Supporting information

Supporting information 1
Methods 1
References 2
Supporting tables 4
Supporting figures 6
Supporting videos 14

Methods

System preparation

The molecular structure for the Tyrosine-protein kinase Lck (UniprotKB P06239) SH2 domain in complex with the Ac-pY-E-E-I phosphopeptide was retrieved from the PDB 1LKK entry [1], including residues from L122 to Y226. The system was parametrized with the standard CHARMM 27 protein forcefield with CMAP corrections [2], preserving the water molecules resolved in the crystal. The ligand was capped with acetylated and amidated termini. The system was oriented such that the vector between the pocket and the peptide was pointing towards the positive *z* direction. The ligand was then displaced by 40 Å along the +*z* direction. The system was finally solvated in TIP3P water, leaving a water buffer of at least 12 A on each side (i.e., there were approximately 52 Å of water buffer from the protein on the positive *z* side). Na^+ and Cl^- ions were then added at a ionic strength of 150 mM.

In order to avoid the orientational diffusion of the protein, a *constrain set* of atoms was defined as the Cα atoms of the protein located on secondary structure elements and at least 9 Å away from the ligand (in its native pose). The system's energy was first minimized with 500 steps of conjugate gradient minimization, restraining the atoms in the constrain set with harmonic potentials of 10 kcal/mol/Å². A constant-pressure simulation was then carried on for 10 ns, at the end of which the periodic simulation box measured 60 × 66 × 98 $Å^3$; the simulation used the Berendsen barostat and long-range electrostatics were computed with the particle-mesh Ewald algorithm [3], [4]. Harmonic restrains were set to 1 kcal/mol/Å². A final 20 ns of equilibration was performed in the constant-volume (NVT) ensemble with the same constrain potential. During the equilibration, the ligand was prevented from diffusing in the bulk through the application of a harmonic potential to its Cα atoms. All dynamics were run with a time step of 4 fs thanks to the use of the hydrogen mass repartition scheme [5] implemented in the ACEMD molecular dynamics software [6]. Individual atom masses do not appear explicitly in the equilibrium distribution, therefore changing them only affects the dynamic properties of the system (marginally) but not the equilibrium distribution [5].

Production simulation were performed in the NVT ensemble at a temperature of 295 K with an analogous setup, except that simulations were conducted for 200 ns and the restrain on the position of the ligand was replaced with a flat-bottom potential (see Figure 2 in the main text); the flat-bottom restrain was null in a orthorhombic region of 40 × 40 × 60 $Å^3$ encompassing the bulk and the binding site, and amounted to 0.1 kcal/mol/Å² outside the box. All of the equilibration steps and the control simulations were carried out on a local cluster, equipped with graphical processing units (GPUs), with ACEMD [6]. Production trajectories were computed with the same software on the GPUGRID.net volunteer distributed computing network [7].

Approximate association rate constant

Assuming first-order kinetics, an order-of-magnitude estimation of the association rate constant can be obtained through the maximum likelihood criterion [8], dividing the number of reactive trajectories by the cumulative unbound time sampled, and the effective ligand concentration "seen" by the protein:

$$k_a \simeq \frac{N}{t_{\text{sampled}} \times c} = 1.6 \cdot 10^6 \ \text{M}^{-1}\text{s}^{-1}$$

where t_{sampled} = 772 trajectories × 200 ns/trajectory = 154.4 μs $\simeq$ t_{unbound}, N = 5 binding events and c = 20 mM, obtained counting the number of water molecules contained in the flat-bottom restrain box. The 95% confidence interval around the estimated value is (0.6 – 3.7) × 10^6 M^{-1} s^{-1}, obtained modelling the binding events as rare and independent [9]. The confidence interval roughly corresponds to a (symmetric) standard error of the mean rate of 0.8 M^{-1} s^{-1}, computed as the second central moment of the likelihood function.

It is worth noting that SH2-peptide association rates are related to the ligand's K_D [10], and therefore direct simulation of a low-affinity SH2 binding event is likely to require prohibitively large computational resources.

Distance, contact and RMSD computations

In Figure 3A and 3B of the main text, residue-residue distances are taken between heavy atoms and smoothed by a moving-average filter with a 5 ns window. In supporting figures S3-S8, residue-residue contact is considered to be present if and only if any pair of heavy atoms is at less than 5 Å distance. Root mean squared deviations (RMSD) were computed for the backbone atoms of the pYEEI peptide after aligning backbone atoms of the protein. Visualization and computation were performed via facilities provided by the VMD [11] and PLUMED [12] software programs.

References

[1] L. Tong, T. C. Warren, J. King, R. Betageri, J. Rose, and S. Jakes, "Crystal Structures of the Human p56 lck SH2 Domain in Complex with Two Short Phosphotyrosyl Peptides at 1.0 Å and 1.8 Å Resolution," *Journal of Molecular Biology*, vol. 256, no. 3, pp. 601–610, Mar. 1996.

[2] A. D. Mackerell Jr., M. Feig, and C. L. Brooks III, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," *Journal of Computational Chemistry*, vol. 25, no. 11, pp. 1400–1415, Aug. 2004.

[3] M. J. Harvey and G. De Fabritiis, "An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware," *Journal of Chemical Theory and Computation*, vol. 5, no. 9, pp. 2371–2377, 2009.

[4] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N log(N) method for Ewald sums in large systems," *The Journal of Chemical Physics*, vol. 98, no. 12, pp. 10089–10092, 1993.

[5] K. A. Feenstra, B. Hess, and H. J. C. Berendsen, “Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems,” *Journal of Computational Chemistry*, vol. 20, no. 8, pp. 786–798, 1999.
[6] M. J. Harvey, G. Giupponi, and G. D. Fabritiis, “ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale,” *J Chem Inf Model*, vol. 5, no. 6, pp. 1632–1639, Jun. 2009.
[7] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, “High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing,” *J Chem Inf Model*, vol. 50, no. 3, pp. 397–403, Mar. 2010.
[8] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, and D. E. Shaw, “How Does a Drug Molecule Find Its Target Binding Site?,” *J. Am. Chem. Soc.*, vol. 133, no. 24, pp. 9181–9183, Giugno 2011.
[9] K. Ulm, “A simple method to calculate the confidence interval of a standardized mortality ratio (SMR),” *Am. J. Epidemiol.*, vol. 131, no. 2, pp. 373–375, Feb. 1990.
[10] M. K. Prakash, “Insights on the Role of (Dis)order from Protein–Protein Interaction Linear Free-Energy Relationships,” *Journal of the American Chemical Society*, 2011.
[11] W. Humphrey, A. Dalke, and K. Schulten, “VMD: visual molecular dynamics.,” *J Mol Graph*, vol. 14, no. 1, pp. 33–38, Feb. 1996.
[12] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello, “PLUMED: A portable plugin for free-energy calculations with molecular dynamics,” *Computer Physics Communications*, vol. 180, no. 10, pp. 1961–1972, Oct. 2009.
[13] M. J. Eck, S. E. Shoelson, and S. C. Harrison, “Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck,” *Nature*, vol. 362, no. 6415, pp. 87–91, Mar. 1993.
[14] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, “ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids,” *Nucleic Acids Research*, vol. 38, p. W529–W533, May 2010.

Supporting tables

Table S1: Timeline of the events for the reactive trajectories. All times are given in ns from the beginning of the simulation. Values marked with *f* indicate that fluctuations still occur after the transition.

		Trajectory (ID)				
Event	**Note**	**T1 (786)**	**T2 (152)**	**T3 (170)**	**T4 (681)**	**T5 (766)**
Color code						
First contact		10	5	9	7	22
pY(+0) in place	(1)	50	–	20	22	40
E(+1) in place	(2)	110	–	50 *f*	40	65 *f*
E(+2) in place	(3)	110	160	45 *f*	18	65 *f*
I(+3) in place	(4)	110	–	75 *f*	110	65 *f*
BC loop opening (Å RMSD)	(5)	< 4 (from 160 ns)	~ 5	> 5	> 5	< 4 (until 160 ns)
BG-EF loop gap / hydrophobic pocket	(6)	Closes at 5, opens at 90	Closes at 10, opens at 60 ns	Open	Open	Open
Final RMSD	(7)	1.5	1.4	1.2	2.0	1.2

(1) reference distance: P of pY(+0) from the center of mass of atoms Cζ of Arg βB5 and Cβ of Ser βC3
(2) reference distances: Cδ of E(+1) from Cε of Lys δD3, Oω of Tyr βD5
(3) reference distance: Cδ of E(+2) from Cζ of Arg βD'1
(4) reference distance: Cδ of I(+3) from Oω of Tyr αB9
(5) RMSD (Å) of the BC loop backbone region with respect to the bound structure PDB:1LKK ("closed")
(6) definition: distance of Cβ of Ser EF1 from Cγ of Leu BG4
(7) average RMSD (Å) of the ligand's backbone between 200 and 208 ns, with respect to the native pose in the PDB:1LKK structure.

Table S2: Equivalencies of residue numbering between Tyrosine-protein kinase Lck and the general nomenclature for SH2 domains proposed by Eck et al. [13] Evolutionary conservation grades, in a scale 0 to 9, are computed with ConSurf [14]. Residues in the ligand are labelled according to their position with respect to the phosphorylated tyrosine: pY(+0) – E(+1) – E(+2) – I(+3). They are numbered from 252 to 255 in the PDB file.

	Lck	Standard name	Conservation grade		Lck	Standard name	Conservation grade
L	Leu122		6	G	Gly175	CD	4
E	Glu123		6	E	Glu176	CD	2
P	Pro124		6	V	Val177	βD	3
E	Glu125		7	V	Val178	βD	7
P	Pro126		4	K	Lys179	βD	9
W	Trp127	βA	9	H	His180	βD	9
F	Phe128	βA	6	Y	Tyr181	βD	9
F	Phe129	βA	7	K	Lys182	βD	7
K	Lys130	AA	5	I	Ile183	βD	9
N	Asn131	AA	2	R	Arg184	βD'	7
L	Leu132	AA	5	N	Asn185	βD'	4
S	Ser133	αA	4	L	Leu186	βD'	6
R	Arg134	αA	9	D	Asp187	DE	8
K	Lys135	αA	5	N	Asn188	DE	3
D	Asp136	αA	5	G	Gly189	DE	7
A	Ala137	αA	8	G	Gly190	βE	7
E	Glu138	αA	8	F	Phe191	βE	(n.s.)
R	Arg139	αA	8	Y	Tyr192	βE	6
Q	Gln140	αA	2	I	Ile193	βE	9
L	Leu141	αA	9	S	Ser194	EF	7
L	Leu142	αA	5	P	Pro195	EF	6
A	Ala143	AB	2	R	Arg196	EF	7
P	Pro144	AB	1	I	Ile197	βF	1
G	Gly145	AB	1	T	Thr198	βF	6
N	Asn146	AB	8	F	Phe199	βF	9
T	Thr147	AB	1	P	Pro200	FB	1
H	His148	AB	1	G	Gly201	αB	4
G	Gly149	AB	9	L	Leu202	αB	8
S	Ser150	βB	7	H	His203	αB	1
F	Phe151	βB	8	E	Glu204	αB	4
L	Leu152	βB	7	L	Leu205	αB	8
I	Ile153	βB	7	V	Val206	αB	8
R	Arg154	βB	9	R	Arg207	αB	1
E	Glu155	βB	6	H	His208	αB	7
S	Ser156	βB	9	Y	Tyr209	αB	9
E	Glu157	BC	8	T	Thr210	αB	1
S	Ser158	BC	7	N	Asn211	αB	1
T	Thr159	BC	4	A	Ala212	αB	1
A	Ala160	BC	3	S	Ser213	BG	3
G	Gly161	BC	7	D	Asp214	BG	7
S	Ser162	βC	4	G	Gly215	BG	9
F	Phe163	βC	3	L	Leu216	BG	9
S	Ser164	βC	8	C	Cys217	BG	7
L	Leu165	βC	8	T	Thr218	BG	4
S	Ser166	βC	9	R	Arg219	BG	1
V	Val167	βC	6	L	Leu220	BG	9
R	Arg168	βC	7	S	Ser221	BG	1
D	Asp169	βC	7	R	Arg222	βG	1
F	Phe170	CD	1	P	Pro223	βG	6
D	Asp171	CD	7	C	Cys224	βG	9
Q	Gln172	CD	1	Q	Gln225		1
N	Asn173	CD	1	T	Thr226		7
Q	Gln174	CD	1				

Supporting figures

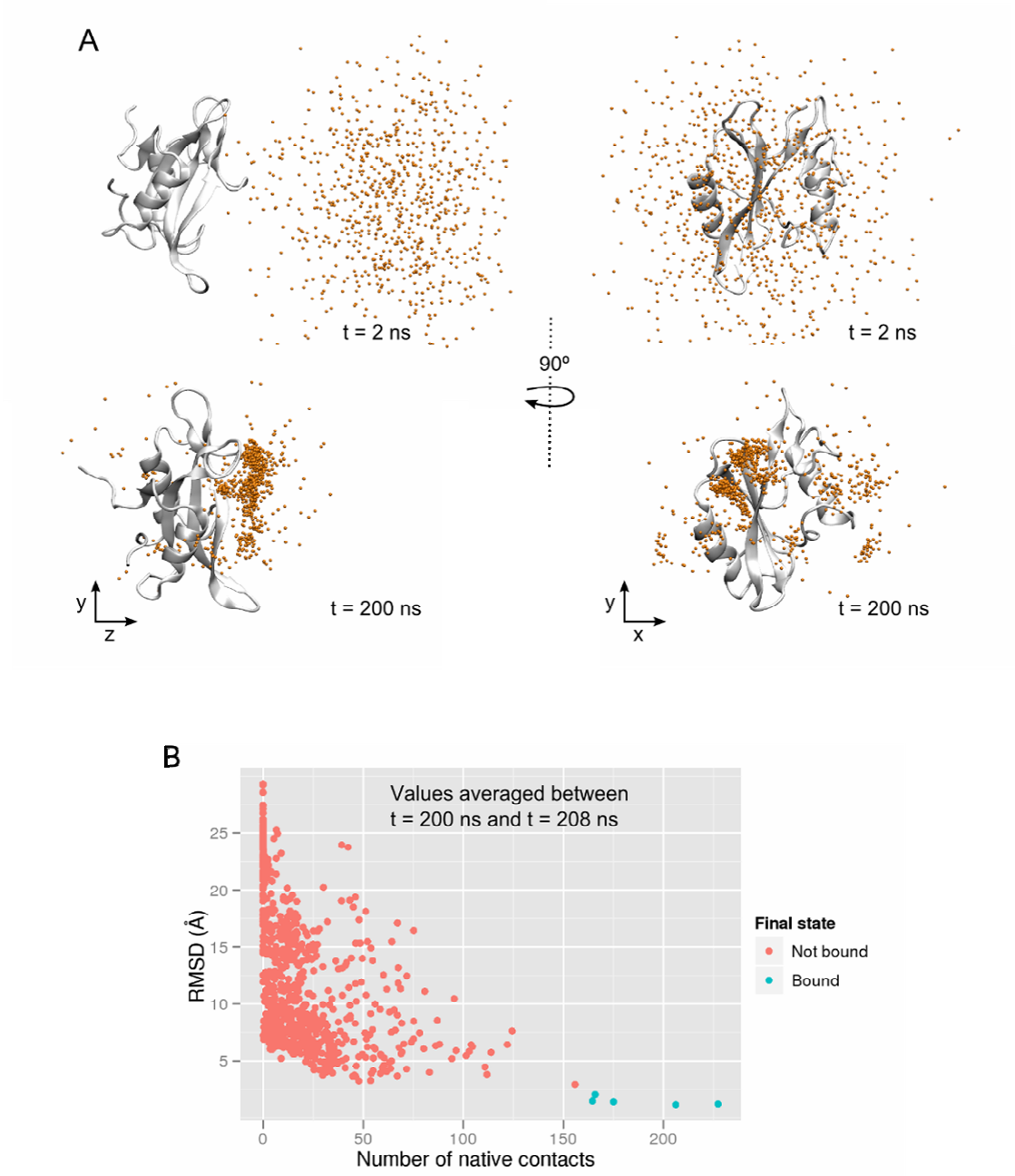


Figure S1: Initial and final states of the ensemble of trajectories. (A) Distribution of the coordinates of the ligand's pY(+0) phosphorus atom in the *yz* and *yx* planes at t = 2 ns (above) and t = 200 ns (below), for all 772 trajectories. Each dot corresponds to a single trajectory. The snapshots show that the configuration of the ligand is essentially randomized in the first few nanoseconds after the beginning of the simulation. (B) Joint distribution of the final values of RMSD and number of native contacts in the ensemble of 772 trajectories, shown as dots. The five binding trajectories that recovered the crystallographic pose (T1-T5, RMSD < 2 Å) are highlighted in blue. In the final state of most of the trajectories the ligand is in contact with the protein; however, only a small fraction of them has clearly reached a bound pose. Both RMSD and the number of native contacts are averages over the interval t = 200-208 ns.

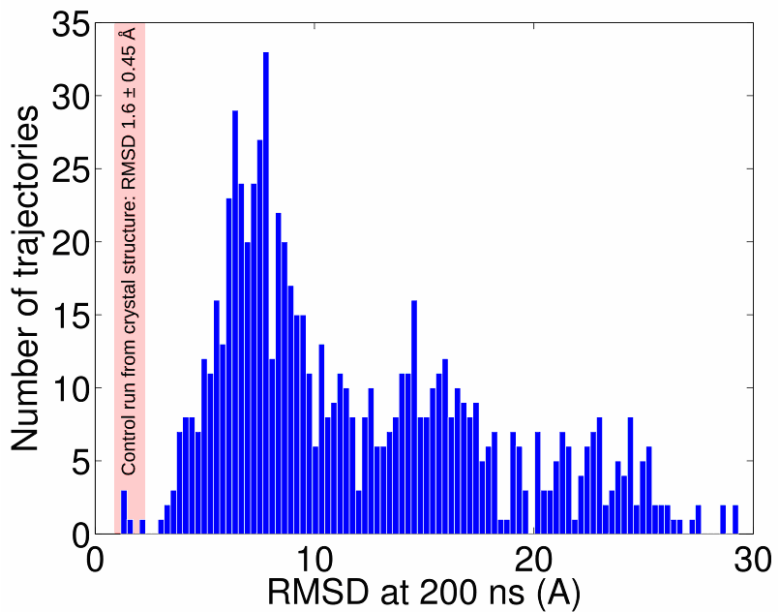


Figure S2: Distribution of the ligand RMSD (averaged between 200 and 208 ns) at the end of the 772 simulations with respect to the crystal structure. The final RMSD of 5 trajectories falls within one standard deviation from the average RMSD of a control run started from the crystal pose (1.6 ± 0.45 Å, shaded region), and are therefore considered reactive trajectories and analyzed in this work. All RMSD figures refer to the ligand's backbone atoms after aligning the proteins' Cα atoms.

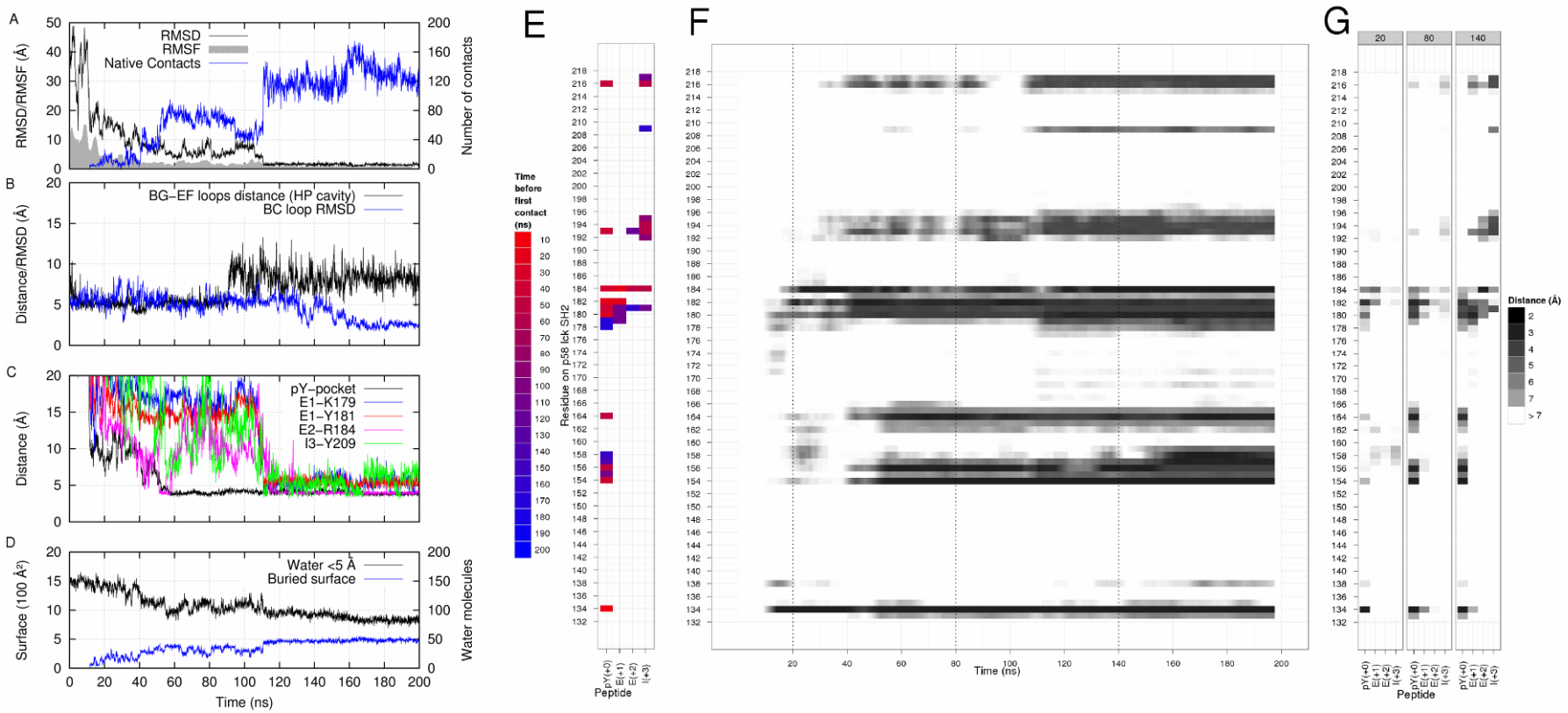


Figure S3: **Trajectory T1** observables with respect to time through the binding event. (A) RMSD and RMSF of the ligand's backbone (black and gray) and number of native contacts (blue). The ligand achieves the native pose (RMSD < 2 Å, low RMSF) at approximately 110 ns. The number of native contacts further increases around 160 ns, when the protein's BC loop locks on the ligand (B, blue). (C) Distances of the ligand's residues from significant partners in the native pose: pY(+0) and E(+2) are the first to form native contacts, 50 ns after the beginning of the simulation (black); E(+2), however, quickly loses these contacts (magenta), and re-established them at 110 ns, cooperatively with the other residues (red, green and blue). (D) Water molecules within 5 Å of the ligand (black) and buried surface (blue). Hydration decreases almost steadily during the binding process, achieving its minimum value (almost half of the bulk value) when the BC loop locks on the ligand (B, blue). The maximum value of the buried surface is reached with the binding pose at 110 ns. (E) Time to first contact for ligand-SH2 residue pairs. Each matrix element shows the time when a contact is first established between a given residue pair (red to purple). The phosphotyrosine achieves contacts with the "capture set" around 10 ns (residues R134(αA2) and K182(βD6), red); shortly afterwards, the phosphotyrosine is buried and makes contacts with the charged pocket (residues R134(αA2), R154(βB5), and S164(βC3), also red). Residues E(+1), E(+2) and I(+3) cooperatively establish their native contacts around 110 ns, achieving the final bound pose when I(+3) forms hydrophobic native contacts with the EF and BG loops (residues Y192 to P195 and L216) (purple). (F) Residue-ligand distance timeline, highlighting the evolution of the ligand's contacts with each of the protein's residues. A swift transition to the native pose is visible around 110 ns, as is BC loop (residues 157-161) locking at 160 ns. (G) Distance maps at 20, 80 and 140 ns, roughly corresponding to captured configurations (20 ns), to pY(+0) in the charged pocket (80 ns), and to the crystal pose (140 ns and beyond).

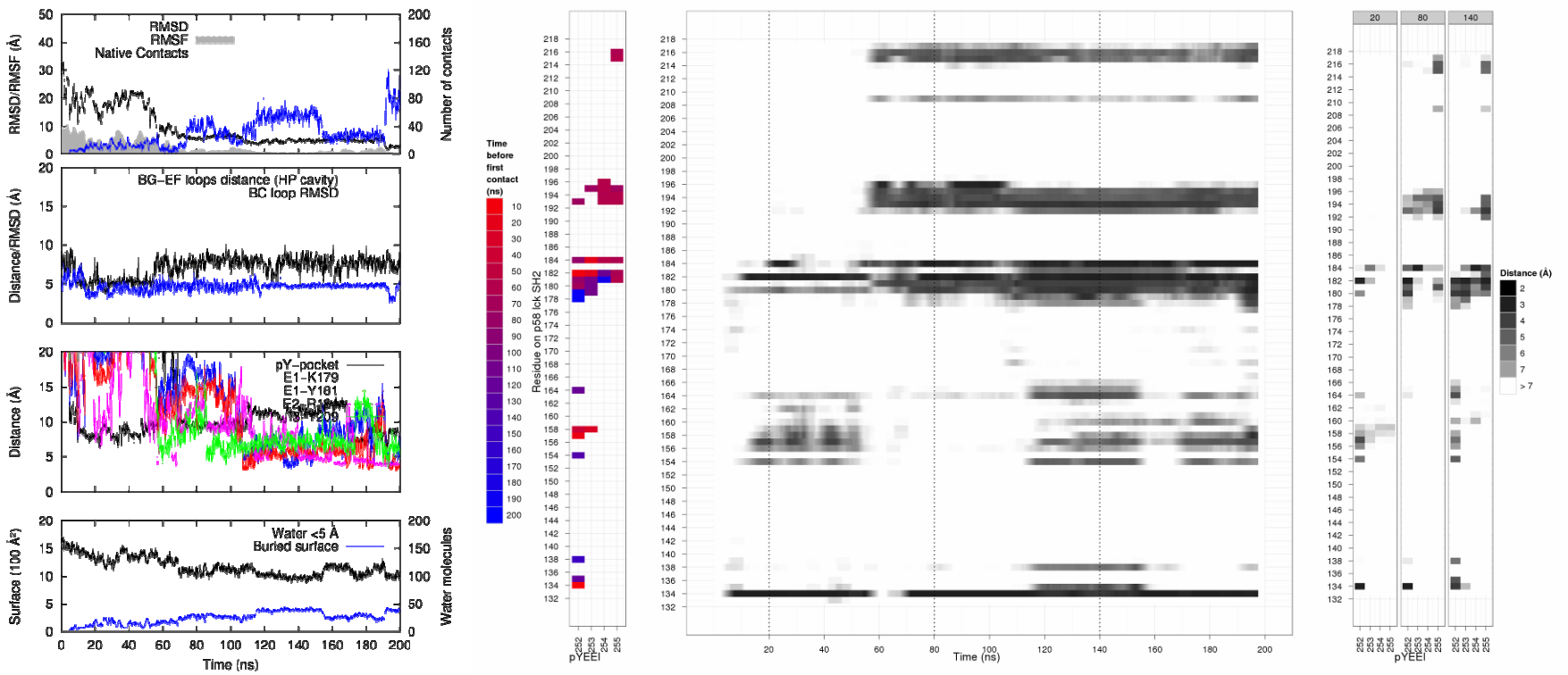


Figure S4: **Trajectory T2** reaches the bound state towards the 190 ns. Although the ligand's final RMSD (1.4 Å) indicates that the native pose is nearly reached, formation of the remaining native contacts is expected to occur beyond 200 ns. Therefore, Figure 3C in the main text displays this trajectory as not having completed the association pathway.

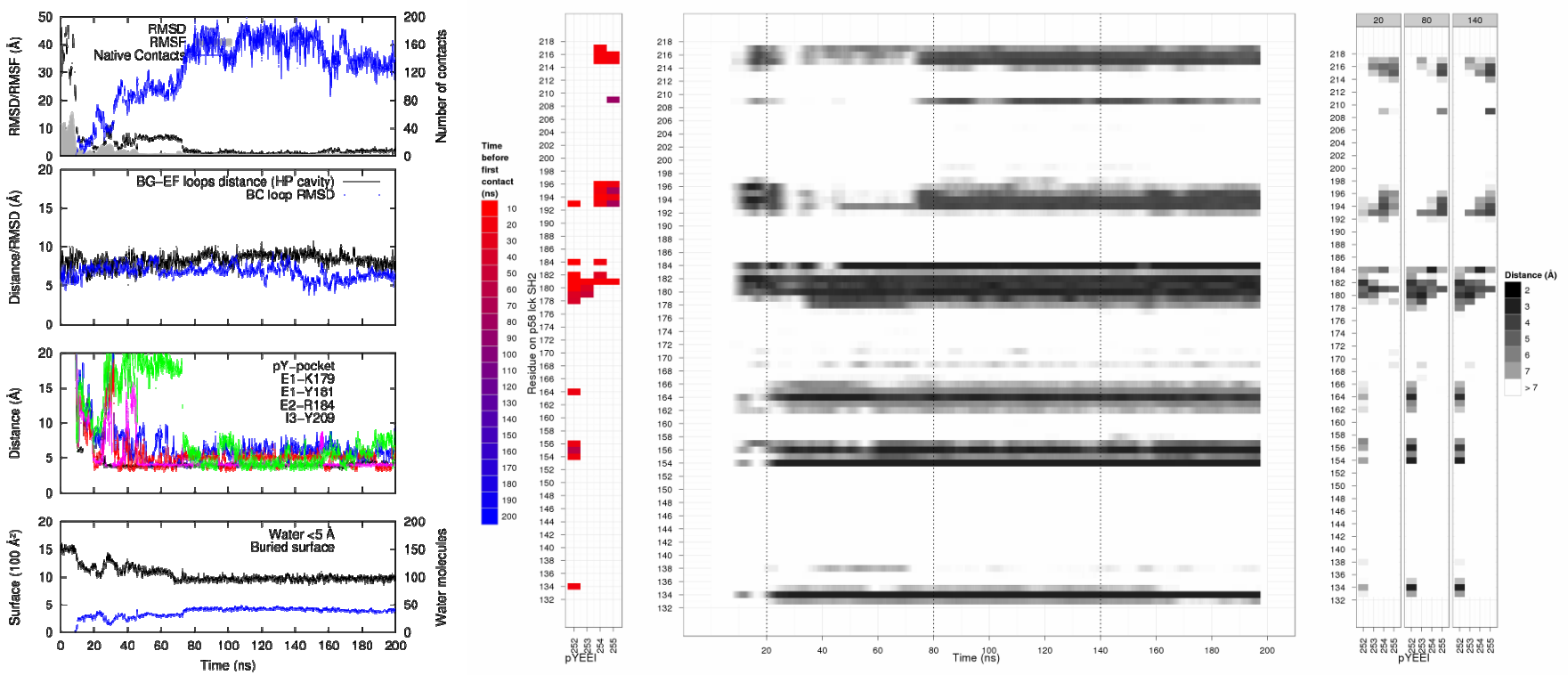


Figure S5: **Trajectory T3** reaches the native conformation at 70 ns. The pY(+0) is accommodated by the proximal pocket at 30 ns; E(+1) and E(+2) display large fluctuations and intermittent contacts until 70 ns, when I(+3) falls into the open hydrophobic pocket.

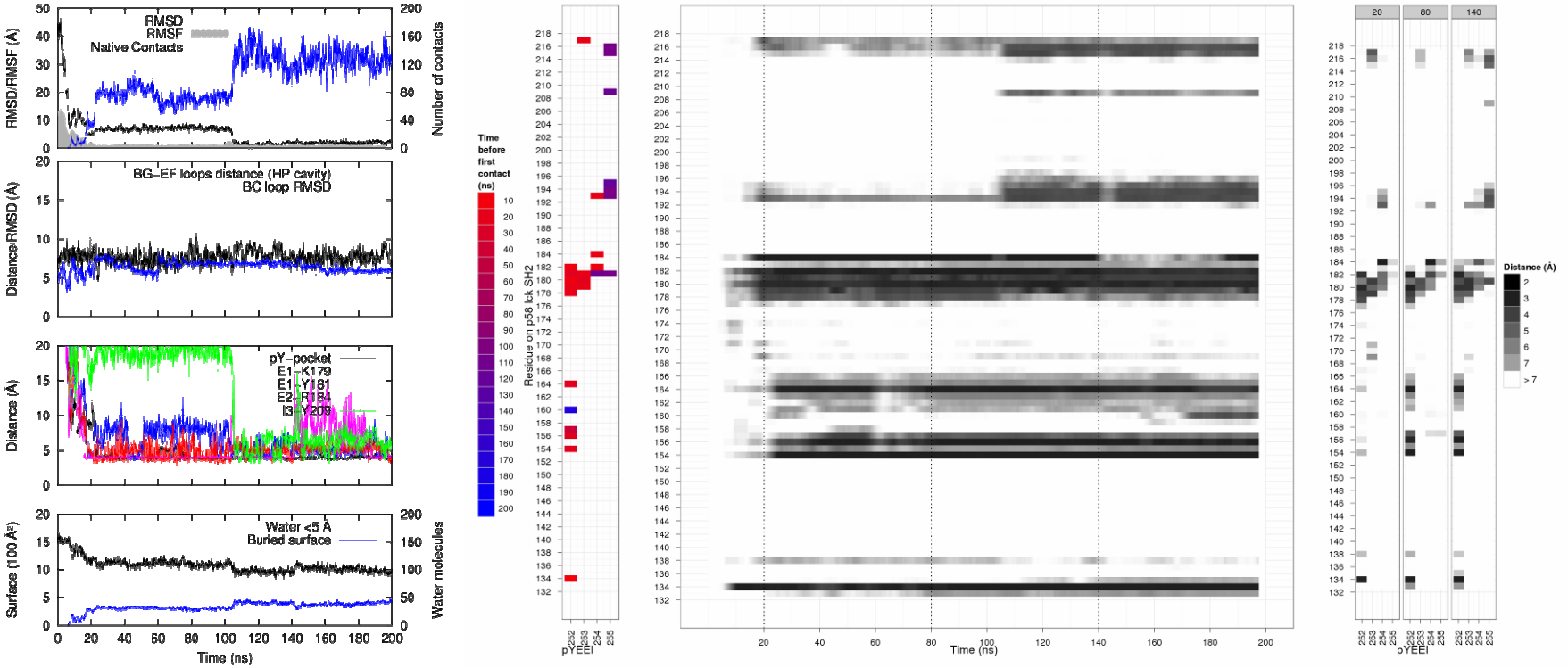


Figure S6: **Trajectory T4** reaches the bound conformation at 105 ns. The total number of native contacts defines a three-state binding process, where the capture set interacts with pY(+0) as early as 10 ns and until 25 ns; it is followed by further accommodation into the proximal pocket accompanied by binding of E(+2) to R184(βD'1). It is not until 105 ns that both E(+1) and I(+3) respectively establish their native contacts, thus leading to the bound pose.

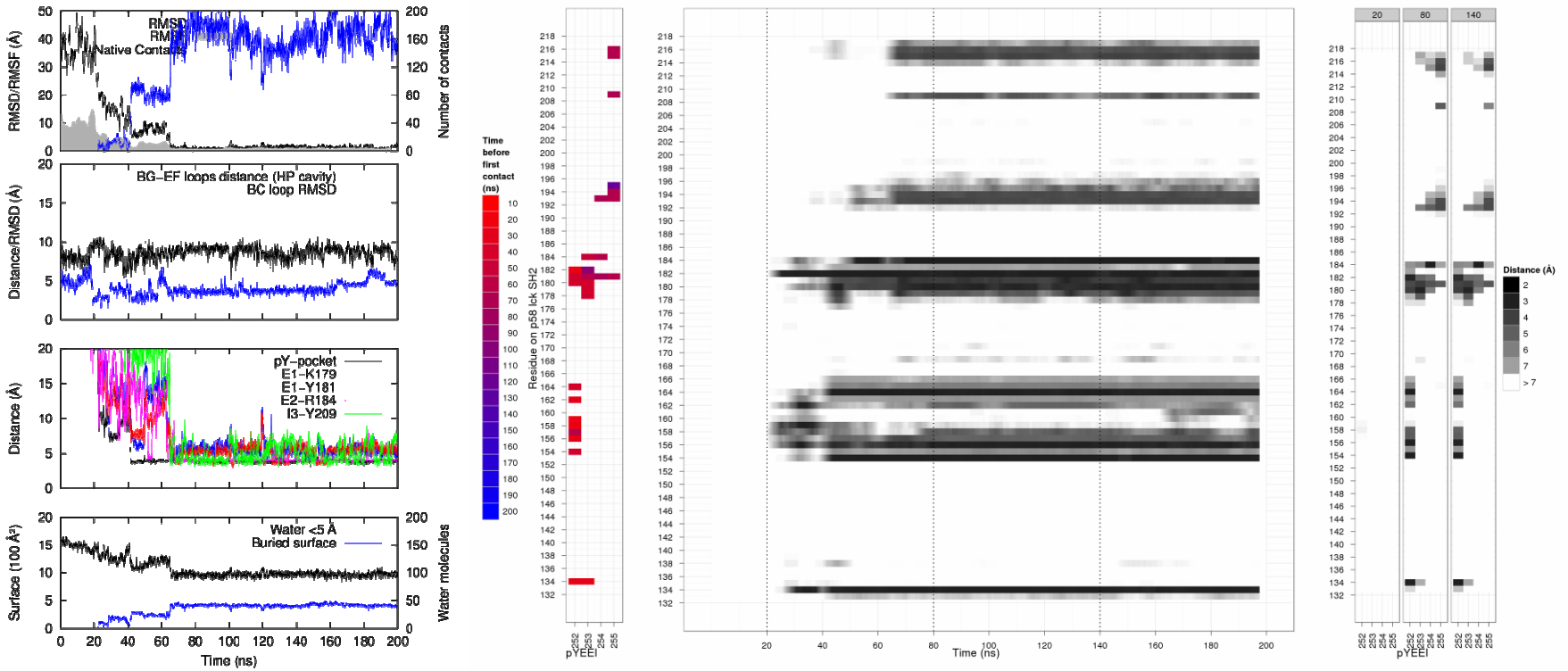


Figure S7: **Trajectory T5** is the fastest (60 ns) binding trajectory. A well defined three state process similar to T3 is also observed. pY(+0) interaction with the capture set is followed by its accommodation into the native pocket at 40 ns. At 60 ns, the three residues E(+1), E(+2) and I(+3) concurrently fall into their native contacts, and stay mostly stable for the remaining 140 ns. A transient closure of the BC loop is seen between 20 and 30 ns.

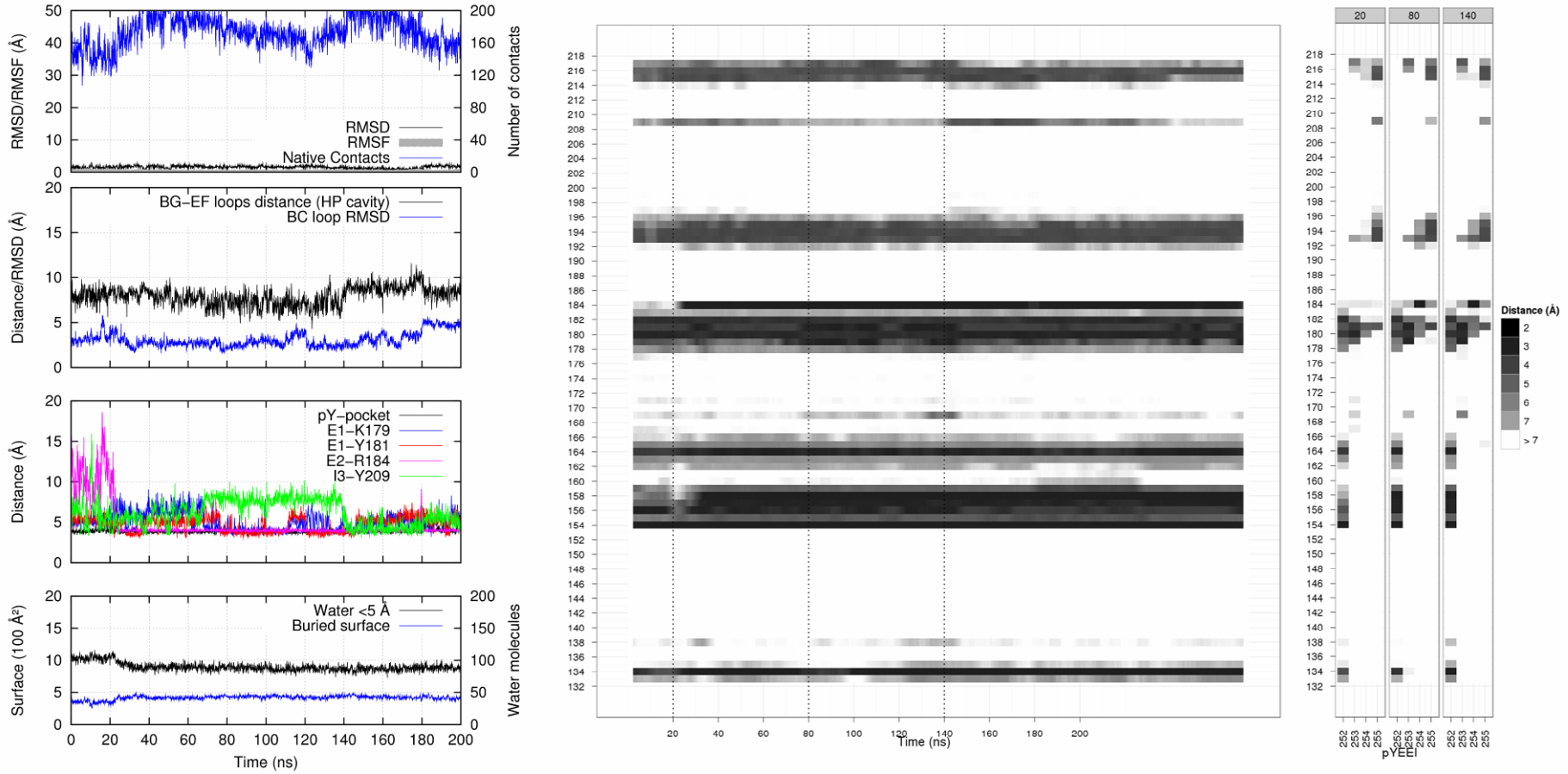
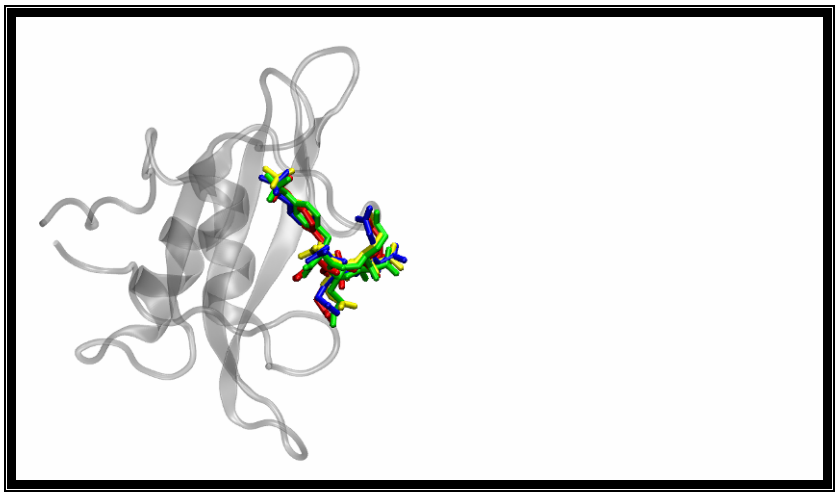
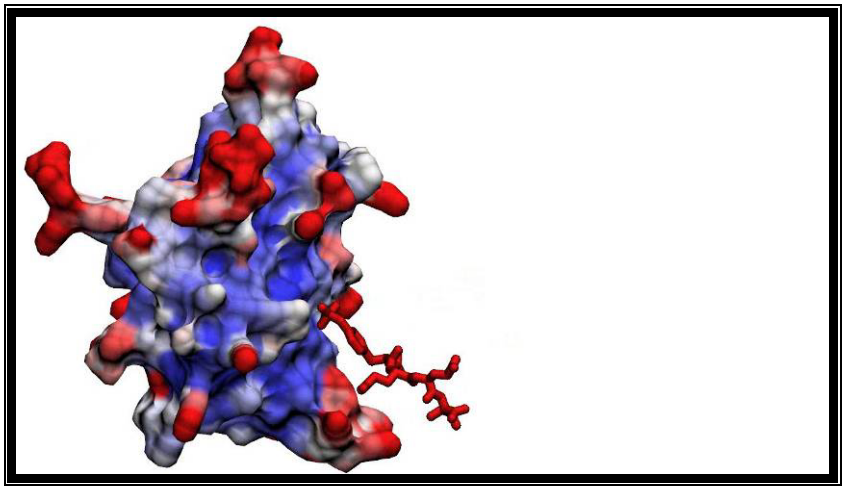


Figure S8: Observables (left) and contact timelines (right) for the **control run** of the bound complex starting from the crystal structure. Intermittent fluctuations in the E(+1), E(+2) and I(+3) native contacts distances are of the order of 3 Å. It is worth noting the stable binding of E(+2) to R184(βD'1) after 20 ns, as well as the higher fluctuation of I(+3) coupled with moderate fluctuations in the BG-EF loop distances. Partial opening of the BC-loop is also observed towards 180 ns producing a minor increase in ligand's RMSD. Average RMSD for the control run is 1.6 ± 0.45 Å. Right: contact maps between protein and ligand residues at 20, 80 and 140 ns.

Supporting videos



Video S1: All of the five binding trajectories, T1-T5, combined. Ligands are color-coded according to Table S1. The ligands' trajectories are shown together in the same videos for the sake of comparison, but they were computed in completely independent trajectories, and they do not interact with each other. One second of movie corresponds approximately to 2.0 ns of simulated time. A high definition version of this video is available online at http://goo.gl/plZDy .



Video S2: Binding trajectory T1 and local flexibility. The protein surface and the ligand's atoms are color-coded to highlight the local RMSF, on a scale from blue (lowest fluctuation) to red (highest fluctuation). Main events (see main text and Table S1) occur at: 0:05, first contact; 0:25, pY(+0) in the proximal pocket; 0:45: opening of the BG-EF distal hydrophobic; 0:55, docking; 1:20, BC loop locks the phosphopeptide. One second of movie corresponds approximately to 2.0 ns of simulated time. A high definition version of this video is available online at http://goo.gl/ZLI42 .