TRAPP: a Tool for Analysis of Transient Binding Pockets in Proteins

Daria B. Kokh[11] , Stefan Richter[1], Stefan Henrich[1], Paul Czodrowski[2], Friedrich Rippmann[2], Rebecca C. Wade[1,3],*

[1]Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany; [2] Global Computational Chemistry, Merck Serono, Merck KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany; [3] Zentrum für Molekulare Biologie (ZMBH), Heidelberg University, 69120 Heidelberg, Germany

Supplementary Information

1. **TRAPP workflow**
2. **Sequence alignment of structures and superimposition of snapshots of a trajectory/ensemble**
3. **Robustness of the pocket shape identification procedure**
    a. **Effect of hydrogen atoms on the pocket shape**
    b. **Effect of the superposition procedure and grid position/orientation on the pocket shape**
4. **Clustering procedure**
5. **Comparison of TRAPP with other programs for analysis of transient/conserved binding pockets**

[1] Corresponding authors: daria.kokh@h-its.org and rebecca.wade@h-its.org

1. TRAPP workflow

A diagram of the TRAPP workflow is shown in **Fig.1S.**

In all the presented calculations, the following (default) parameters for the pocket shape identification procedure were used:

Parameter	Value
Grid spacing (Å)	0.75
σ_{Prot} (Å)	1.2, 1.5, and 1.7 for H, O, and other atom-types respectively
σ_{Lig} (Å)	1.5
Total number of vectors	64/144 (for 1st/2nd step)
α_0	0.6
$\|\boldsymbol{u}\|$ (Å)	7/3.5 (for 1st/2nd step)
Atomic radius for pocket averaging (Å)	3
G_{vdW}	0.6

2. Sequence alignment of structures and superimposition of snapshots of a trajectory/ensemble.

A trajectory or ensemble may represent conformational changes with respect to a reference structure. The structures may have mutations or missing residues with respect to the reference structure. In this case, a new set of binding site residues is generated for each ensemble or trajectory using a sequential alignment procedure.

The alignment procedure is implemented in TRAPP as follows:

(i) The binding site residues of a reference structure are defined and saved as a set of N_{ref} residue numbers {NU} and names {NA}.

(ii) The coordinate file of a representative structure from a trajectory or ensemble (for example, the 1st snapshot) is screened until the same residue name as the first element of {NA} is found. Starting from this residue, all residues of the representative

structure are sequentially compared with the list {NA}. If the name of a residue and its relative position (counted from the "starting residue") agree with the corresponding residue in lists {NU} and {NA}, the "index of the site similarity", $N_{sitesim}$, is increased by 1. The procedure is finished and the structure is aligned if $N_{sitesim} = (N_{ref} - 1)$.

(iii) If $N_{sitesim} < (N_{ref} - 1)$, the next residue with the same residue name as the first element of {NA} is found and the procedure (ii) is repeated. If $N_{sitesim} = (N_{ref} - 1)$, the alignment procedure is finished and the corresponding set of residues is defined as the binding site of the structures of the trajectory/ensemble analyzed.

(iv) The procedure (iii) is repeated until all residues with the same residue name as the first element of {NA} have been found in the representative structure.

(v) If the largest $N_{sitesim} < (N_{ref} - 1)$, the complete procedure (ii) - (iv) is repeated, but without the first residue from the list {NA}/{NU}.

(vi) At the end, the set of residues that gives the largest value of $N_{sitesim}$ is defined as the binding site of the structures of the trajectory/ensemble analyzed. A pairwise comparison of the reference binding site residues and those identified by the alignment procedure is shown in the program output.

When a set of binding site residues has been generated for each trajectory/ensemble, all snapshots are superimposed with the reference structure using the corresponding binding site residues. The BioPython function "Bio.PDB.Superimposer()" [33] is used for this purpose.

Using only binding site residues for superimposition enables the pocket analysis to be more robust than using all residues. In particular, large conformational changes of some elements in a protein structure may affect the position of a binding site on a grid if all the protein residues are used for superimposition, see Fig. 2S.

3. Robustness of the pocket shape identification procedure

a. Effect of hydrogen atoms on the pocket shape

Hydrogen atoms can be generated by the user or using TRAPP (external call of pdb2gmx GROMACS tool [40]). Alternatively, TRAPP can be run using protein structures without hydrogen atom positions assigned. In the latter case, a fixed Lennard Jones radius ($\sigma_{Prot} =$ 1.8Å) is used for all atom types in a protein structure. Since the bonds to H atoms are about 1 Å long, the uncertainty in the computed shape of the pockets and the identification of transient and conserved regions due to the lack of H atoms or the assignment of different hydrogen atom positions is less than ~1 Å . The results of TRAPP calculations for a structure of P38 with and without hydrogen atoms are shown in **Fig.3S**.

b. Effect of the superposition procedure and grid position/orientation on the pocket shape.

To reduce sensitivity to the pocket position on the grid, the most stable elements of the binding site are used for superimposition. Only backbone atoms are employed for superposition of snapshots. If most of the backbone around a binding site is stable, a change in the set of binding site residues does not affect the pocket shape notably. Moreover, any rotation or translation of the grid relative to the protein should not cause more than a grid spacing's difference in the pocket shape. This is illustrated in Fig. 4S where a pocket in a reference structure and identified transient pockets are shown for two different sets of binding site residues (9 and 46 residues) and, therefore, two different grids (the 9 and 46 residues correspond to the small and large grids, respectively, whose origins are 6.45 grid spacings apart). Computations are done using 15 MD snapshots in which only motion of the β-sheet is observed (sub-pocket A). One can see, however, that some difference appears at the grid boundary because the two last slabs (about one Lennard-Jones radius) of the grid are not

included in the calculations (otherwise the procedure for pocket shape smoothing would not work).

4. Clustering procedure

The clustering procedure is used to split transient pocket regions at a particular level of occurrence (or iso-value $G_0^t(\boldsymbol{r})$) into sub-pockets or sub-regions. This splitting enables the user to easily trace the opening or closing of a particular transient sub-region along a protein trajectory or in an ensemble of protein structures. The clustering procedure consists of two steps: (1) dividing the transient region into sub-pockets that do not contact each other using a *hierarchical clustering* procedure; and (2) additional splitting of the large sub-pockets found in the first step into smaller regions using a *k-means clustering* procedure, in which small sub-pockets found in the first step remain unchanged (default volume of a large pocket to be splitted is above ~350 Å^3 for a default grid spacing of 0.75 Å). For each cluster found, the overlap with the binding pocket of each snapshot is calculated and visualized as a matrix or as a function (**Fig.5S** A' and B'). The overlap value equals 1 if the transient region is completely open in a particular structure.

The clustering procedure is implemented as follows:

1) Points of the distribution $G^t(\boldsymbol{r})$ mapped on a 3D grid are scanned and grid points where $G^t(\boldsymbol{r}) >= G_0^t(\boldsymbol{r})$ are stored as an array of pointers to grid nodes that is used in the clustering procedure. In the clustering procedure, several sub-sets of pointers are generated, each describing a separate transient sub-pocket.

2) *Hierarchical clustering* is used to split transient regions into sub-pockets that do not contact each other (the result of clustering is illustrated in Fig.5B):

(i) Initially, each point is assigned to a separate cluster.

(ii) The distance between each point *(i,j,k)* of a particular cluster and each point of another cluster *(i',j',k')* is compared sequentially and if they directly contact each other ($\sqrt[2]{(i-i')^2+(j-j)^2+(k-k)^2} < 2$), these two clusters are merged.

(iii) The procedure is repeated until no contacts between clusters are found.

(iv) Sub-pockets smaller than 30 grid cells (~12.66 $Å^3$ with a grid spacing of 0.75 Å that corresponds to a pocket of about 2.3 Å radius) are eliminated.

3) *k-means clustering* is used for splitting large sub-pockets into compact regions (the result of clustering is illustrated in Fig. 5C).

(i) The starting number of centers is defined as a pocket size (number of grid cells) divided by a minimum pocket size (that is 150 grid cells or about 63 $Å^3$ if the grid spacing is 0.75 Å , which corresponds to a pocket of about 3 Å radius and is the default minimum pocket size). Positions of the centers are chosen randomly, but the distance between two centers must not be smaller than 10 Å, otherwise a new center is chosen.

(ii) Each point of a sub-pocket is assigned to the nearest cluster center;

(iii) The geometric center of each cluster is re-calculated; if a new center is outside the cluster, the nearest cluster point is used as a new center.

(iv) If the distance between two cluster centers is less than 1.5 times the average distance from the center to all points of a cluster, these two clusters are merged; if the distance between cluster centers is less than 20 grid spacings (15 Å with a grid spacing of 0.75 Å), they are merged; if the cluster is smaller than double the minimum pocket

size (300 grid cells which corresponds to a volume of ~ 126 Å ^{3} with a grid spacing of 0.75 Å), it is merged with the nearest cluster.

5. Comparison of TRAPP with other programs for analysis of transient or conserved binding pockets

(i) TRAPP focuses on one selected binding site, whereas most other methods scan the complete protein for possible binding cavities. For this reason, the procedure of pocket shape identification in TRAPP is designed to reproduce precisely the physical pocket boundary, and therefore, even small roughnesses on the pocket surface as well as small sub-pockets of the size of only a few heavy atoms. In contrast, other methods have to ensure fast screening of the complete protein structure and finding potential binding pockets (that may accommodate a small ligand) often at the expense of accuracy of the pocket boundary description.

(ii) The new pocket detection procedure implemented in TRAPP enables identification of the shapes of all types of cavities without parameter adjustment (see paper, Sec. Introduction). A comparison of the conserved and transient regions generated by TRAPP and MDPocket [17] (we have used Mobyle@RPBS webserver: *http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py?form=mdpocket#forms::mdpocket*) using default parameters for both programs is shown in Fig. 6S for IL-2. Whereas the FPocket [27] algorithm used in the MDPocket program identifies mostly well-buried, relatively large cavities, TRAPP detects open shallow pockets on the protein surface and small completely buried pockets equally well.

(iii) Our tests with FPocket and EPOS [13] showed that a small variation of an atom position in a binding site may induce a relatively large variation in the computed pocket shape. One possible reason is the cavity representation by pseudo-atoms, which leads to an uncertainty in the pocket boundary of about the radius of the pseudo-atom used. Another probable reason in

the case of the alpha-sphere algorithm used in the FPocket [27] detection program, is that the shift of the alpha-sphere center may under certain conditions become notably larger than the movement of the protein atom causing this shift. This shortcoming is partially overcome by setting minimum and maximum radius thresholds for the α-spheres and by performing several steps of clustering on several tens of thousands of alpha spheres, which leads to an effective averaging of the pocket curvature over the α-spheres built on many neighboring atoms of a binding site. In contrast, in the TRAPP algorithm, a shift of the position of a binding site atom leads directly to the same change of the corresponding pocket boundary.

(iv) TRAPP uses superposition of protein structures based on binding site residues only, which enables some of the deficiencies of grid-based approaches for protein pocket identification to be overcome. This procedure gives the most accurate representation of the motion of a particular binding pocket, independent of structural variations of the rest of protein (see **Fig. 2S**). Most other methods rely upon a superposition of snapshots performed by the user.

(v) We developed a new definition of the transient pocket parts (ARDR), which uses a reference protein structure. This procedure is designed specifically for a complete workflow that includes simulation of protein flexibility and then analysis of possible pocket variations (e.g. when one crystal structure is available and possible variation in a binding site must be derived from simulation of protein conformational dynamics).

(vi) Some additional tools are implemented in TRAPP (computation of pocket physico-chemical characteristics, ligand-pocket overlap measure, tracing of pocket opening etc.), which give a basis for further development of TRAPP as a tool for selection of binding compounds taking into account protein flexibility.

The **Reference** numbering is as given in the manuscript's Bibliography

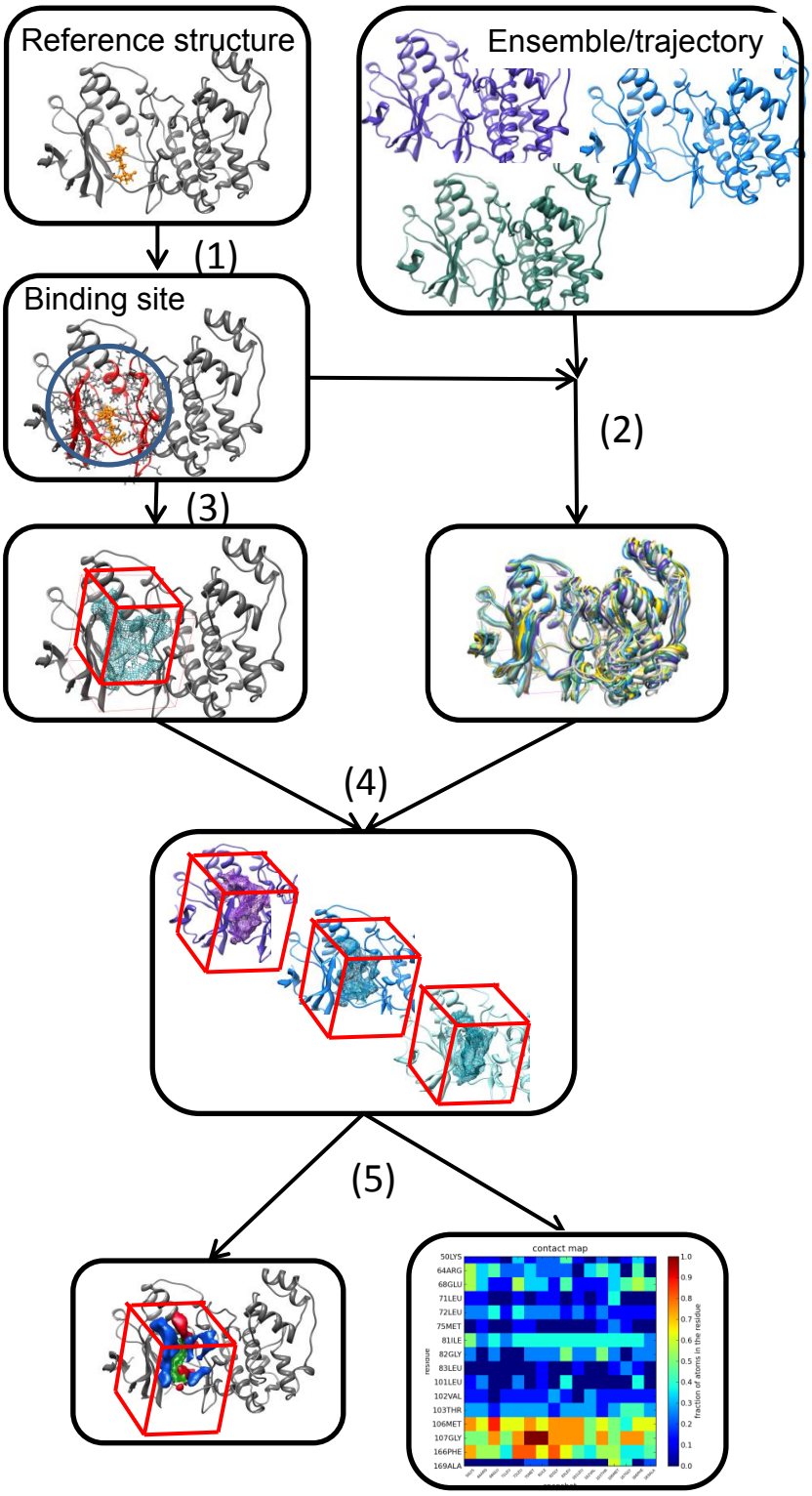

Figure 1S: TRAPP workflow: (1)- a set of binding residues is defined for a reference structure; (2) binding residues are used for sequential alignment and superposition of different structures; (3) a pocket shape in a reference structure is computed; (4) the shape and physicochemical properties of the binding site region are computed for each structure and stored on a grid; (5) A set of pocket shapes is used for analysis of the pocket dynamics and properties (pocket similarity, pocket – ligand complementarity), as well as for data visualization.

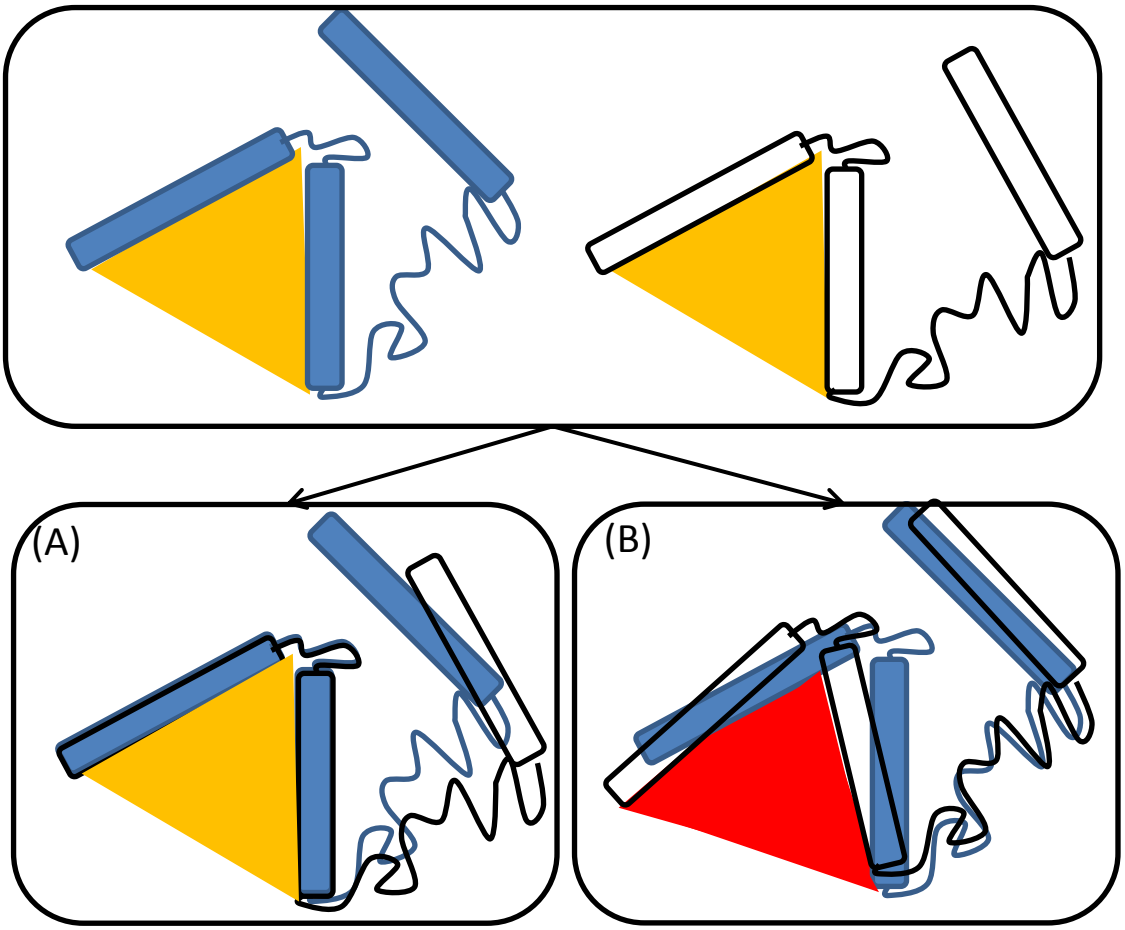

Figure 2S: Illustration of different superposition procedures using (A) pocket or (B) all protein residues. The pocket region is shown in yellow for two protein conformations (blue and white). If superimposition is done using all residues, the position and shape of the pocket identified may change even if the binding site is unchanged. Thus, only part of the pocket will be considered as conserved (shown in red).

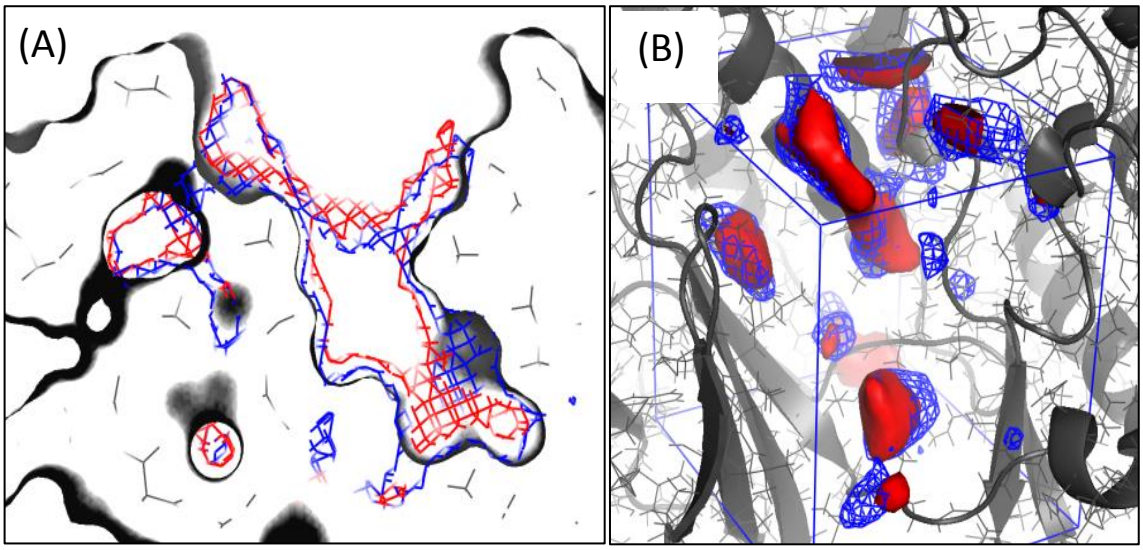

Figure 3S: Illustration of the influence of the assignment of hydrogen atoms on TRAPP results. (A): A 2-D section of the protein surface of the reference structure of P38 generated by PyMol [29] is shown in black; the pocket shapes computed by TRAPP with and without H atoms (but with larger Lennard-Jones radius, see text) are shown by red and blue meshes, respectively. (B): transient pockets shown with the reference structure (red surface and blue mesh show calculations with and without H atoms, respectively)

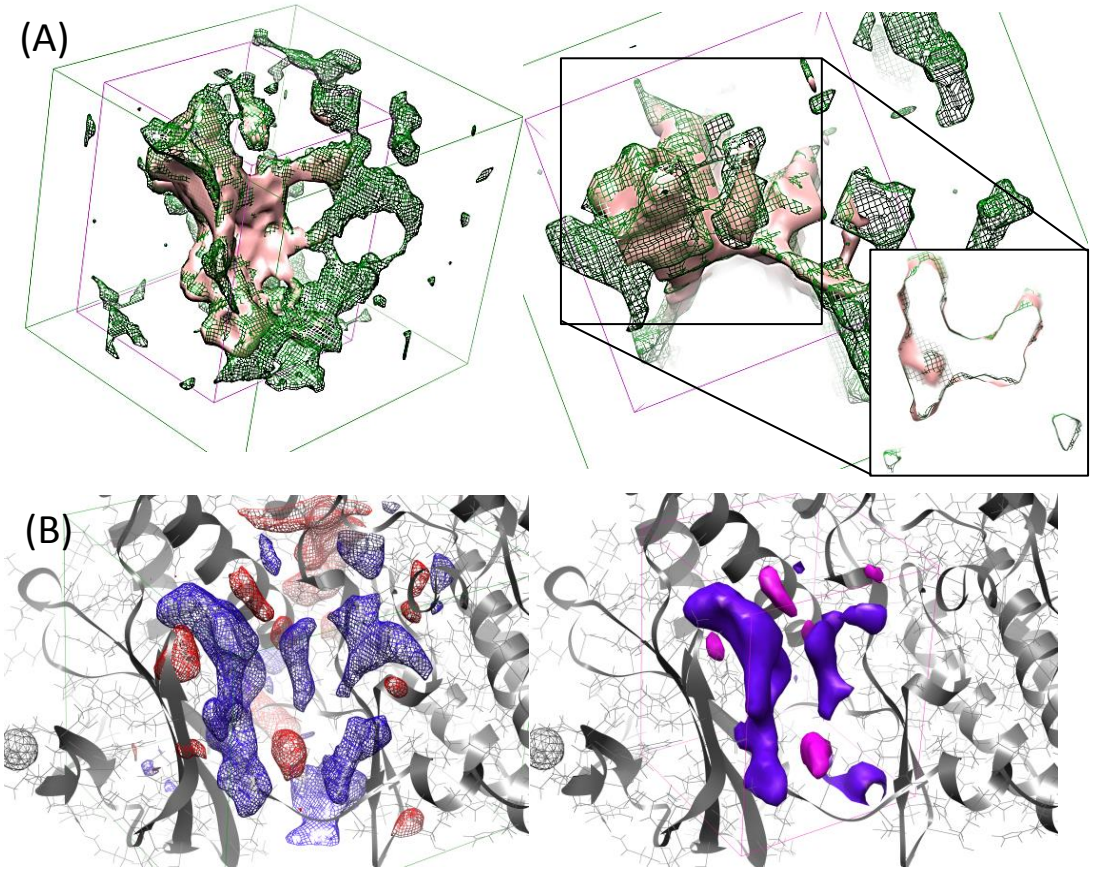

Figure 4S: Illustration of the influence of the assignment of grid position and size on the computed transient pockets. Calculations are shown for one structure on two grids with different origin position and different sizes; the grids have the same grid spacing of 0.75 Å, but as the grid origins are 6.45 grid spacings apart, grid nodes are shifted by about 0.2 Å in the overlapping region. (A): the shape of the binding pocket of the reference structure is shown from two different views and a cross-section of the pocket is shown in the insert (the green mesh and the pink surface correspond to the large grid and the small grid, respectively; the grid boundaries are shown in the corresponding colours); (B): transient pocket regions identified for (left, meshes) the large and (right, surfaces) the small grid; red and blue correspond to appearing and disappearing regions, respectively. Within the small grid, the transient pockets computed are very similar.

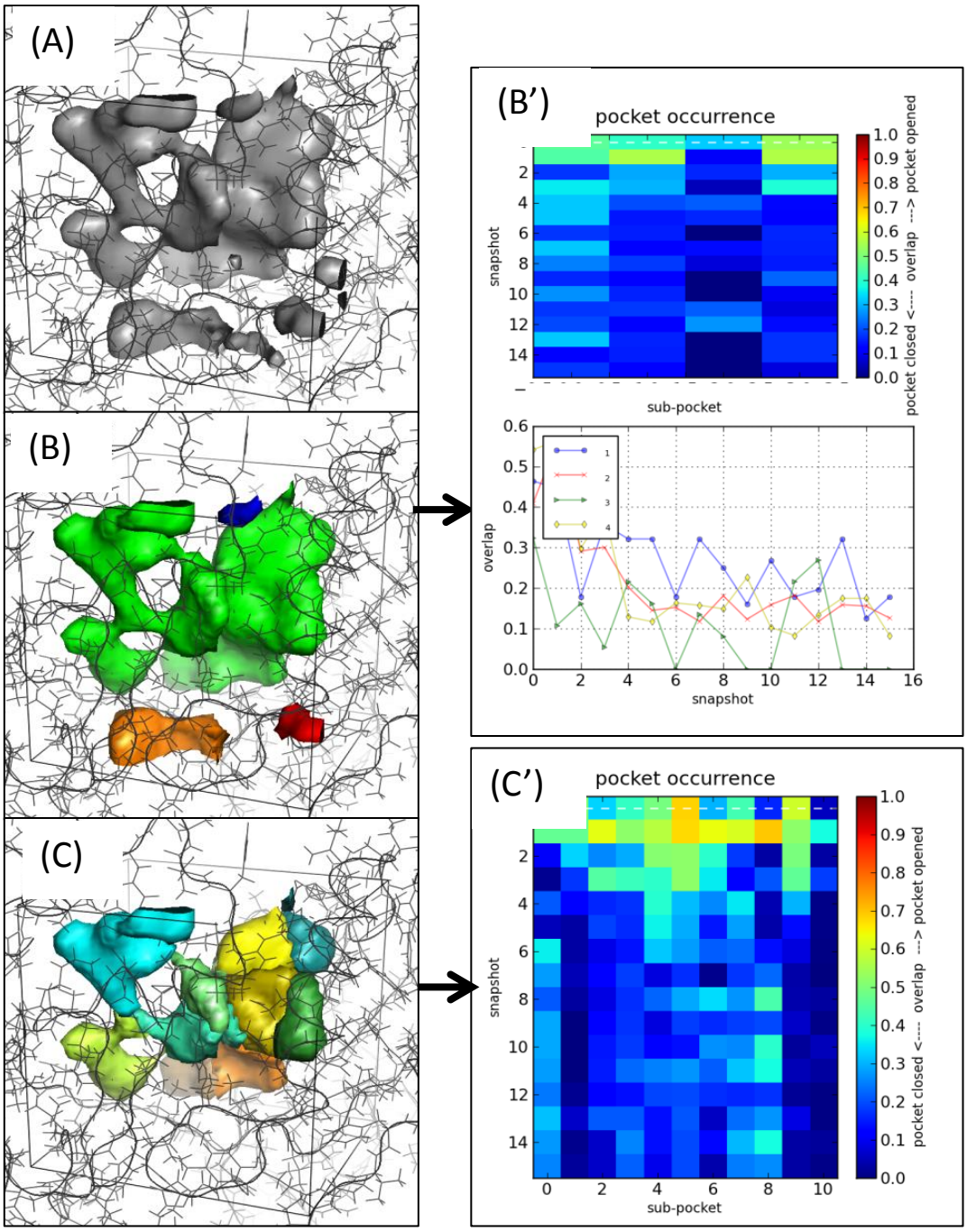

Figure 5S: Illustration of the clustering procedure for pocket analysis. The transient pocket shape was computed using 15 MD snapshots for P38 during which the pocket closed. (A) – all identified transient regions. Transient regions were separated into sub-pockets using (B) hierarchical clustering, and (C) k-means clustering of the large sub-pocket only (small sub-pockets are not shown in this plot). (B') and (C') – Overlap between transient sub-pockets and binding pockets in MD snapshots: (B') for the sub-pockets in (B) and (C') for those in (C).

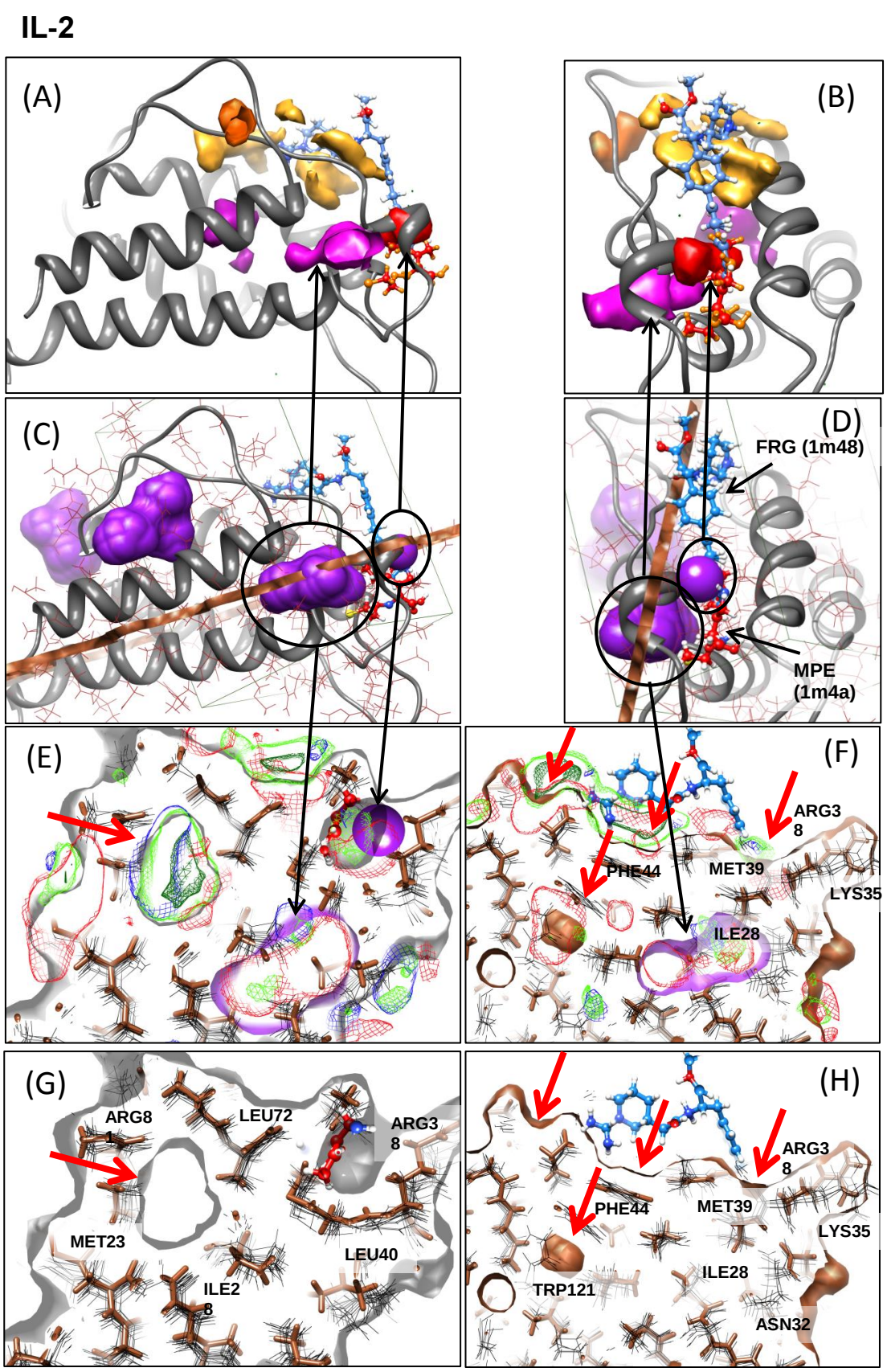

Figure 6S: Comparison of the conserved and transient pockets identified by the TRAPP and MDPocket procedures (using default simulation parameters) for IL2 protein. 20 MD snapshots were used for simulations in

both methods, the first snapshot (obtained after minimization, PDB code 1m48) was considered as a reference structure in ARDR analysis of TRAPP. Upper plots (A/B and C/D) – show two views of the protein reference structure with transient sub-pockets identified by TRAPP (A/B) and pockets identified by MDpocket (C/ D); the lower plots, E/F and G/H, show the protein structure and transient/conserved pockets in two different sections shown in brown in Figs. (C) and (D), respectively. The reference structure is represented by brown sticks; other MD snapshots used for analysis are shown as black wires; two ligands (from 1m48 and 1m4a co-crystallized structures) are shown in ball and stick representation with blue and red carbon atoms, respectively; a protein surface for the first MD snapshot (used as a reference structure in ARDR analysis) is shown as generated by Chimera [35]. About ten transient pocket regions that appear in MD snapshots in addition to the binding pocket of the reference structure were identified by TRAPP; the largest ones are shown in different colours in Fig. A/B). Four pockets were found by MDPocket (violet surfaces generated by Chimera based on the position of pseudo-atoms from MDPocket) and indicate pockets opening in some MD snapshots. Only two of the pockets (shown by circles in Figs. C/D) are close to the binding site analysed by TRAPP. The smallest one (just one pseudoatom center found by MDPocket) is in the direct vicinity of the position occupied by the ligand in structure 1m4a, shown in plot (E). In plots (E) and (F), mesh surfaces indicate regions identified by TRAPP as follows: conserved pocket regions (dark green), pocket shape of the reference structure (light green); transient regions that appear in at least 20% of snapshots and regions that disappear in 20% of the snapshots (but are observed in the reference structure) are shown by red and blue mesh, respectively. Some pockets that are observed in the reference structure and identified by TRAPP, but have not been detected by MDPocket, are shown by red arrows in Figs E/F and G/H.