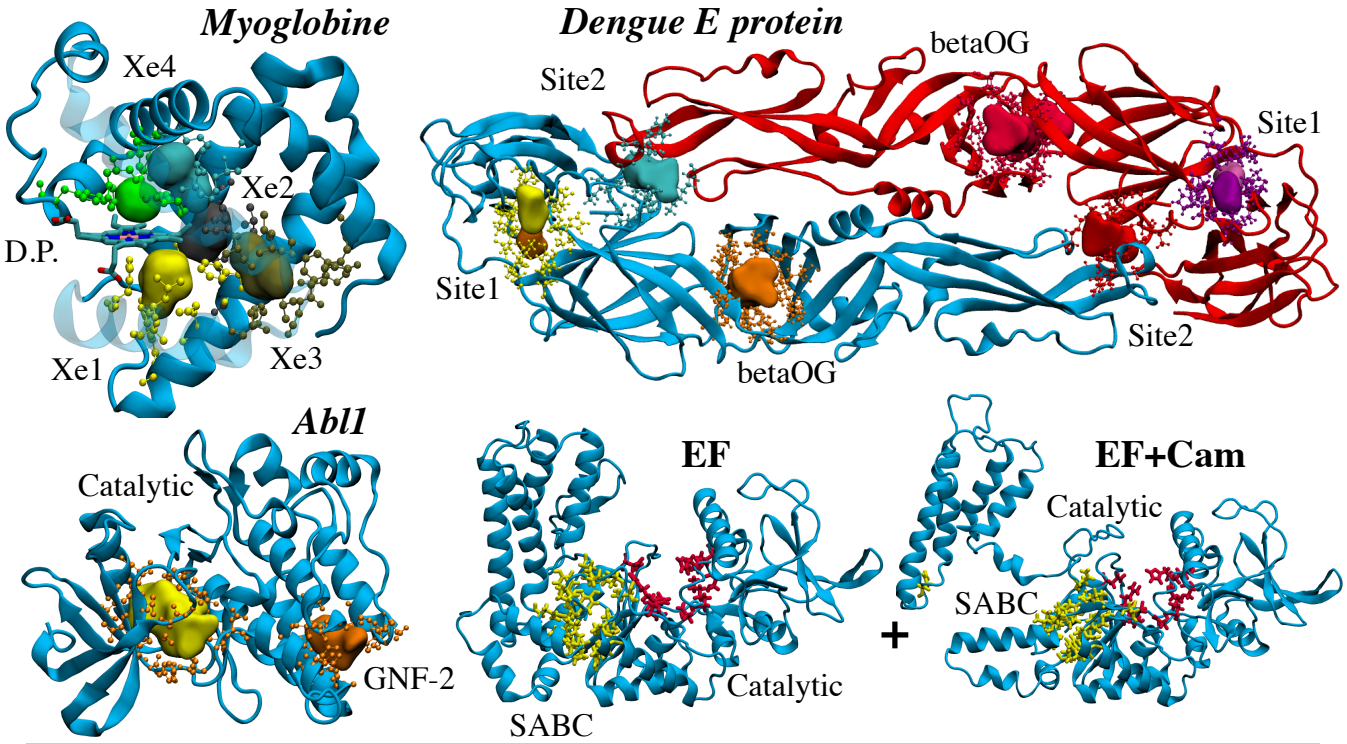
Instantaneous Pocket Groups of atoms of the protein delineating an instantaneous cavity (can either be listed as residues, individual atoms, or other groupments).

Transverse Pocket A set of pockets from different protein conformations that are clustered together based on the similarity of their list of groups of atoms.

Consensus Pocket A pocket defined by the average of the pockets descriptors found in a transverse pocket.

Transverse Cavity The ensemble of instantaneous cavities from different protein conformations that are delineated by the same transverse pocket.

Site Refers to a transverse pocket and its associated transverse cavity.



Reference Site	Residues
Myoglobin	
Distal Pocket	L29 L32 F43 H64 V68 I107
Xe1	L89 A90 H93 L104 F138 I142 Y146
Xe2	L72 L104 I107 S108 L135 F138 R139
Xe3	W7 I75 L76 K79 G80 H82 A134 L135 L137 F138
Xe4	G25 I28 L29 G65 V68 L69 L72 I107 I111
Dengue E protein	
Site1	P39 T40 H144 S145 G146 E147 Y178 L294 K295 T353 V354 N355 P356 I357 T359 S363 V365
Site2	chain-A: D98 R99 G100 G102 N103 K246
	chain-B: R2 I4 G5 I6 G152 D154
βOG	T48 E49 A50 P53 K128 V130 L135 G190 L191 F193 L198 Q200 A205 L207 T268 I270 Q271
	L277 F279 T280 G281
Abl1	
Catalytic	L248 G249 Y253 V256 A269 V270 K271 E286 M290 V299 I313 T315 E316 F317 M318 G321
	N322 L370 A380 D381 F382
GNF-2	A337 L340 L341 A344 L429 I432 A433 Y435 E462 G463 C464 P465 V468 F493
EF	
SABC	A496 P499 I538 E539 P542 S544 S550 W552 Q553 T579 Q581 L625 Y626 Y627 N629 N709
Catalytic	R329 K346 H351 S354 K372 D491 D493 H577 N583

Figure 5: Reference Sites Definition. Pockets are displayed on respective crystal structures: Myoglobin, PDB:1J52; Dengue E protein, PDB:1OKE; Abl1, PDB:2HZI; EF, PDB:1K8T; EF+Cam PDB:1K93 (calmodulin not shown). The cavity volumes are displayed as solid volumes for the first 3 systems. The distal pocket of myoglobin is labeled D.P. Dengue E system is symmetric and each site has 2 instances. Their reference pockets, also symmetric, are listed once only. Sites 2 of Dengue E protein are found at the interface between chain A and B.

Cavity Tracking Workflow The general workflow of the cavity tracking is presented in Figure 4. A molecular trajectory (desolvated) is taken as input. It can be aligned to facilitate subsequent analysis and visualization. The algorithm calculates instantaneous cavities and associated pockets, then clusters the latter to group the cavities, yielding an exhaustive enumeration of the protein potential binding site along the trajectory.

3.3 Reference Pockets/Sites to Benchmark the Method

Although this method rapidly gave promising results, inconsistencies remained, and no clear clustering approach emerged to reach full consistency (see SI, section "Classification of cavities by protein environment in dynamics lacks self-evident parameters", Figure S5). Therefore, it appeared necessary to use reference cavities or sites to benchmark the method and establish relevant settings.

Reference sites were selected among systems for which one or more binding pockets were described in the literature. We chose four systems with diverse size, function, and involving

Table 2: Evaluation of the Performance of Clustering Option Combinations. All the combinations are ranked by $F1_{prot}$ on each protein system. Then, the ranks on the 4 systems of each combination are summed (*RSum*). The combinations having the lowest *RSum* are ranked best. Clustering options, Group, Dist., Clust., Thr. are described in Figure 6. Thresholds are given in percentage. Assignment, Ass., is applied when necessary. $F1_{prot}$ scores are given for the different proteins. The average, Aver., and the worst, Worst, scores are also given. *Spectral clustering for Abl1 using all footprints failed due to memory errors, and a sampling of 1/10 was used in that case.

Rank	RSum	Group	Dist.	Clust.	%Thr.	Ass.	Myo	EF	DENV	Abl1	Aver.	Worst
1	112	byatom	cosine	upgma	50	min	0.958	0.797	0.901	0.735	**0.848**	0.735
2	124	byatom	cosine	upgma	50	mean	0.958	0.779	0.902	0.735	0.843	0.735
3	213	byatom	cosine	upgma	55	min	0.958	0.779	0.872	0.728	0.834	0.728
4	225	byatom	cosine	upgma	55	mean	0.958	0.762	0.878	0.728	0.831	0.728
5	269	byatom	cosine	upgma	60	mean	0.923	0.764	0.867	0.728	0.820	0.728
6	270	byatom	cosine	upgma	60	min	0.923	0.779	0.860	0.728	0.822	0.728
7	272	byatom	jaccard	upgma	65	mean	0.955	0.665	0.889	0.754	0.816	0.665
8	285	byatom	jaccard	upgma	75	mean	0.895	0.668	0.918	0.744	0.806	0.668
9	308	byatom	jaccard	upgma	65	min	0.955	0.659	0.902	0.754	0.818	0.659
10	322	byatom	jaccard	upgma	60	min	0.957	0.659	0.907	0.739	0.815	0.659
11	322	byres	jaccard	upgma	60	mean	0.931	0.659	0.921	0.738	0.813	0.659
... 14	337	byres	cosine	upgma	45	mean	0.899	0.662	0.913	0.745	0.805	0.662
15	339	B.S.	jac-loc	upgma	75	min	0.914	0.799	0.851	0.720	0.821	0.720
... 20	397	byres	cosine	upgma	50	mean	0.875	0.749	0.885	0.721	0.808	0.721
... 30	443	B.S.	cosine	upgma	55	mean	0.885	0.739	0.858	0.727	0.802	0.727
... 47	507	byres	cosine	upgma	35	min	**0.963**	0.657	0.834	0.745	0.800	0.657
... 61	567	B.S.	cosine	upgma	50	mean	0.895	0.645	0.894	0.725	0.790	0.645
... 66	590	B.S.	cosine	complete	90	min	0.845	0.751	0.823	0.829	0.812	0.751
... 104	746	B.S.	jaccard	*spectral	70	mean	0.867	0.641	0.846	0.733	0.772	0.641
... 124	860	byres	jaccard	upgma	75	mean	0.797	0.764	0.821	0.771	0.788	**0.764**
... 198	1127	byres	jac-loc	complete	95	min	0.815	**0.825**	0.756	0.706	0.776	0.706
... 248	1311	byres	euclid.	meanshift	auto-05	min	0.921	0.514	0.879	0.558	0.718	0.514
... 270	1418	B.S.	jaccard	dbscan	20	mean	0.840	0.573	0.733	0.736	0.721	0.573
... 764	3184	byres	jaccard	dbscan	25	min	0.368	0.665	**0.924**	0.105	0.515	0.105
... 779	3232	byres	jaccard	upgma	auto-01	min	0.903	0.110	0.150	**0.865**	0.507	0.110
...	...	...	...	...	...	...	...	...	...	...	...	...
2189	8694	byres	jaccard	complete	auto-99	min	0.055	0.061	0.044	0.008	0.042	0.008

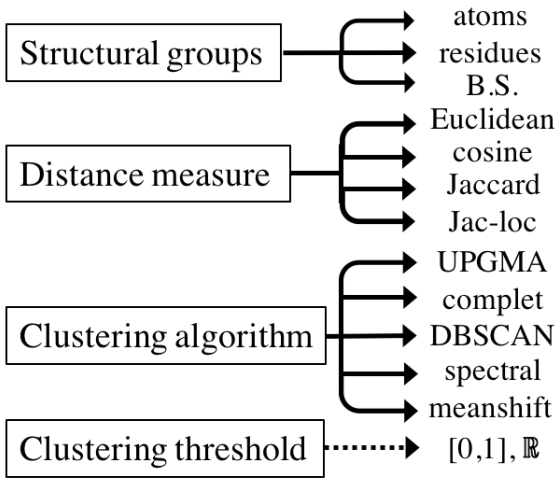


Figure 6: Options Combined in the Method. B.S.: Backbone-Sidechain.

small to large structural transitions. Hence, Myoglobin is relatively static (Table 1) while the protein E of Dengue and Abl1 kinase present larger fluctuations. Finally the EF component of Anthrax toxin was studied on two sets of conformations, one in the inactive and one on the active states, which are about 10 Å RMS distance from one another.

Binding pockets were defined as described in Materials and Methods starting either from co-crystallized ligand or from pocket delineation given in the literature. The lining residues used in the validation of the method are summarized in Figure 5.

In Myoglobin, they were marked by CO (distal Pocket) and Xenon atoms (Figure 5, Myoglobin). β-Octyl-Glucoside marked the βOG site in the Dengue E protein ([44,45]) (Figure 5, Dengue E protein). PD180970[46] and GNF-2[47] marked respectively the catalytic and allosteric sites in the Abl1 tyrosine-kinase (Figure 5, Abl1). Finally, 3'dATP marked the catalytic site of EF.[48]

Additionally, we also used the so-called site 1 and site 2 of the Dengue E protein given in Refs.[13,49] They are putative binding sites discovered by cavity analysis (from pre/post fusion crystallographic states comparison or molecular dynamic simulation). Their definitions (list of residues) differed in the two publications. Site1 is formed of two adjacent cavities in the crystallographic structure (PDB:1OKE) according to the list of residues given in.[49] In both cases, we choose the definition given in Ref.[49] as it delineated the binding sites with a smaller and more focused list of residues.

We used the definition given in Ref.[32] for the EF SABC site, which was discovered by transition path calculation between active/inactive states.

3.4 Clustering Options for Consistent Site Identification

As seen, clustering settings strongly impact the results. Hence, to identify the most consistent site delineation options, we tested the 2286 clustering combinations implied by Figure 6 on each reference protein system (see M&M). 2189 completed for all sites and could be evaluated with the $F1_{prot}$ score.

Depending on the objective (best performing, most robust, etc...), scores can be combined in different ways. Here, to identify combinations that best preform globally, but also do not under-perform on "difficult" sites, we computed the average site scores per system, rank those scores among the methods, summed the ranks ($\sum rank(F1_{prot})$: *RSum*), and ordered from smaller (the better) to larger, Table 2.

Table 3: Evaluation of Clustering Performance without Ambiguous Sites. Hence, Abl1 scores correspond to catalytic site only, and EF scores to the combination of SABC site in frames 1 to 1000, and catalytic site for frames 1001 to 2000. Same quantity definitions as in Table 2.

Rank	RSum	Group	Dist.	Clust.	%Thr.	Ass.	Myo	EF	DENV	Abl1	Aver.	Worst
1	200	byatom	jaccard	upgma	55	min	0.890	0.979	0.904	0.896	0.917	0.890
2	258	byatom	cosine	upgma	50	mean	0.958	0.966	0.902	0.898	0.931	0.898
3	280	byatom	cosine	upgma	50	min	0.958	0.965	0.901	0.898	0.930	0.898
4	294	byatom	cosine	upgma	45	mean	0.926	0.965	0.896	0.898	0.921	0.896
5	300	byres	cosine	upgma	30	min	0.925	0.977	0.835	0.897	0.908	0.835
6	305	byatom	cosine	upgma	45	min	0.926	0.963	0.903	0.898	0.922	0.898
7	355	byatom	jaccard	upgma	55	mean	0.890	0.974	0.881	0.896	0.910	0.881
8	355	byatom	jaccard	upgma	65	mean	0.955	0.959	0.889	0.906	0.927	0.889
9	363	byres	cosine	upgma	30	mean	0.925	0.976	0.828	0.897	0.906	0.828
10	397	byatom	jaccard	upgma	60	min	0.957	0.954	0.907	0.906	0.931	0.906
... 16	432	B.S.	cosine	upgma	30	min	0.891	0.975	0.858	0.886	0.902	0.858

Noticeably, the best two methods, only differing in the assignment of remaining footprints, had an *RSum* about half the following ones. The best combination happened to be: *UPGMA*/cosine dis./real footprints/*by atoms*/d_{th}=0.5/*min.* It was also the combination yielding the best average score ($^1/_4 \sum F1_{prot}$; "Aver." in Table 2).

The first occurrence for various groups, distance metrics and clustering are given in Table 2). Clustering methods respectively appeared: *complete* (66^{th}), *Spectral* (104^{th}), *MeanShift* (248^{th}) and *DBSCAN* (270^{th}).

In the top ranking combinations (Table 2), myoglobin and E protein consistently had high scores, while EF and Abl1 yielded moderate scores, below 0.8. To understand why, we analyzed the scores site by site.

3.5 Accuracy for Individual Sites

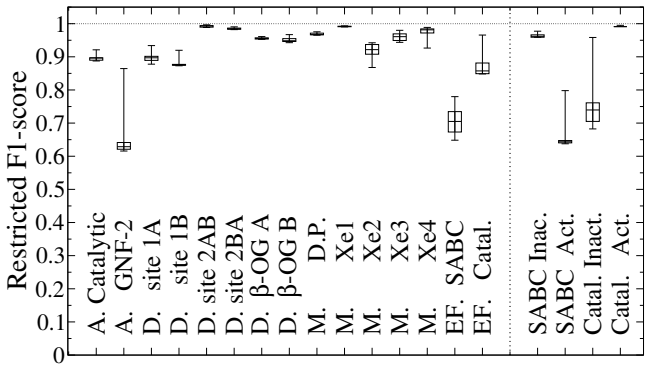


Figure 7: Distribution of the Best 100 $F1_{site}$ Scores among all the Combinations for Each Site. Proteins are abbreviated by A., D., M. and EF. for Abl1, Dengue E protein, Myoglobin, and EF respectively. Boxplots display: min, 1^{st} quartile, median, 3^{rd} quartile and max values. On the right, $F1_{site}$ are calculated ether on the first 1000 steps "Inact.", or steps 1000-2000, and "Act." of the trajectory for the SABC and Catalytic sites of EF.

The $F1_{site}$ score distribution for the 100 best scoring settings are depicted by boxplots for each site in Figure 7. Myoglobin sites are consistently well assigned (mean scores > 0.95 for all sites, except for Xe2, $\simeq 0.92$). The two copies of Site2 and βOG of the Dengue E protein scored better than the Site1 ones. Interestingly, symmetric sites of the Dengue system had similar scores. No method could predict the SABC site of EF with better scores than 0.8 and the top-100 average score was relatively low (0.70). GNF-2 site of Abl1 scored the worst with a top-100 average at 0.63. Thus, the latter two sites appeared as intrinsically difficult to identify consistently.

Noticeably, the EF trajectory is composed of two halves. The protein is in an inactive form in the first 1000 steps, and in an active form in the last 1000 steps. We analyzed the $F1_{site}$ scores independently in the two phases (Figure 7, right). Following the activation mechanism, the catalytic site is better formed in the second half. Conversely, the SABC site is well formed in the first half, but split in the second half as expected from previous study.[32] Accordingly, SABC was well identified in the first phase and the catalytic site in the second one ($\simeq 0.96$ or more), while they were poorly delineated in the other respective parts ($\simeq 0.65$).

3.6 Assessment for Unambiguous Sites

To check whether the "optimal" settings for all sites would also best predict consensual sites, we calculated the ranks without ambiguous sites. Hence, the allosteric site of Abl1 was removed, the SABC site of EF was used on the first half of the trajectory only and the catalytic site of EF was used on the second half (see Table 3). As a result, Abl1 and EF scores improved greatly. Here, discrimination with *RSum* proved uneasy as scores are close to their maximum. Our previous combination, *byatom-cosine-UPGMA-50*, ranked second and was outperform by *byatom-jaccard-UGPMA-55* due to improved score for EF. However, *byatom-cosine-UPGMA-50* was better for average score (0.931) and worst score of (0.898). The 10^{th} ranking method, *byatom-jaccard-ugpma-60* appeared better for those indexes (0.931/0.906), but it performs much worse on the "all systems" benchmark (also 10^{th} in Table 2). *byres* and *backbone-sidechain* footprints appeared in the top 10 options with similar methods and thresholds, which showed their suitability for cavity tracking.

From now on, the *byatom-cosine-UPGMA-50* combination, which appeared the most robust is considered.

3.7 Correct Location and Delineation of Predicted Sites

Modest $F1_{site}$ scores (e.g. Abl1 GNF-2 site or SABC of EF in active form, Figure 7) could either be caused by wrong assignment for some time steps, which would be a flaw, or by instantaneous protein deformation inherently altering cavity delineation and, thus, close matching with the reference, which would then be acceptable. This was tested (SI, "Correct location of predicted sites", Table S2). Actual misassignment never occurred in unambiguous sites and only occurred rarely in "difficult" ones (1 % of the time for GNF-2 site of Abl1, an unstable bundle of helices, and 9 % the

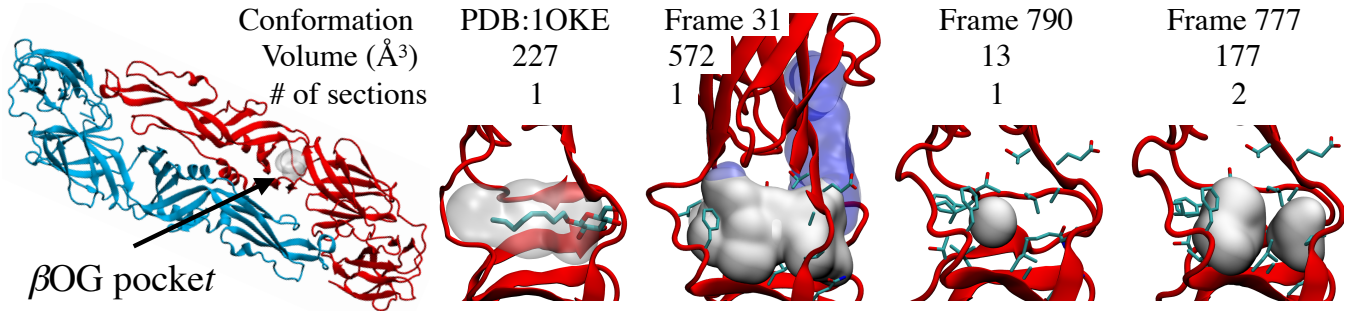


Figure 8: Geometry Diversity of the Reference Cavity of the βOG Site (light gray volume) Along the Trajectory (chain A of Dengue E protein). The βOG molecule is shown in sticks in 1OKE structure (cavity in transparency). The volume and number of pieces are displayed. An example of fusion with a neighboring cavity is given in Frame 31. The extension of the instantaneous cavity beyond the reference cavity is displayed in transparent blue volume.

time for SABC site, when it is scattered at different locations in the activated form).

Further testing was made to check if the method predicted the correct pockets (lists of atoms or residues). This was done by comparing the consensus pocket and its reference. For most of the sites, delineations closely matched (see SI, Table S3). However, some sites had average matches. To distinguish if this was due to a weakness of the method or simply because better match is not possible due to the intricacy of the benchmark, we made the following comparison. We assembled the best possible set of cavities along the protein trajectory by selecting instantaneous cavities located in the reference pocket. We calculated their instantaneous pockets and the resulting consensus pocket. Finally, we compared that consensus pocket with the reference one. This gave surprisingly similar results (Table S3). Hence, cavity delineation appeared to be as good as it could be, given the intrinsic variability introduced by the dynamics.

3.8 Mapping Cavities to Reference Sites Requires Splitting

The cavity found in the reference site did not always match the reference pocket, for example, due to fusion with a neighboring cavity. In an attempt to quantify and solve this point, we computed reference cavities tightly mapping the reference pocket definition for each reference site as explained in SI. Statistics on their volume, number of segments and presence show relatively large variance (see SI, Table S4). Noticeably, the cavity geometry varied largely as can be seen for examples on the βOG pocket in Figure 8.

We compared the reference cavities with the best matching transverse cavity using the volume and the geometric distance, d_{Geo}. An example is given in Figure 9 for an interval of the βOG (chainB) site trajectory. Volumes are close most of the time, but in some instances the cavity vanishes (31/1000 intermediates) or it is bigger than the reference cavity (by more than 200Å^3 24 times in 1000). d_{Geo} largely increases on these instances (purple bars on Figure 9). Closer inspection revealed large cavities covering multiple sites, as for example, in Frame 31 of Figure 8. In these cases, assignment of the instantaneous cavity to a single transverse cavity did not appear relevant. This led to introduce cavity splitting (see M&M).

Applying splitting reduced discrepancies (Figure 9). Worst volume difference for the βOG (chainB), 670 Å^3, was reduced down to 21 Å^3. The mean volume difference along the whole trajectory decreased from 31Å^3 to 9Å^3. Overall, cavity volumes fit more closely to the reference

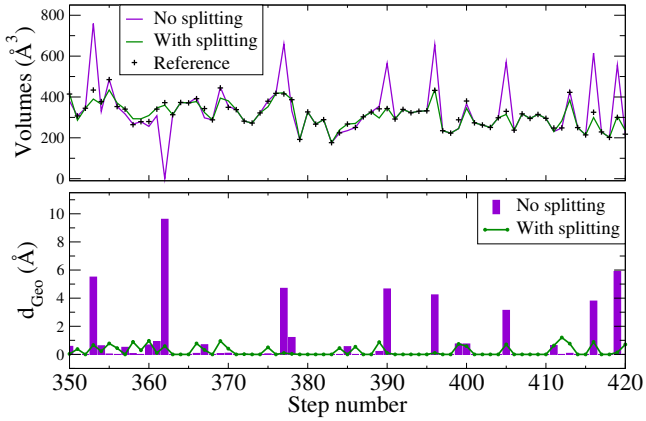


Figure 9: Volume and Geometry Comparison of the Reference and Predicted Cavity for the Dengue E Protein βOG Site, chain B, Given for an Illustrative Time Interval. Top: volume monitoring; "+", reference; purple, predicted; green, predicted with splitting. Bottom: Geometric distance d_{geo} between the reference cavity and the predicted one, without splitting: purple, and with splitting, green.

cavity trajectory for all the systems (see SI, Figure S6). Some sites (Catalytic in Abl1, catalytic in EF, and Site2 from Dengue E protein) still displayed variations, suggesting intrinsic difference between the reference sites and the cavities that can be found in dynamics trajectories.

3.9 Mapping some Abl1 Ligand Binding Sites

Table 4: Co-Crystallized Ligands and Matching Predicted Sites. Ligands found in other crystal structure of Abl1 were associated to the identified sites (see Table S5). The site identifier, the number of associated co-complexes, the number of different ligands (in parenthesis), and the types of site are indicated. Site matching the homologous kinase allosteric site found in PDB 5MRD (poc20) is also mentioned.

Site ID			Complexed Ligands
poc40	18	(12)	complexes in catalytic site (DFG-in like)
poc113	16	(13)	complexes in catalytic site (DFG-out like)
poc55	3	(2)	complexes in GNF-2 allosteric site
Others	16	(10)	Small molecules sites (Buffer, salt, etc...)
poc20	1	-	Homologous site for PS267 (S26) in 5MRD

Structures of Abl1 from the PDB were retrieved (see Table S5) and residues interacting with each ligand were used to find corresponding sites from the analysis made on the Abl1 molecular dynamics. Most ligands were associated to the catalytic groove (Table 4). They were associated to

two sites subdividing the groove, one in the catalytic part (poc40) one extending towards the extension found in the DFG-out forms (poc113; see SI, Figure S7). Other small molecules ligands were found in various other sites, including the GNF-2 site (poc55), in which myristic acid, involved in the kinase regulation,[50] also binds (see Table S5 for more details).

In addition, the binding site of an allosteric inhibitor, PS267 in complex with a related kinase, PDK1–PKCι (5MRD), was mapped on the Abl1 structure. To overcome difference in protein sequence, sequence-structure alignment was performed to identify the matching transverse cavity and pocket. Interestingly, a site, absent from the initial structure (2HZI) of the Abl1 molecular dynamics, was found in the dynamic site analysis (poc20, see SI, Figure S8).

3.10 Comparison with Existing Methods

Table 5: Site Identification Methods for Structure Ensembles. Channel identification methods were not included. First author of the reference is used when method name is lacking: [a]Krone and [b]Lindow. [c]*mkgrigXf*: this work. Cavity Detection: geometrical algorithm family. Cavity Accretion: approach to relate cavities from one step to another; For pocket based methods only the type of atom groups list is specified; Geometrical: sites delineated within the volume given in "Pocket Identification". Pocket identification can be "Exhaustive": automatically identify all site candidates; performed from a type of user selection (sel.), or; "No": not performed.

Method	Cavity Detection	Cavity Accretion	Pocket Identification
EPOS^{BP} [16]	Sphere	Atoms	Exhaustive
MDocket [51]	$\alpha-$sphere	Spatial	Cavity sel.
[a]Krone [52,53]	Gaussian	Spatial	No
Provar [54]	Sphere/Grid	Atoms/Residues	No
[b]Lindow [55]	Voronoi	Spatial	Cavity sel.
TRAPP [56]	Grid	Spatial	Pocket sel.
trj_cavity [57]	Grid	Spatial	No
POVME [58]	Grid	Geometrical	Sphere sel.
Epock [59]	Grid	Geometrical	Geometric sel.
[c]*mkgridXf*	Grid	Atoms/Residues	Exhaustive

Methods to study cavities in dynamics are listed in Table 5. Methods performing cavity accretion by protein environment, e.g. list of atoms or residues, are by construct the only ones to primarily perform pocket identification. They are also far less sensitive to orientation and protein conformation than methods using spatial or geometrical cavity accretion. The latter also require small time separation and original order of the molecular dynamics trajectory steps to ensure relevant overlap as can be read from their respective reference articles. Among pocket based methods, Provar does not pursue the actual pocket classification. Hence, only EPOS^{BP} and *mkgridXf* perform an exhaustive classification/identification of potential binding pocket, and could be compared.

Comparison of Performance and Site Delineation Completion time and number of sites are given in Table 6. *mkgridXf* proved more than 60 times faster on single processor, and more than 1200 time faster if 18 processors could be used. Memory requirement prevented EPOS^{BP} completion on large systems (Dengue E protein, EF).

Noticeably, EPOS^{BP} predicted more sites. This is consistent with the fact that fusion is not taken into account and classification depends on transient aggregations with neighboring cavities. In the same line, $F1_{prot}$ recalculated for EPOS^{BP} (see SI, Table S6) were 0.582 and 0.561 for Myoglobin and Abl1 respectively, lower than for *mkgridXf* (0.958 and 0.735; Table 2). Even if fusion is virtually allowed in EPOS^{BP}: more than one site can match the reference, these scores were not high: 0.642 and 0.630 (see, e.g. Figure 7 or Table 2 for comparison). Similarly, EPOS^{BP} cavities were often absent and did not fit reference ones as closely as *mkgridXf* allowed (see SI, Table S7).

Table 6: Comparison of Programs Performance and Number of Predicted Site. Single processor times are given. Time for 18 processors in parallel is also given in parenthesis for *mkgridXf*. For large systems, Dengue E protein and EF, EPOS^{BP} crashed with memory error (256 GB memory machine).

	mkgridXf		**EPOS^{BP}**	
	CPU time	#sites	CPU time	#sites
Myoglobin	35mn (2mn)	34	43h	65
Abl1	4h (12mn)	123	248h	137
Dengue E p.	28h (2h)	193	Mem. err.	-
EF	45h (3h)	159	Mem. err.	-

Hence, both in term of performance and in term of tight delineation, *mkgridXf* proved advantageous to perform consistent analysis of the cavities/pockets on an ensemble of protein structures.

4 Discussion

The elusive nature of cavities in protein dynamics and the resulting great difficulty to self-consistently identify them as specific sites were marking surprises. After unsuccessful attempts to develop an algorithm based on spatial overlap, use of instantaneous "pockets" appeared more effective. No comparable method exists except for one similar approach proposed previously,[16] but which lacked some fundamental features, such as the notion of fusion/splitting, and was by far not as consistent and efficient in our hands. In effect, obtaining self-consistency proved a challenge. It could only be achieved after testing various ways to define pockets, to measure their distances and to cluster them. Finally, assembling and use of reference sites/cavities/pockets to probe and optimize the method turned out to be essential to reach high accuracy.

Interestingly, our algorithm could make fairly accurate predictions on benchmarks including "difficult" cases. Difficulty arises from the internal motions of the site that can shift or create ambiguity in their definition. This stresses the importance of having reference sites displaying different behaviors, especially significant variability, to push optimization until robustness is reached.

Use of pockets as descriptors of cavities rather than use of spatial overlap makes the method less sensitive to internal motions, and insensitive to the orientation and the order in the conformational ensemble. Similarly, it also makes it insensitive to pruning of long trajectories to any extent.

Automation by pocket classification resolves site delineation, but it also allows to perform the analysis on hundred of thousand or even millions of transient cavities. This, with the possibility to prune extensively, gives access to the exploitation of extremely long simulations for large systems.

Use of pocket as descriptors also allows to analyze the whole surface of the protein without *a priori* or bias. Finally, the method returns the sites definitions as list of atoms or residues, allowing the analysis of their composition (hydrophobicity, charge, polarity, etc...). This, with the geometrical information, is convenient for subsequent use, for

example, to setup virtual screening.

4.1 Reference Sites

As noted,[33] gold standard libraries for cavity computation are sparse. Despite numerous databases for assessment of ligand binding sites (see e.g. Ref.[60]), to our knowledge, reference sites have never been described for cavity tracking in ensembles such as molecular dynamics. In the present work, we collected a set of 12 reference sites defined by a reference pocket. We calculated a trajectory of cavities along a molecular dynamics simulation for each system, and even, in one case, EF, two trajectories involving different functional states were generated.

The reference systems were selected for their diversity and are well described in the literature. They involved proteins of various sizes (Myoglobin, Abl1, EF, DENV). The first 3 are monomers while DENV is a dimer. In all cases molecular motions are important, ranging from relaxation allowing ligand diffusion in Myoglobin, to an extensive allosteric transition for EF. Some sites are at the surface while other are buried (DP & Xe1-4 in Myoglobin; GNF-2 in Abl1, SABC in inactive form of EF). All the sites are defined by ligands except for Site 1 & Site 2 of DENV. The nature of the ligands varies and correspondingly pocket composition differs with different hydrophobicity, polarity or charge (see Figure 5). Correspondingly, the definition of sites by the pockets (atoms or residues) provided by *mkgridXf* can help to select those with desired physico-chemical profile, and thus probable binding propensity.

Another aspect is the existence of "difficult sites". Examples presented here, suggest various causes for the "difficulty". In EF catalytic site and in Abl1 allosteric GNF-2 site, difficulty seems to originate from breathing motions, significantly changing the local cavities delineation. For Site 1 & 2 of the Dengue E protein, ambiguity arises from fusion and splitting. The SABC site is yet even more challenging. The pocket clearly delineated in the inactive form is scattered in pieces during the large allosteric reorganization of the activation.

These difficult reference sites are essential for the identification of a robust method, and its validation emphasizes its performance. Noticeably, the results question the very definition of some of the reference sites, suggesting that consideration of the protein dynamics may call for a refined definition.

4.2 Cavities and Dynamics

Literature, reports 0.06 to 2.26% cavity volume in proteins[41,43]). These numbers are for static structures, and in our experience larger figures are obtained in dynamics: from 2 to ∼6 % (Table 1). This difference may be due to the amount of sampling used for the analysis (see data characterizing sampling in SI, Figure S2 and S3), and could increase with the sampling.

Additionally, the studied proteins could have larger voids than average as we have selected them for their significant functional motions. Hence, Myoglobin displays breathing motion; Abl1, allosteric modulation; EF, large activation transconformation; and Dengue E protein, large motion between pre- and post-fusion conformations.

If we now consider the domains of definition of the cavities, we have much larger figures (Table 1). This illustrates that the course of cavities spans a wide volume than could not be anticipated on static structures alone. This has important implications for the diffusion of ligands, like O_2 in Myoglobin. This characteristics can also explain why spatial methods to track cavities are likely to fail.

An essential dynamics characteristics, absent in static structures, is that cavities move, split, merge, disappear and (re)appear in the sampled conformations. This is a strong challenge, creating ambiguities, and it required thorough optimization of the clustering method to reach satisfying identification consistency.

Another characteristics is the possible presence of states significantly differing, as found in the EF trajectory (inactive and activated states). If the different sub-states are known in advance, it is possible to make independent analyses for each sub-state. However, it would be better if the analysis could be performed without having to stratify sub-states. In effect, i) it may be difficult to make a relevant site stratification based on conformations only, ii) it might be more interesting to have a unified analysis. Here again, the robustness of the algorithm is essential/critical, but this open the way to the use of novel relevant conformations for known binding sites.[20]

Differences in definitions can appear through dynamic analysis. For example, reference pockets, consensus pockets predicted for the reference sites and consensus pockets identified for the reference cavities are similar, but not identical (see Tables S7-S8). Noticeably, crystal structures are structural averages and may not depict satisfactorily instantaneous cavities appearing in dynamics. Hence, methods taking functional molecular motions into account without requiring prior information, could lead to an enrichment of sites definition over that obtained from static structures.

4.3 Towards the Identification of Novel Effector Sites ?

Dynamics analysis can reveal new sites. For example, the numbers of cavities detected on static initial structures are 13, 10, 29/30 and 52 for Myoglobin, Abl1, Inactive/Activated-EF and Dengue E protein respectively. The average numbers of cavities per conformation in dynamics are 12, 27, 36 and 58 for the respective systems (Table 1), a significant increase (except for Myoglobin). Now, if we consider all the identified sites appearing more than 25% of the time in the course of the analysis, the above figures raise to 18, 89, 107 and 127 respectively. Hence, this approach can reveal new sites, which could be exploited, for example to bind effectors. To evaluate the potential of a novel site, the user can consider the exposition frequency of the delineating groups through their weights in the consensus pocket. The availability of a sufficient volume to bind a ligand can be evaluated from the times series of the cavity analysis.

Interestingly, the method could identify the reference sites despite the absence of ligand in the molecular dynamics simulations (except for myoglobin). In the latter case, one CO molecule only was hopping from one site to another.[20] Furthermore, for Abl1, the simulation was started from a crystal structure (2HZI) in which the allosteric site is in an apo form.

As another illustration of this capability, binding pockets of all ligands found in a series of other crystal structures of Abl1 could be mapped on the sites identified in the analysis of the Abl1 trajectory. Most of the ligands are found in the catalytic groove, divided in two regions (see Table 4 and SI, Table S5 and Figure S7). More interestingly, all the

other co-crystallized compounds, either allosteric or from the buffer, could be associated to an identified site. Although the matching score for smaller ligands was weaker (SI, Table S5), they could in some instances be indicative of interesting sites as exemplified by the site shared by myristic acid and GNF-2.

As a further illustration, preliminary analysis showed that the pocket binding compound PS267 in a related kinase (crystal structure 5MRD) was not found in the initial structure of the Abl1 trajectory (2HZI) by *mkgridXf*, but was identified in the course of the dynamics analysis (See SI, Figure S8). This successful identification suggests that our approach could be useful in drug design, where the identification of cryptic sites draw increasing interest.[61,62]

4.4 Conclusion

We conceived a method to identify sites in a consistent way on an ensemble of structures such as molecular dynamics trajectories. It overcomes the challenges due to site intrinsic ambiguity and flexibility, and can alleviate the burden of analyzing massive amount of data. It required thorough tuning, so we assembled a set of reference sites derived from the literature. This essential benchmark, helped to reach self-consistency, robustness and high accuracy. We could not find as efficient, memory economical and accurate methods preforming similar analysis in the literature. It appears to have the ability to identify novel binding sites, as illustrated on a few examples.

Acknowledgement This work has been supported by the European "Horizon 2020 HBP SGA1" (Grant Agreement No. 785907) grant, and the ANR NICOFIVE (ANR-17-CE11-0030) grant. ND was supported by the AXA Research Fund, "Don 2011", P766169, and DM by a PhD fellowship of "Ecole Doctorale Complexité du Vivant (ED-515)".

Supporting Information Available: Further details for Materials and Methods; Further data characterizing protein structure ensembles and representative intermediates; Intermediate results justifying the methodological choices; Further details on the match of delineated cavities/pockets with reference ones; Further details on improvement brought by splitting; data on identification of an homologous allosteric site on Abl1; and further comparison with existing methods. Movies illustrating the principle of the method and applications. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Perozzo, R.; Folkers, G.; Scapozza, L. Thermodynamics of Protein–Ligand Interactions: History, Presence, and Future Aspects. *J. Recept. Signal Transduction* **2004**, *24*, 1–52.

(2) Du, X.; Li, Y.; Xia, Y.-L.; Ai, S.-M.; Liang, J.; Sang, P.; Ji, X.-L.; Liu, S.-Q. Insights Into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *International journal of molecular sciences* **2016**, *17*, 144.

(3) Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7*, 1884–1897.

(4) Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **2003**, *10*, 787–797.

(5) Congreve, M.; Murray, C. W.; Blundell, T. L. Keynote Review: Structural Biology and Drug Discovery. *Drug discovery today* **2005**, *10*, 895–907.

(6) Christopoulos, A. Allosteric Binding Sites on Cell-Surface Receptors: Novel Targets for Drug Discovery. *Nat. Rev. Drug Discovery* **2002**, *1*, 198–210.

(7) Conn, P. J.; Christopoulos, A.; Lindsley, C. W. Allosteric Modulators of GPCRs: a Novel Approach for the Treatment of CNS Disorders. *Nat. Rev. Drug Discovery* **2009**, *8*, 41–54.

(8) May, L. T.; Leach, K.; Sexton, P. M.; Christopoulos, A. Allosteric Modulation of G Protein–Coupled Receptors. *Annu. Rev. Pharmacol. Toxicol.* **2007**, *47*, 1–51.

(9) Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* **2009**, *5*, 789–796.

(10) Nussinov, R.; Tsai, C.-J. Allostery in Disease and in Drug Discovery. *Cell* **2013**, *153*, 293–305.

(11) Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. Multiple Diverse Ligands Binding at a Single Protein Site: A Matter of Pre-Existing Populations. *Protein Sci.* **2002**, *11*, 184–197.

(12) Globisch, C.; Pajeva, I. K.; Wiese, M. Identification of Putative Binding Sites of P-glycoprotein Based on its Homology Model. *ChemMedChem* **2008**, *3*, 280–295.

(13) Fuzo, C. A.; Degrève, L. New Pockets in Dengue Virus 2 Surface Identified by Molecular Dynamics Simulation. *J. Mol. Model.* **2013**, *19*, 1369–77.

(14) Diskin, R.; Engelberg, D.; Livnah, O. A Novel Lipid Binding Site Formed by the Map Kinase Insert in p38α. *J. Mol. Biol.* **2008**, *375*, 70–79.

(15) Lindow, N.; Baum, D.; Bondar, A.-N.; Hege, H.-C. Exploring Cavity Dynamics in Biomolecular Systems. *BMC Bioinf.* **2013**, *14*, S5.

(16) Eyrisch, S.; Helms, V. Transient Pockets on Protein Surfaces Involved in Protein- Protein Interaction. *J. Med. Chem.* **2007**, *50*, 3457–3464.

(17) Tilton, R. F.; Kuntz, I. D.; Petsko, G. A. Cavities in Proteins: Structure of a Metmyoglobin Xenon Complex Solved to 1.9 Å. *Biochemistry* **1984**, *23*, 2849–2857.

(18) Elber, R.; Karplus, M. Enhanced Sampling in Molecular Dynamics: Use of The Time-Dependent Hartree Approximation for a Simulation of Carbon Monoxide Diffusion Through Myoglobin. *J. Am. Chem. Soc.* **1990**, *112*, 9161–9175.

(19) Elber, R. Ligand Diffusion in Globins: Simulations Versus Experiment. *Curr. Opin. Struct. Biol.* **2010**, *20*, 162–167.

(20) Desdouits, N.; Nilges, M.; Blondel, A. Principal Component Analysis Reveals Correlation of Cavities Evolution and Functional Motions in Proteins. *J. Mol. Graphics Modell.* **2015**, *55*, 13–24.

(21) Brunori, M.; Gibson, Q. H. Cavities and Packing Defects in the Structural Dynamics of Myoglobin. *EMBO Rep.* **2001**, *2*, 674–679.

(22) Ruscio, J. Z.; Kumar, D.; Shukla, M.; Prisant, M. G.; Murali, T.; Onufriev, A. V. Atomic Level Computational Identification of Ligand Migration Pathways between Solvent and Binding Site in Myoglobin. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 9204–9209.

(23) Bossa, C.; Amadei, A.; Daidone, I.; Anselmi, M.; Vallone, B.; Brunori, M.; Di Nola, A. Molecular Dynamics Simulation of Sperm Whale Myoglobin: Effects of Mutations and Trapped CO on The Structure and Dynamics of Cavities. *Biophys. J.* **2005**, *89*, 465–474.

(24) Tomita, A.; Sato, T.; Ichiyanagi, K.; Nozawa, S.; Ichikawa, H.; Chollet, M.; Kawai, F.; Park, S.-Y.; Tsuduki, T.; Yamato, T.; Koshihara, S.-Y.; Adachi, S.-I. Visualizing Breathing Motion of Internal Cavities in Concert with Ligand Migration in Myoglobin. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 2612–2616.

(25) Scorciapino, M. A.; Robertazzi, A.; Casu, M.; Ruggerone, P.; Ceccarelli, M. Breathing Motions of a Respiratory Protein Revealed by Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2009**, *131*, 11825–11832.

(26) Gabba, M.; Abbruzzetti, S.; Spyrakis, F.; Forti, F.; Bruno, S.; Mozzarelli, A.; Luque, F. J.; Viappiani, C.; Cozzini, P.; Nardini, M.; Germani, F.; Bolognesi, M.; Moens, L.; Dewilde, S. CO Rebinding Kinetics and Molecular Dynamics Simulations Highlight Dynamic Regulation of Internal Cavities in Human Cytoglobin. *PLoS One* **2013**, *8*, e49770.

(27) Nayal, M.; Honig, B. On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 892–906.

(28) Pérot, S.; Sperandio, O.; Miteva, M. a.; Camproux, A.-C.; Villoutreix, B. O. Druggable Pockets and Binding Site Centric Chemical Space: A Paradigm Shift in Drug Discovery. *Drug discovery today* **2010**, *15*, 656–667.

(29) Claußen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient Molecular Docking Considering Protein Structure Variations. *J. Mol. Biol.* **2001**, *308*, 377–395.

(30) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **2004**, *47*, 5076–5084.

(31) Cosconati, S.; Marinelli, L.; Di Leva, F. S.; La Pietra, V.; De Simone, A.; Mancini, F.; Andrisano, V.; Novellino, E.; Goodsell, D. S.; Olson, A. J. Protein Flexibility in Virtual Screening: the Bace-1 Case Study. *J. Chem. Inf. Model.* **2012**, *52*, 2697–2704.

(32) Laine, E.; Goncalves, C.; Karst, J. C.; Lesnard, A.; Rault, S.; Tang, W.-J.; Malliavin, T. E.; Ladant, D.; Blondel, A. Use of Allostery to Identify Inhibitors of Calmodulin-Induced Activation of Bacillus Anthracis Edema Factor. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 11277–82.

(33) Krone, M.; Kozlíková, B.; Lindow, N.; Baaden, M.; Baum, D.; Parulek, J.; Hege, H.-C.; Viola, I. Visual Analysis of Biomolecular Cavities: State of the Art. *Comput. Graph. Forum* **2016**, *35*, 527–551.

(34) Simões, T.; Lopes, D.; Dias, S.; Fernandes, F.; Pereira, J.; Jorge, J.; Bajaj, C.; Gomes, A. Geometric Detection Algorithms for Cavities on Protein Surfaces in Molecular Graphics: a Survey. Computer Graphics Forum. 2017; pp 643–683.

(35) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379–400.

(36) Connolly, M. L. Solvent-Accessible Surfaces of Proteins and Nucleic Acids. *Science* **1983**, *221*, 709–713.

(37) Desdouits, N. Concepts et Méthodes d'Analyse Numérique de la Dynamique des Cavités au Sein des Protéines et Applications à l'Élaboration de Stratégies Novatrices d'Inhibition. **2015**, https://www.theses.fr/2015PA066250 (accessed May 17, 2016), https://tel.archives-ouvertes.fr/tel-01316546/document, (accessed May 17, 2016), Bio-Informatique, Biologie Systémique [q-bio.QM]. Université Pierre et Marie Curie - Paris VI, 2015. Français. <NNT: 2015PA066250>. <tel-01316546>.

(38) Laine, E.; Yoneda, J. D.; Blondel, A.; Malliavin, T. E. The Conformational Plasticity of Calmodulin upon Calcium Complexation Gives a Model of its Interaction with the Oedema Factor of Bacillus Anthracis. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1813–1829.

(39) Larsen, B.; Aone, C. Fast and Effective Text Mining Using Linear-Time Document Clustering. *Proc. fifth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '99* **1999**, 16–22.

(40) Rockafellar, R. T.; Wets, R. J.-B. *Variational Analysis*; Springer Science & Business Media, 2009; Vol. 317.

(41) Hubbard, S. J.; Argos, P. Cavities and Packing at Protein Interfaces. *Protein Sci.* **1994**, *3*, 2194–206.

(42) Hubbard, S. J.; Gross, K.-H.; Argos, P. Intramolecular cavities in globular proteins. *Protein Eng., Des. Sel.* **1994**, *7*, 613–626.

(43) Sonavane, S.; Chakrabarti, P. Cavities and Atomic Packing in Protein Structures and Interfaces. *PLoS Comput. Biol.* **2008**, *4*, e1000188.

(44) Modis, Y.; Ogata, S.; Clements, D.; Harrison, S. C. A Ligand-Binding Pocket in the Dengue Virus Envelope Glycoprotein. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 6986–91.

(45) Poh, M. K.; Yip, A.; Zhang, S.; Priestle, J. P.; Ma, N. L.; Smit, J. M.; Wilschut, J.; Shi, P.-Y.; Wenk, M. R.; Schul, W. A Small Molecule Fusion Inhibitor of Dengue Virus. *Antiviral Res.* **2009**, *84*, 260–6.

(46) Cowan-Jacob, S. W.; Fendrich, G.; Floersheimer, A.; Furet, P.; Liebetanz, J.; Rummel, G.; Rheinberger, P.; Centeleghe, M.; Fabbro, D.; Manley, P. W. Structural Biology Contributions to the Discovery of Drugs to Treat Chronic Myelogenous Leukaemia. *Acta Crystallogr., Sect. D: Struct. Biol.* **2007**, *63*, 80–93.

(47) Zhang, J.; Adrián, F. J.; Jahnke, W.; Cowan-Jacob, S. W.; Li, A. G.; Iacob, R. E.; Sim, T.; Powers, J.; Dierks, C.; Sun, F.; Guo, G.-R.; Ding, Q.; Okram, B.; Choi, Y.; Wojciechowski, A.; Deng, X.; Liu, G.; Fendrich, G.; Strauss, A.; Vajpai, N.; Grzesiek, S.; Tuntland, T.; Liu, Y.; Bursulaya, B.; Azam, M.; Manley, P. W.; Engen, J. R.; Daley, G. Q.; Warmuth, M.; Gray, N. S. Targeting Bcr-Abl by Combining Allosteric with ATP-Binding-Site Inhibitors. *Nature* **2010**, *463*, 501–6.

(48) Drum, C. L.; Yan, S.-Z.; Bard, J.; Shen, Y.-Q.; Lu, D.; Soelaiman, S.; Grabarek, Z.; Bohm, A.; Tang, W.-J. Structural Basis for the Activation of Anthrax Adenylyl Cyclase Exotoxin by Calmodulin. *Nature* **2002**, *415*, 396–402.

(49) Yennamalli, R.; Subbarao, N.; Kampmann, T.; McGeary, R. P.; Young, P. R.; Kobe, B. Identification of Novel Target Sites and an Inhibitor of The Dengue Virus E Protein. *J. Comput. Aided. Mol. Des.* **2009**, *23*, 333–41.

(50) Nagar, B.; Hantschel, O.; Young, M. A.; Scheffzek, K.; Veach, D.; Bornmann, W.; Clarkson, B.; Superti-Furga, G.; Kuriyan, J. Structural Basis for the Autoinhibition of c-Abl Tyrosine Kinase. *Cell* **2003**, *112*, 859–871.

(51) Schmidtke, P.; Bidon-Chanal, A.; Luque, F. J.; Barril, X. MDpocket: Open-Source Cavity Detection and Characterization on Molecular Dynamics Trajectories. *Bioinformatics* **2011**, *27*, 3276–3285.

(52) Krone, M.; Falk, M.; Rehm., S.; Pleiss, J.; Ertl, T. Interactive Exploration of Protein Cavities. *Comput. Graph. Forum* **2011**, *30*, 673–682.

(53) Krone, M.; Reina, G.; Schulz, C.; Kulschewski, T.; Pleiss, J.; Ertl, T. Interactive Extraction and Tracking of Biomolecular Surface Features. Computer Graphics Forum. 2013; pp 331–340.

(54) Ashford, P.; Moss, D. S.; Alex, A.; Yeap, S. K.; Povia, A.; Nobeli, I.; Williams, M. A. Visualisation of Variable Binding Pockets on Protein Surfaces by Probabilistic Analysis ff Related Structure Sets. *BMC Bioinf.* **2012**, *13*, 39.

(55) Lindow, N.; Baum, D.; Bondar, A.-N.; Hege, H.-C. Exploring Cavity Dynamics in Biomolecular Systems. *BMC Bioinf.* **2013**, *14*, S5.

(56) Kokh, D. B.; Richter, S.; Henrich, S.; Czodrowski, P.; Rippmann, F.; Wade, R. C. TRAPP: A Tool for Analysis of Transient Binding Pockets in Proteins. *J. Chem. Inf. Model.* **2013**, *53*, 1235–1252, PMID: 23621586.

(57) Paramo, T.; East, A.; Garzón, D.; Ulmschneider, M. B.; Bond, P. J. Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: trj_cavity. *J. Chem. Theory Comput.* **2014**, *10*, 2151–2164.

(58) Durrant, J. D.; Votapka, L.; Sørensen, J.; Amaro, R. E. POVME 2.0: an Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory Comput.* **2014**, *10*, 5047–5056.

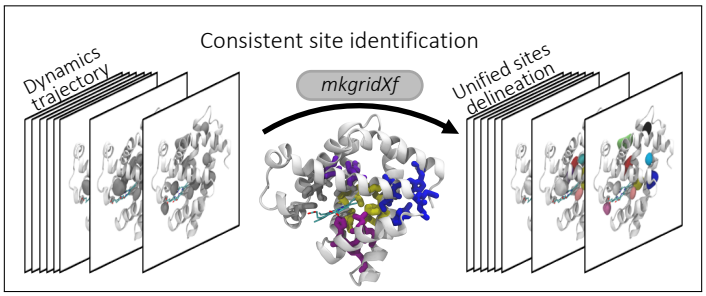
(59) Laurent, B.; Chavent, M.; Cragnolini, T.; Dahl, A. C. E.; Pasquali, S.; Derreumaux, P.; Sansom, M. S.; Baaden, M. Epock: Rapid Analysis of Protein Pocket Dynamics. *Bioinformatics* **2014**, *31*, 1478–1480.

(60) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; ; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.

(61) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. a.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. a.; Fraser, J. S.; Sali, A. Cryptosite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Biol.* **2016**, *428*, 709–719.

(62) Beglov, D.; Hall, D. R.; Wakefield, A. E.; Luo, L.; Allen, K. N.; Kozakov, D.; Whitty, A.; Vajda, S. Exploring the Structural Origins of Cryptic Sites on Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E3416–E3425.

Graphical TOC Entry



Graphical TOC