Supporting Information for

Perturbation Approaches for Exploring Protein Binding Site Flexibility to Predict Transient Binding Pockets

Daria B. Kokh[1], Paul Czodrowski[2], Friedrich Rippmann[2], Rebecca C. Wade[1,3,4]*.*

[1]Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

[2]Global Computational Chemistry, Merck Serono, Merck KGaA, Darmstadt, Germany

[3]Zentrum für Molekulare Biologie, DKFZ-ZMBH Alliance, Heidelberg University, Heidelberg, Germany

[4]Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany

Short title: Protein pocket flexibility from L-RIP and RIPlig

*Corresponding authors: Daria.kokh@h-its.org, Rebecca.wade@h-its.org.

S1. MD simulation of unbound HSP90.

The stability of α-helix3 in HSP90 was additionally explored in 400ns standard MD simulations using the OPLS-AA [3] force field which is known to underestimate the stability of α-helices in peptides [1]. Conventional explicit solvent MD simulations of HSP90 (PDB: 2UYD) were carried out using the GROMACS 4.5.3 software package [2]. In the structure preparation step, the ligand was removed from the protein structure and hydrogen atoms were added. The protein was immersed in a periodic cubic box of TIP3P water molecules extending at least 1 nm beyond the protein surface. The protein structure was energy minimized with 200 steepest descent (SD) steps (for water only), and 500 SD steps and 500 conjugate gradient steps (for the whole system). The system was heated to 300 K in steps of 40 K with 200ps simulation at each step and equilibrated at 300 K for 2 ns in the NVT ensemble (the number of particles N, the volume V, and the temperature T were fixed in simulations). All bond lengths were constrained using the Linear Constraint Solver (LINCS) algorithm [4]. A cut-off of 10 Å was used for non-electrostatic forces and a particle-mesh Ewald method was applied for long-range electrostatic forces. Temperature was maintained with a Nose-Hoover thermostat (τ=0.1 ps) [5]. The RMSD along the 400ns standard MD trajectory is shown in Fig.S2

S2. Placement of the pseudo-ligand in the binding pocket

The positions of the pseudo-ligand in the binding pocket were defined in the following procedure: (i) the pocket shape of the starting reference structure was defined as described in section 2.5 of the main text; (ii) to ensure extensive sampling of the protein pocket, the binding pocket was divided into smaller compact sub-pockets with a size of less than half that of the pseudo-ligand (about 4 Å, a Lennard-Jones radius of 1.6 Å was used); (iii) then only sub-pockets that satisfied the following conditions were selected: the solvent-exposed area was less

than 5% of the total pocket surface area (for IL-2, a value of 20% was used to include all solvent-exposed sub-pockets as well), and the protein-exposed area was more than 70% of the total pocket surface area. The latter condition enables the central part of a large pocket, where the pseudo-ligand cannot contact the protein, to be discarded; (iv) finally, the phenylalanine pseudo-ligand was placed in each selected sub-pocket with its CA-CZ axis aligned with the longest dimension of the sub-pocket (as determined by principle component analysis of the sub-pocket function) with the aromatic ring oriented towards the closest protein atoms (the placement of the pseudo-ligands is illustrated in Figure S4).

S3. Clustering of protein structures by similarity of the binding site conformations

A k-means clustering procedure was applied to analyze the sampling of conformational space in standard MD and perturbation MD trajectories. As a metric for the comparison of two structures, we employed the maximum value of the backbone atom RMSD amongst all the binding site residues. This metric emphasizes the maximum local distortion rather than the average deviation over all binding site residues. A standard k-means clustering procedure requires the number of clusters to be specified as an input parameter. To define this number, we first carried out a single-linkage hierarchical clustering with the same metric and a threshold of 3 Å. To reduce computation time, we did not compare all cluster members. Instead, only the first element of a cluster was used as a reference to decide whether the next structure considered belonged to this cluster or not. After hierarchical clustering, the distances between the mean centers of the generated clusters were compared pairwise and clusters with center-to-center distances less than the clustering threshold value (3Å was used in the present work) were merged into one cluster and a new center of the combined cluster computed. This procedure

provided a rough, but fast estimate of the number of clusters, whose mean centers were then used as starting centers for k-means clustering. After each iteration step of the k-means clustering procedure, the structure with the smallest deviation of the binding site backbone coordinates from the mean value in a cluster was selected as the cluster representative and used as the new k-means center. The maximum change in position of the binding site residues of cluster representative structures generated in two subsequent clustering steps was used as a criterion for convergence, and was set at 0.2 Å.

S4. Simulation time

In the L-RIP approach, perturbation is applied only to residues lining the binding site of interest. Consequently, the number of trajectories to be generated is defined by the size of the binding site. In the present examples, 28 (PR) to 64 (Src) trajectories were used, which is equivalent to about 0.9-2.1 ns or 2.7-6.3 ns of MD simulations for L-RIP-0.1ps or L-RIP-0.3ps, respectively, assuming that the trajectories consist of 300 pulses. For RIPlig, the number of trajectories necessary for exploring binding pocket flexibility is defined by the total number of initial pseudo-ligand positions that usually does not exceed 5-10. Thus, computational time can be reduced by several-fold relative to L-RIP.

Table S1. Details of the setup of the L-RIP and RIPlig simulations used for method evaluation, Sec.3.3.

	IL2	**PR**	**HSP90**	**SRC**
PDB structure used as a reference	1M47	1A28	1UYD	3U4W
Pocket radius (Å)	6	5.5	7	6.5
Binding site residues used in simulations	27GLY-28ILE 31TYR-45TYR 65PRO 69VAL-82PRO	715LEU-719ASN 721LEU-725GLU 755TRP-766ARG 778PHE-782LEU 794PHE 797LEU 801MET 887LEU-891CYS 894THR-895PHE 903VAL 905MET 913ILE	22PHE 26ILE 47GLU-56LEU 91ILE-98MET 102ASP-108GLY 110ILE-111ALA 133PHE-139TYR 150VAL-154HIS 162TRP 170PHE 183GLY-185LYS	272LYS-284GLY 292VAL-298LYS 301THR-302MET 307PHE 323VAL-325LEU 334ILE-348ASP 351LYS 384HIS-394VAL 403ALA-411ILE 416TYR 422ALA-425 PRO
Reference ligand	1M47, 1M4A, 1NBP	1A28	1UYD	3U4W
Number of RIPlig trajectories[1)]	9[2)]	5	6	8
Number of L-RIP trajectories[3)]	28	33	41	64

[1)] Each trajectory was started with the pseudo-ligand positioned in a different part of the pocket, see Figure S4

[2)] Solvent exposed sub-pockets and those outside the central binding pocket were included in simulations.

[3)] Number of L-RIP trajectories is defined by the number of rotatable binding site residues used for perturbation

Table S2. Number of clusters obtained from the hierarchical clustering procedure for the crystal structures (listed in Table S3) and in the different types of simulation.

	IL2	**PR**	**HSP90**	**SRC**
Crystal structures	19	2	6	20
MD	8	1	7	3
RIPlig	7	7	17	13
L-RIP	40	71	44	>66*

* An RMSD threshold of 4Å (instead of 3Å for all the other simulations) was used because of the very large number of clusters

Table S3. Crystal structures used in the simulations

Protein	**PDB ID**
IL2	1M47; 1M48; 1M49; 1M4B; 1M4A; 1NBP; 1PW6; 1PY2; 1QVN; 1M4C
PR	1A28; 1E3G; 1SQN; 1SR7; 1ZUC; 2W8Y; 3D90; 3G80; 3HQ5; 3KBA; 3ZR7; 3ZRA; 3ZRB; 4A2J; 4APU
HSP90	1UYD; 1YES; 3T0H; 3VHD; 2VCJ; 1UYL; 1UYF; 4EEH; 2QFO; 2XAB; 2XDK; 2XDL; 2XDS; 2XDU; 2XJJ; 2XJX; 2XK2; 2YJW; 2YK9; 2YKB; 2YKE; 3HEK; 3HZ5; 3INW; 3K97; 3MNR; 3OW6; 3OWD; 3QTF; 3R91; 3R92; 3RKZ; 3RLP; 3RLQ; 3RLH; 3T1K; 3T2S; 3T10L; 4QWQ; 4B7P; 4DRH; 4DRJ; 4HY6; 4JQL
SRC	3U4W; 1KSW; 1YOL; 1YOM; 2BDF; 2BDJ; 2H8H; 2OIQ; 2SRC; 3EL7; 3EL8; 3EN4; 3EN5; 3F3T; 3F3U; 3F3V; 3SVV; 3QLG; 3QLF; 3G6H; 3G6G

Table S4. Crystal structures used in Figs.4 and 5

Protein	**PDB ID**
HSP90	Helix: 1UYD; 1UYF Loop-in: 4EEH; 4AWP;1YER;2VCJ; 3T0H Loop-out: 1YES; 3T0H; 3VHD; 4B7P; 1YET
SRC	DFG-in: 2SRC; 3U4W; 2PTK; 2H8H; 1KSW; 1FMK; 3SVV DFG-out: 2HW0; 2QQ7; 3OE7; 3F3T; 3F3V;3F3W; 2BDJ; 2OE7; 3EN5; 3EN4

Table S5. RMSD (Å) of HSP90 from three types of conformations in crystal structures (helix, PDB:1UYD; loop-in: PDB:1YER; and loop-out: PDB: 1YES) as observed in one representative structure generated by L-RIP and after its equilibration by 10ns standard explicit solvent MD. For comparison, the RMSD between crystal structures is included. (A): the α-helix3 and (B) a whole protein.

A. α-helix3: residues 100-121

	helix	loop-In	loop-Out
L-RIP	5.32	4.2	4.94
L-RIP+MD	3.04	2.59	4.1
helix		2.67	3.24
loop-In			3.9

B. Protein: residues 17-221

	helix	loop-In	loop-Out
L-RIP	2.74	2.66	2.76
L-RIP+MD	1.9	1.95	2.03
helix		0.98	1.42
loop-In			1.66

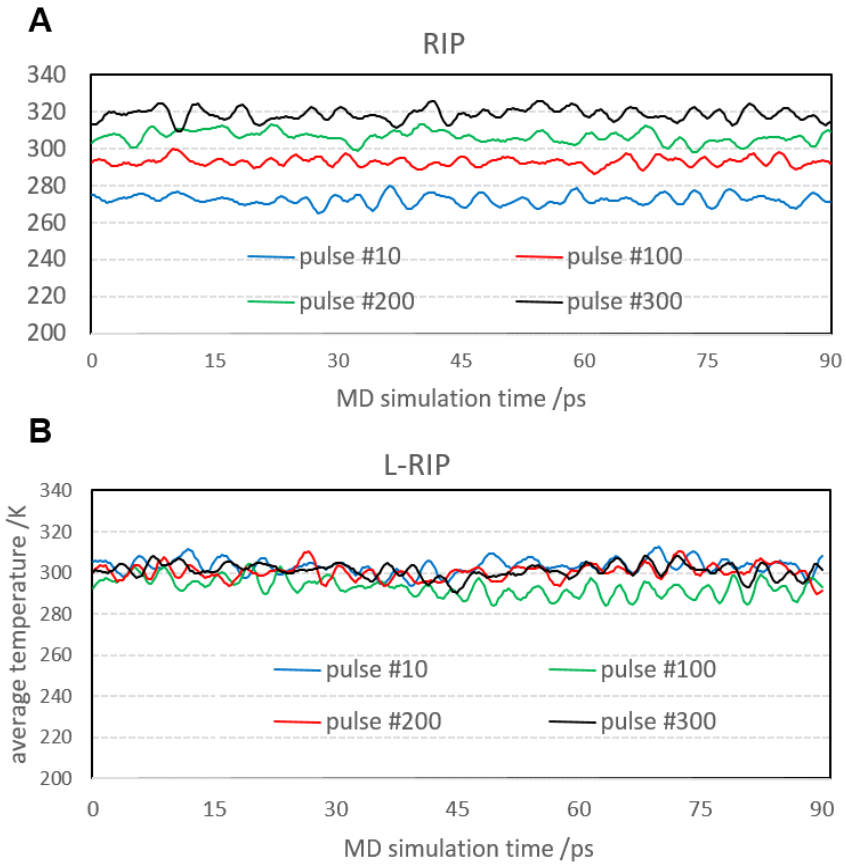

Figure S1 Temperature variation along MD trajectories in RIP (A) and L-RIP (B) simulations of HSP90. The average temperature increases with increasing pulse number in RIP whereas the average temperature is preserved in L-RIP. L107 is perturbed; the length of perturbation pulse used in simulations is 0.3ps.

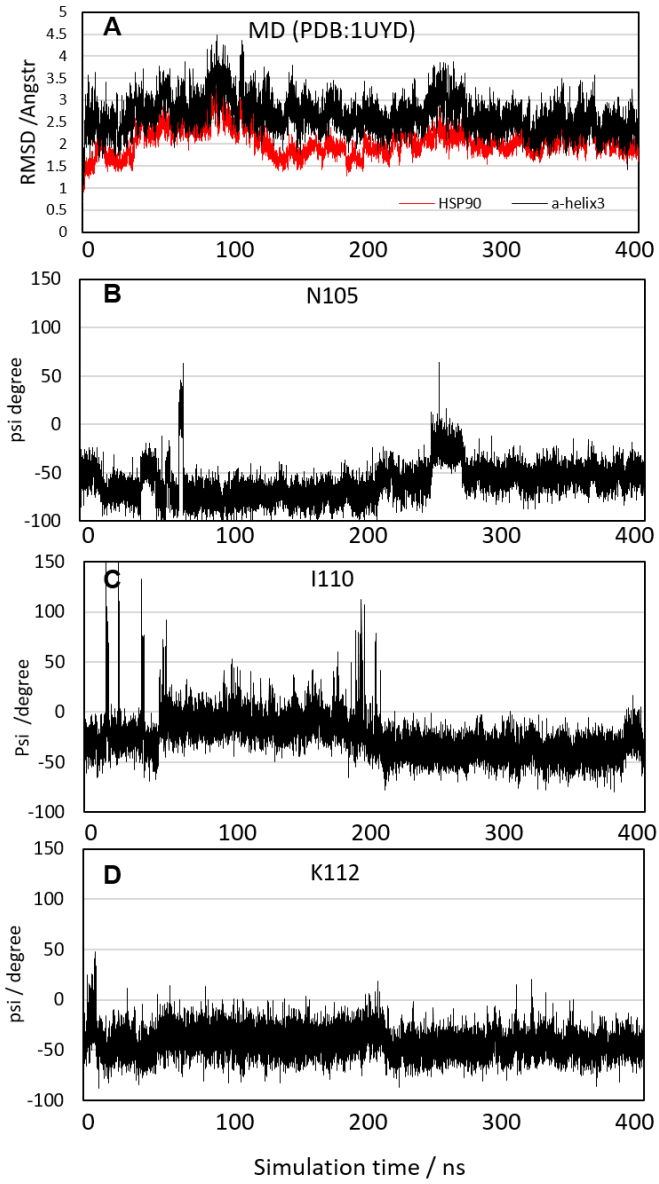

Figure S2 Stability of α-helix3 in 400ns plain MD simulations (OPLS force field used). **A:** Backbone RMSD of HSP90 (plots for complete HSP90-NTD and only α-helix3 are shown by black and red lines, respectively) and in a 400ns MD trajectory starting from a structure with a complete α-helix3 (PDB: 1UYD, the ligand was removed from the structure); **B-D**: Evolution of the conformation of α-helix3 along the same MD trajectory shown by the variation of the ψ dihedral angles (backbone atoms N-C^{α}-C'-N) of N105, I110, and K112 with time. The ψ dihedral angles of the loop-In and loop-Out structures are around 50/75/50 degrees and -50/150/150 degrees (N105/I110/K112), respectively.

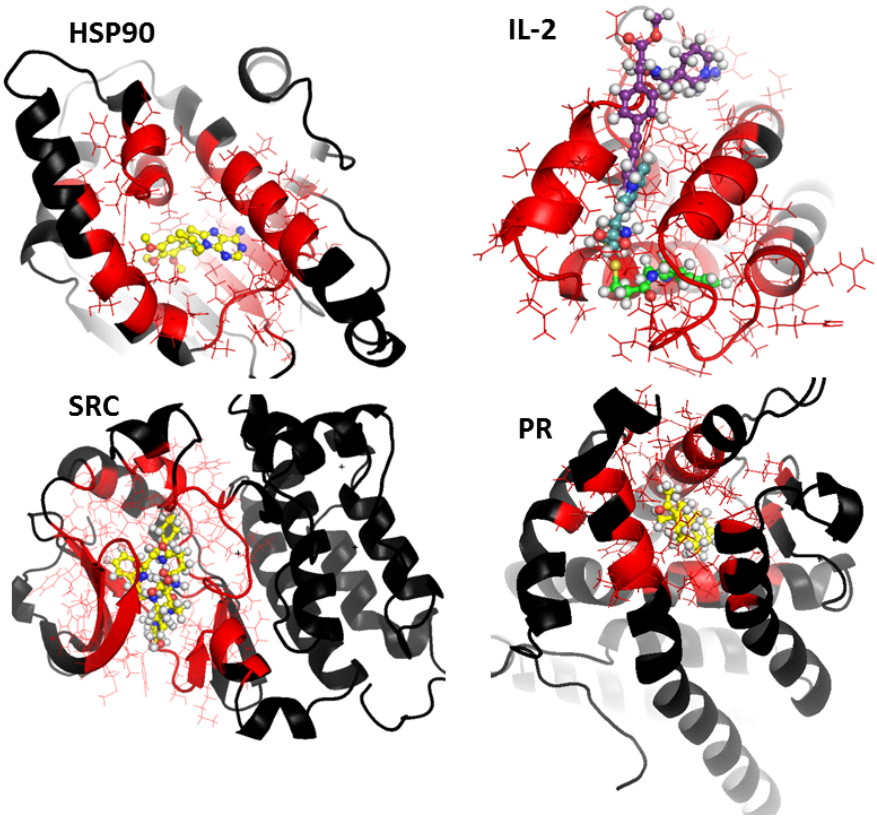

Figure S3 Binding site residues defined for perturbation in the L-RIP simulations for the test targets: the selected binding site residues are shown by lines and colored in red. The ligands for HSP90, SRC, and PR are shown with carbon atoms in yellow; the three ligands used for definition of the binding site in IL2 are colored in green (PDB 1NBP), blue (1M4A) and purple (1M48).

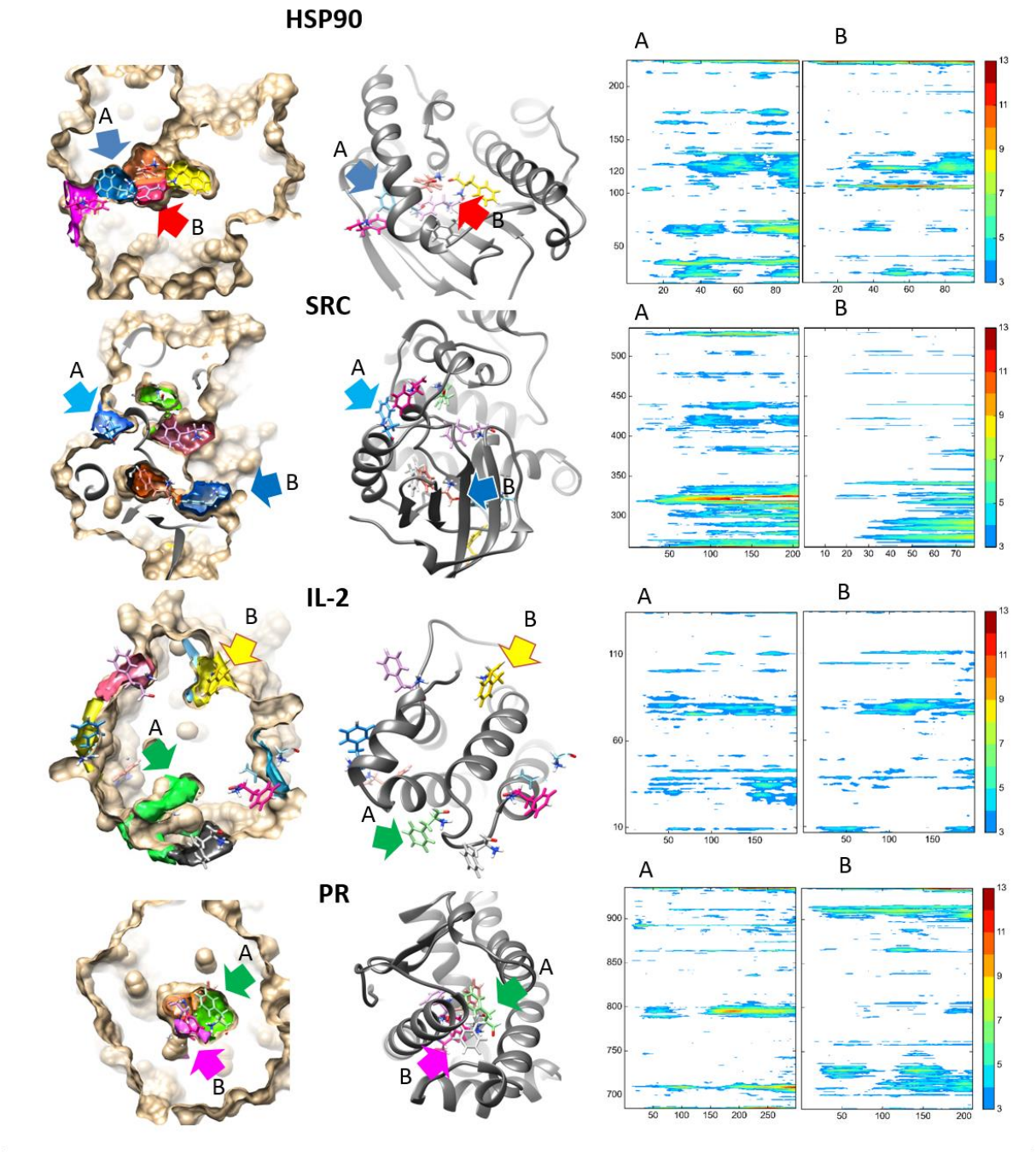

Figure S4 Placement of the pseudo-ligand in different parts of the binding pocket in the RIPlig approach. Left: Sub-pockets for the four test proteins are shown in different colors, whereas the protein surface is shown in wheat. Middle: Corresponding positions of the pseudo-ligands with the protein structure shown in gray cartoon. Right: Residue-wise RMSD relative to the starting structures plotted against pulse number for two (A and B) representative RIPlig trajectories for each target. Trajectories A and B correspond to perturbation of the ligands placed as indicated by the respectively denoted arrows.

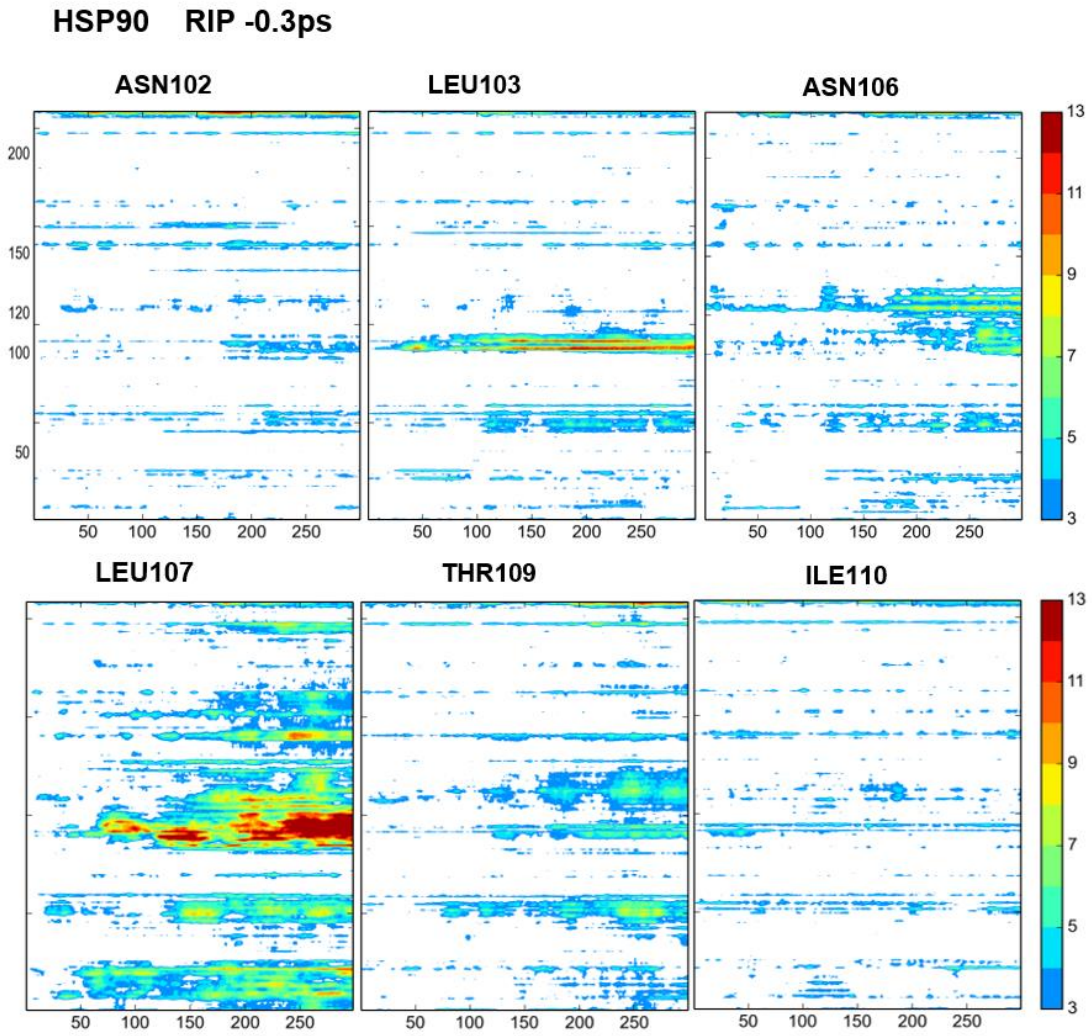

Figure S5 Residue-wise RMSD relative to the starting HSP90 structure with an unperturbed α-helix3 (PDB: 2UYD) against pulse number as observed in 6 RIP trajectories (each with a 0.3ps MD step in each pulse and 300 pulses) upon perturbation of residues of the binding site that located at the α-helix3 (N102, L103, N106, L107, T109, I110).

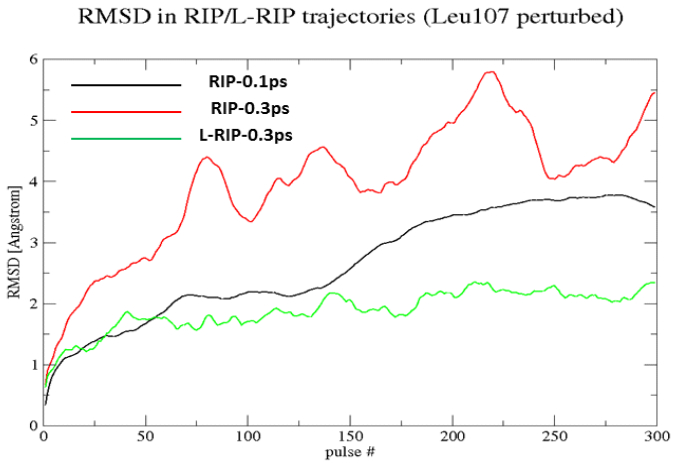

Figure S6 Backbone RMSD of HSP90 along two RIP trajectories (black, red) and one L-RIP trajectory (green) upon perturbation of L107. The overall distortion of the protein is smaller with L-RIP than with RIP.

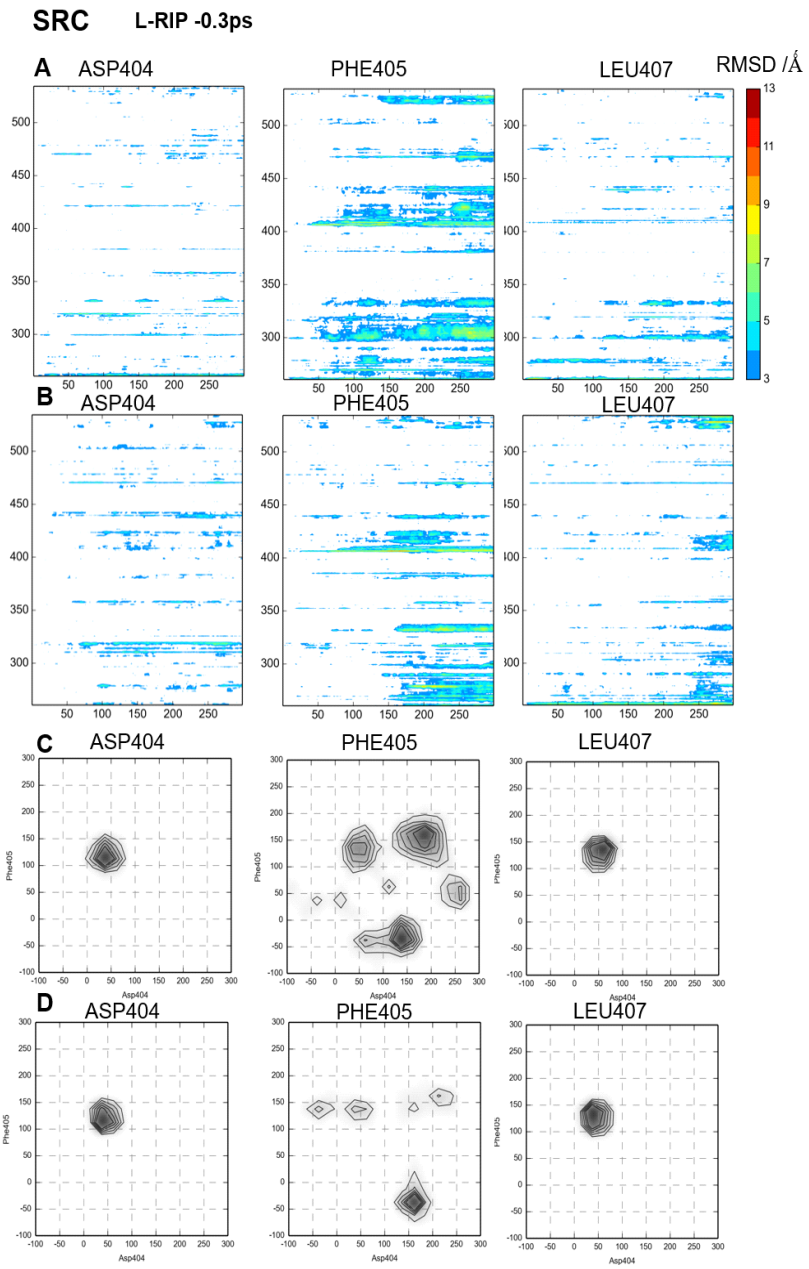

Figure S7 A-B: Residue-wise backbone RMSD relative to the starting SRC structure (PDB: 3U4W) as observed in two series of L-RIP simulations (0.3ps MD step in each pulse; 300 pulses) upon perturbation of three residues: D404 and F405 in the DFG loop, and L407 (G406 was not used for perturbation since its side-chain is not rotatable). C-D: Conformational distribution of the L-RIP structures in two trajectories mapped onto the 2D space of the D404/F405 dihedral angles. Plots A and C show analysis of one series of L-RIP trajectories and B and D of another. These plots demonstrate that perturbation of F405 initiates conformational changes in the DFG loop (the degree of conformational change can vary in different simulation trajectories), while perturbation of D404 or L407 does not.

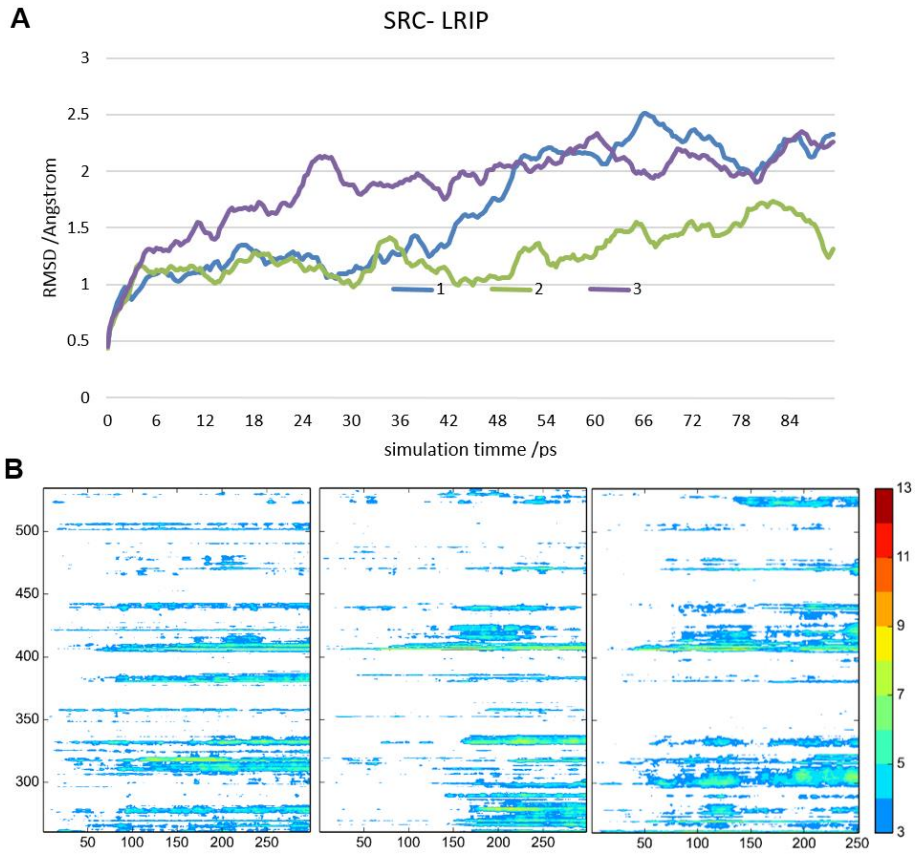

Figure S8. A- backbone RMSD of simulated Src conformations along three different L-RIP trajectories relative to the starting structure (PDB: 3U4W), and B - per-residue RMSD along the same trajectories as shown in (A) (pulse length 0.3ps; the perturbation is applied to F405 in all simulations).

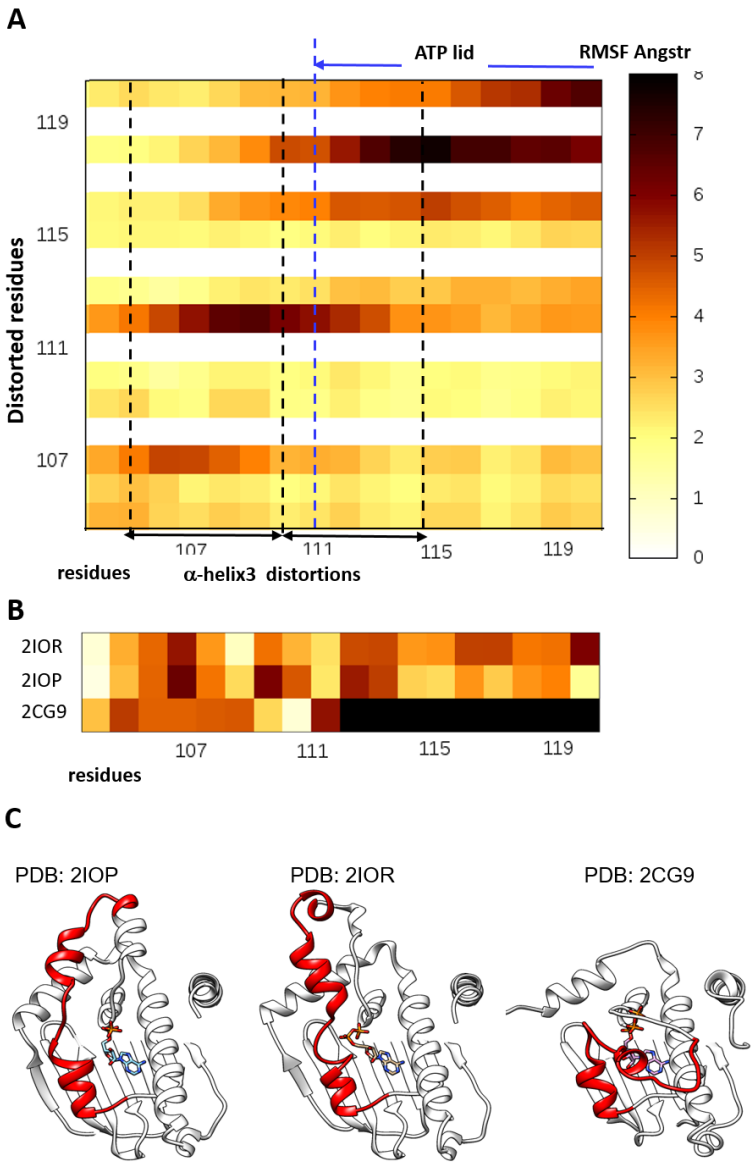

Figure S9 A: Backbone RMSF of the α-helix3 in HSP90 observed in L-RIP trajectories generated by perturbation of residues of α-helix3 (perturbed residues are given on the y axis, while the per-residue response is shown along the x axis). Fluctuation in the region I104-I110 (loop-In conformation) arises mostly from the perturbation of L107, while perturbation of several residues of the ATP lid (specifically, K112, K116, and F120) causes motion of the region N105-T115 (loop-out conformations) and the ATP lid itself. B: backbone RMSD of the α-helix3 from the helical conformation of human HSP90 N-terminal domain (PDB: 2UYD) observed in the bacterial (PDB: 2IOP and 2IOR) and yeast (PDB:2CG9) HSP90 N-terminal domain co-crystallized with ATP/ADP. The corresponding structures are visualized in (C). The α-helix3 segment is colored in red.

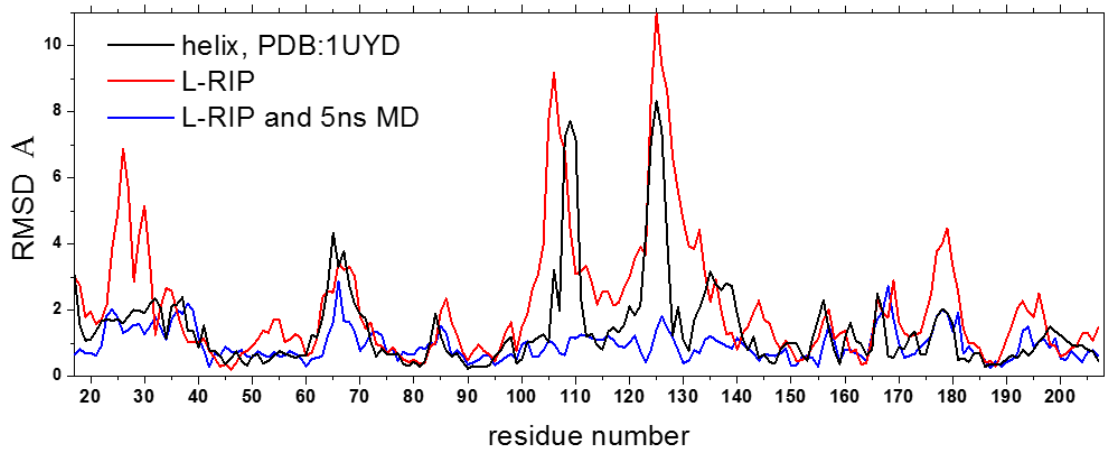

Figure S10. Illustration of the convergence of a representative L-RIP structure generated from the helical conformation to the loop-in conformation. The backbone RMSD from the loop-in conformation of HSP90 in the crystal structure (PDB: 1YER) is given for the starting structure with a helical conformation of α-helix3 (PDB: 2UYD), a structure generated by L-RIP that is close to the loop-in crystal structures, and the structure after a subsequent 5ns explicit solvent equilibration. The starting, L-RIP, and equilibrated L-RIP structures are displayed in the insets in Figures 4 A, B, and C, respectively. Residues 104-111 correspond to the α-helix3 flexible region that changes its conformation from the helical to loop-In structure; residues 65-75 are in the flexible N-terminus of α-helix2; residues 112-138 belong to the mobile ATP lid.

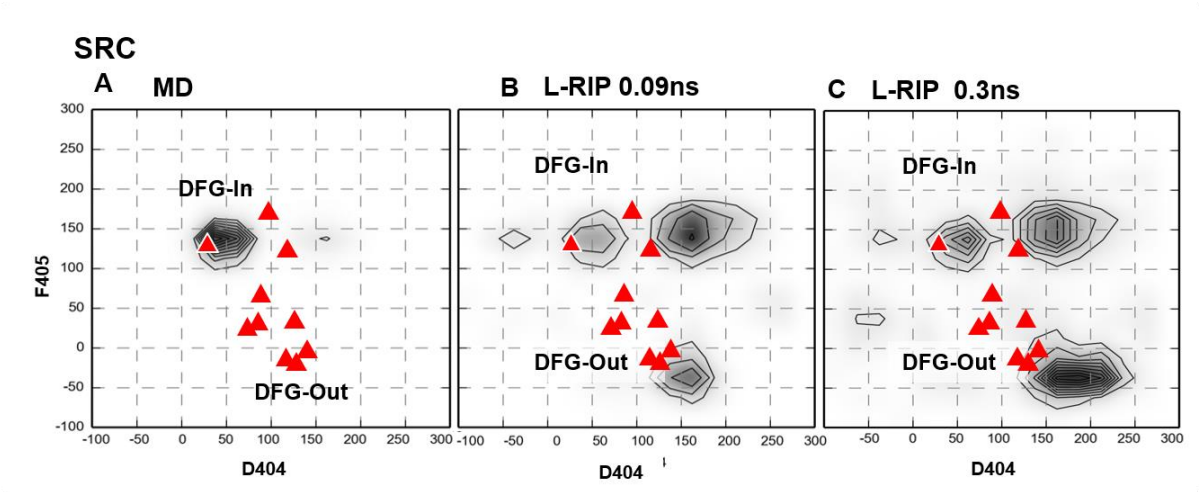

Figure S11 Distribution of backbone ψ angles of two DFG loop residues, D404 and F405, in SRC as observed in: 100ns explicit MD simulations (A); 2 L-RIP trajectories (B, C). B- in the first 300 pulses, and C- in pulses from 300 to 1000. Red triangles indicate positions of crystal structures (the PDB files used are given in Table S4).

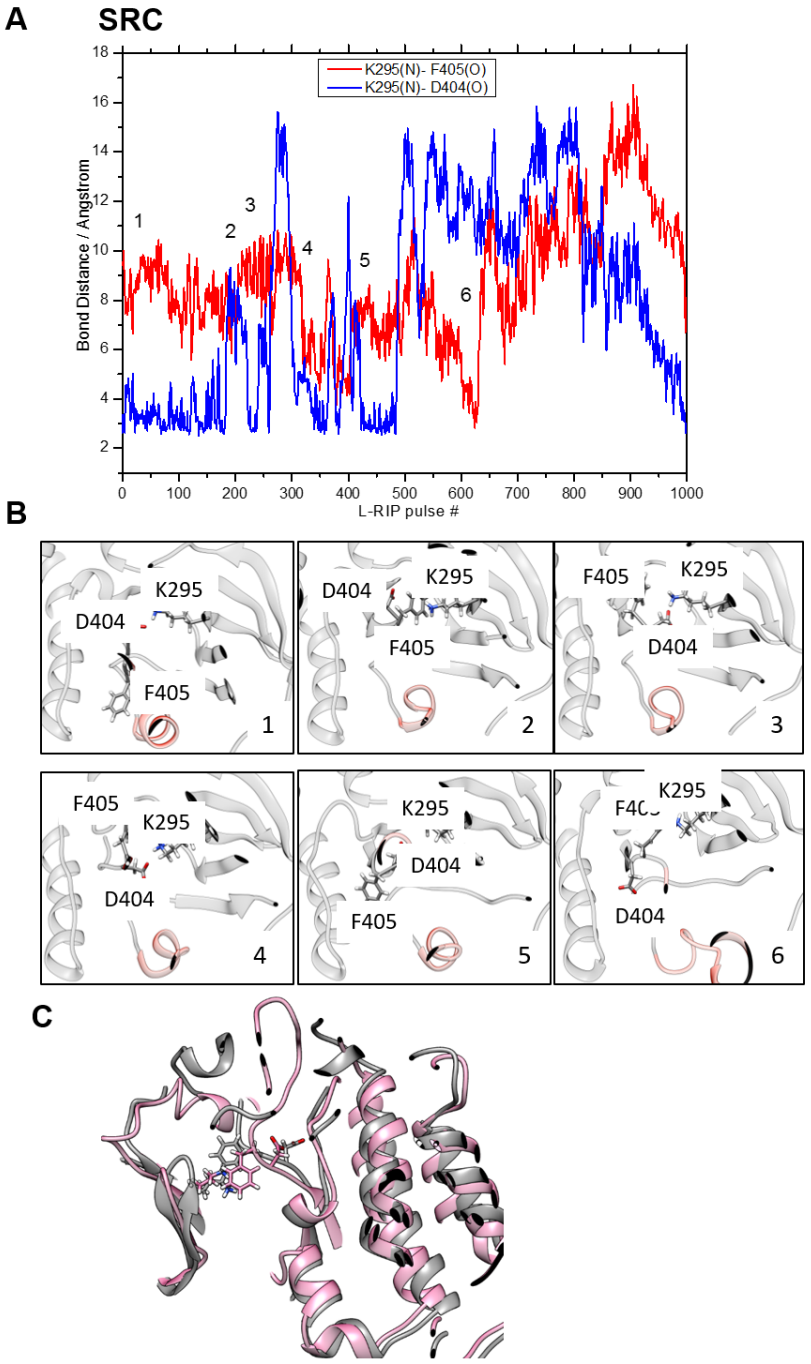

Figure S12 Evolution of the D404-K295 and F405-K295 salt-bridges along a L-RIP perturbation of SRC. A: The time dependence of the distances between the side chain N atom of K295 and the backbone O of F405 and the sidechain OD of D404 along the L-RIP trajectory shows breaking of the K295-D404 contact around pulse 300 and formation of the K295-F405 contact at 300-400 and 600-630 perturbation pulses; B: 6 snapshots at the times indicated in (A) showing the relative positions of K295, D404, and F405; C: comparison of the DFG loop conformation in snapshot 630 of the L-RIP simulation (shown in grey) and in a crystal structure (PDB:3EL8, shown in pink).

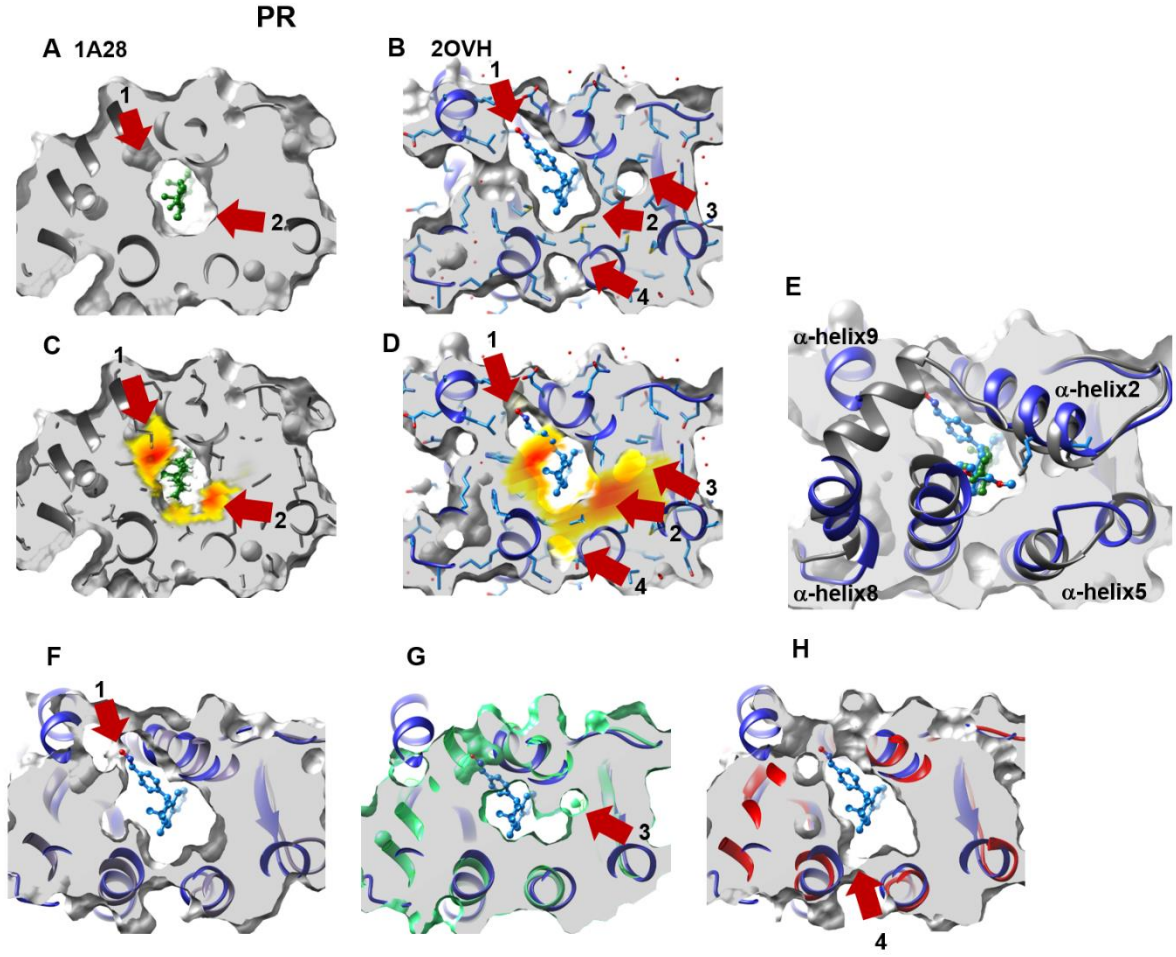

Figure S13 Opening of transient pockets in PR. Transient binding pocket regions detected are shown as the cross-section of the opening and closing transient regions (color variation from yellow to red indicates increasing number of structures in which a particular pocket is open; cross-section plane is the same as in Fig.7). Red arrows show positions of transient regions. Regions denoted as 1 and 2 are detected in crystal structures and in L-RIP simulations. Additional regions 3 and 4 are detected in L-RIP simulations. These regions appear as adjacent pockets in the structure of PDB:2OVH. A, B: binding pocket in two different structures: PDB: 1A28 (co-crystallized ligand shown in green, A) and PDB:2OVH (co-crystallized ligand shown in blue, B); C and D – transient regions detected from a set of crystal structures (C) and from L-RIP simulations (D); E –superimposed crystal structures, PDB: 1A28 (grey) and PDB:2OVH (blue); F-H: binding pocket conformations observed in L-RIP snapshots, demonstrating opening of different sub-pockets in the simulations (superimposed with the crystal structure PDB:2OVH shown in blue along with a co-crystallized ligand).

References

1. Patapati KK, Glykos NM. Three force fields' views of the 3(10) helix. Biophys J. Biophysical Society; 2011;101: 1766–71.
2. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013;29: 845–54.
3. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J Am Chem Soc. 1996;118: 11225–11236.
4. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. J Comput Chem. 1997;18: 1463–1472.
5. Hoover W. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A. 1985;31: 1695–1697.