SUPPORTING INFORMATION

The role of conformational dynamics in the evolution of novel retro-aldolase activity

Adrian Romero-Rivera[1‡], Marc Garcia-Borràs[1,2‡]* & Sílvia Osuna[1]*

[1]Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, Carrer Maria Aurèlia Capmany 69, 17003 Girona (Spain)

[2]Department of Chemistry and Biochemistry, University of California, Los Angeles (UCLA), 607 Charles E. Young Drive, CA 90095 (USA)

[‡]These authors have equally contributed to the work

Corresponding Authors:

silvia.osuna@udg.edu, +34 664284535
marcgbq@gmail.com

COMPUTATIONAL METHODS WITH FULL REFERENCES

Molecular Dynamics Simulations. Long-timescale MD simulations in explicit water were performed using AMBER 16 package[1] in our in-house GPU cluster *Galatea*.
Schiff base parameters for the MD simulations were generated within the *antechamber* module of AMBER 16 using the general AMBER force field (GAFF),[2] with partial charges set to fit the electrostatic potential generated at the HF/6-31G(d) level by the restrained electrostatic potential (RESP) model.[3] The charges were calculated according to the Merz-Singh-Kollman scheme[4-5] using Gaussian 09.[6] Amino acid protonation states were predicted using the H++ server (http://biophysics.cs.vt.edu/H++).[7] Then, the enzyme was solvated in a pre-equilibrated truncated cuboid box with a 10-Å buffer of TIP3P[8] water molecules using the AMBER16 *leap* module, resulting in the addition of ~9,000 solvent molecules. The systems were neutralized by addition of explicit counterions (Na^+ and Cl^-). All subsequent calculations were done using the widely tested Stony Brook modification of the Amber 99 force field (ff99SB).[9] The structures used were taken from the protein data bank (PDB); RA95.0 (4A29), RA95.5 (4A2S), RA95.5-5 (4A2R) and RA95.5-8F (5AN7). RA95.5-8 structure was generated by introducing the corresponding mutations on the RA95.5-5 (4A2R) variant using RosettaBackrub software (https://kortemmeweb.ucsf.edu/backrub).[10-12] An homology model for the complete RA95.5-8F structure (4 amino acids missing in the original 5AN7 crystal structure, from 57 to 60) was generated using the SWISS-MODEL workspace (https://swissmodel.expasy.org)[13-16] and taking the RA95.0 (4A29) and RA95.5-8F (5AN7) X-ray as template structures.
A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on solute by a harmonic potential with a force constant of 500 kcal mol^{-1} $Å^{-2}$, and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems are gently heated using six 50-ps steps, incrementing the temperature 50 K each step (0–300 K) under constant volume and periodic boundary conditions. Water molecules were treated with the SHAKE algorithm such that the angle between the hydrogen atoms is kept fixed. Long-range electrostatic effects were modeled using the particle-mesh-Ewald method.[17]
An 8-Å cutoff was applied to Lennard-Jones and electrostatic interactions. Harmonic restraints of 10 kcal/mol were applied to the solute, and the Langevin equilibration scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated without restrains for 2 ns with a 2-fs timestep at a constant pressure of 1 atm and temperature of 300 K. After the systems were equilibrated in the NPT ensemble, 3 independent one microsecond MD simulations (i.e. 3 microsecond accumulated) were performed under the NVT ensemble and periodic-boundary conditions using our Galatea cluster (composed by 178 GTX1080 GPUs). With Galatea, RA95 simulations were performed at a speed of ca. 190 ns/day.

Correlation Analysis of protein motions. Correlations between the carbon alpha (C_α) of all residues of the protein variants were analyzed along the whole microsecond MD simulations using the *cpptraj* module (*matrix correl* keyword) of AMBER16:[1]

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle \langle \Delta r_j^2 \rangle}}$$

where Δr_i and Δr_j is the displacement of the C_α of the ith residue of the protein along the analyzed trajectory with respect to its position at the most populated cluster. This matrix has been called *dynamic cross-correlation matrix*. The correlation values fall in the [-1,1] range.

A correlation value of 1 indicates that the considered residues move in the same direction in most of the frames, *i.e.* they are correlated. In contrast, a value of -1 denotes that the residues move in opposite directions (anti-correlated), and values of 0 indicate that the movement of the 2 residues is uncorrelated.

Similarly, the mean distances between C_α of all residues along the MD simulations are computed with the *cpptraj* module (*matrix dist* keyword) of AMBER 16.[1] Both the correlation and distance matrices have been computed using the accumulated three microsecond MD simulation.

Input files for cpptraj module for computing the correlation and proximity matrices:

```
# We take as reference the most populated cluster from the MD trajectory
reference RA95_0_apo_3ms.c0.pdb
trajin RA95_0_apo_3_micro.nc  1 last 1

rms reference @CA
matrix dist @CA out distmat_RA95_0_apo_3_micro.dat
matrix correl name WT_corrCA @CA out corr_RA95_0_apo_3_micro.dat
```

Shortest Path Map analysis. The first step of the Shortest Path Map (SPM) calculation relies on the construction of a graph based on the computed mean distances and correlation values, in a similar fashion as done in previous studies. [18-19] For each residue of the protein a node will be created and centered on the C_α. The next step is to define edges between pairs of nodes. An edge will be drawn between those pairs of nodes whose C_α distances are at less than 5 Å during the whole simulation time. The edge distance will be derived from the computed correlation values, which define the information transfer across a given edge: $d_{ij}=-\log |C_{ij}|$. Thus, those pairs of nodes presenting larger correlation values (closer to 1 or -1) will have shorter edge distances, whereas less correlated residue pairs (values closer to 0) will have edges with long distances. The residue auto-correlation value is 1, thus the edge distance is 0 (no edge will be created). At this point, a graph with nodes and edges based on proximity and correlation is created, which is further simplified. We make use of Dijkstra algorithm as implemented in igraph module[20] to identify the shortest path lengths. The algorithm goes through all nodes of the graph and identifies which is the shortest path to go from the first until the last protein residue. The method therefore identifies which are the edges of the graph that are shorter, i.e. more correlated, and that are more frequently used for going through all residues of the protein, i.e. they are more central for the communication pathway. Pairs of residues (i,j) that are more frequently used have a higher frequency score (f_{ij}). For more details about the Dijkstra algorithm implemented in igraph, check: http://igraph.org/python/doc/igraph.Graph-class.html#shortest_paths_dijkstra. We then identify which is the edge that has the maximum frequency score (f_{max}), and compute the weight for each edge (i,j) as: f_{ij}/f_{max} (f_{ij}=frequency score of the edge between residue i and j). Only those edges that are more central are represented, and weighted according to their f_{ij}/f_{max} score. The shortest path map (SPM) is then represented on the protein structure using PyMoL.[21]

Table S1 | Mutations and kinetic properties of the different RA variants studied.

Enzyme	N° of mutations	Mutations	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M (M^{-1} s^{-1})	k_{cat}/k_{uncat}
RA95.0[22]	16	K10E, F22V, E51V, K53E, S70A, L83T, K110S, E159L, N180S, L184F, L187G, E210K, S211L, G233S, F246L, L247E	0.00005	300	0.17	4.8×10^3
RA95.5[23]	6	V51Y, E53S, T83K, M180F, R182M, D183N	0.0043	270	16	6.6×10^5
RA95.5-5[24]	6	R23H, R43S, E53T, T95M, S110N, G178S,	0.17	230	320	1.1×10^7
RA95.5-8[23]	5	S43R, F72Y, K135N, S178V, G212D	0.36	230	1600	5.5×10^7
RA95.5-8F[23]	13	T53L, R75P, N90D, N135E, S151G, V178T, F180Y, A209P, K210L, I213F, S214F, R216P, L231M	10.8 ± 0.6	320 ± 36	34000	1.7×10^9

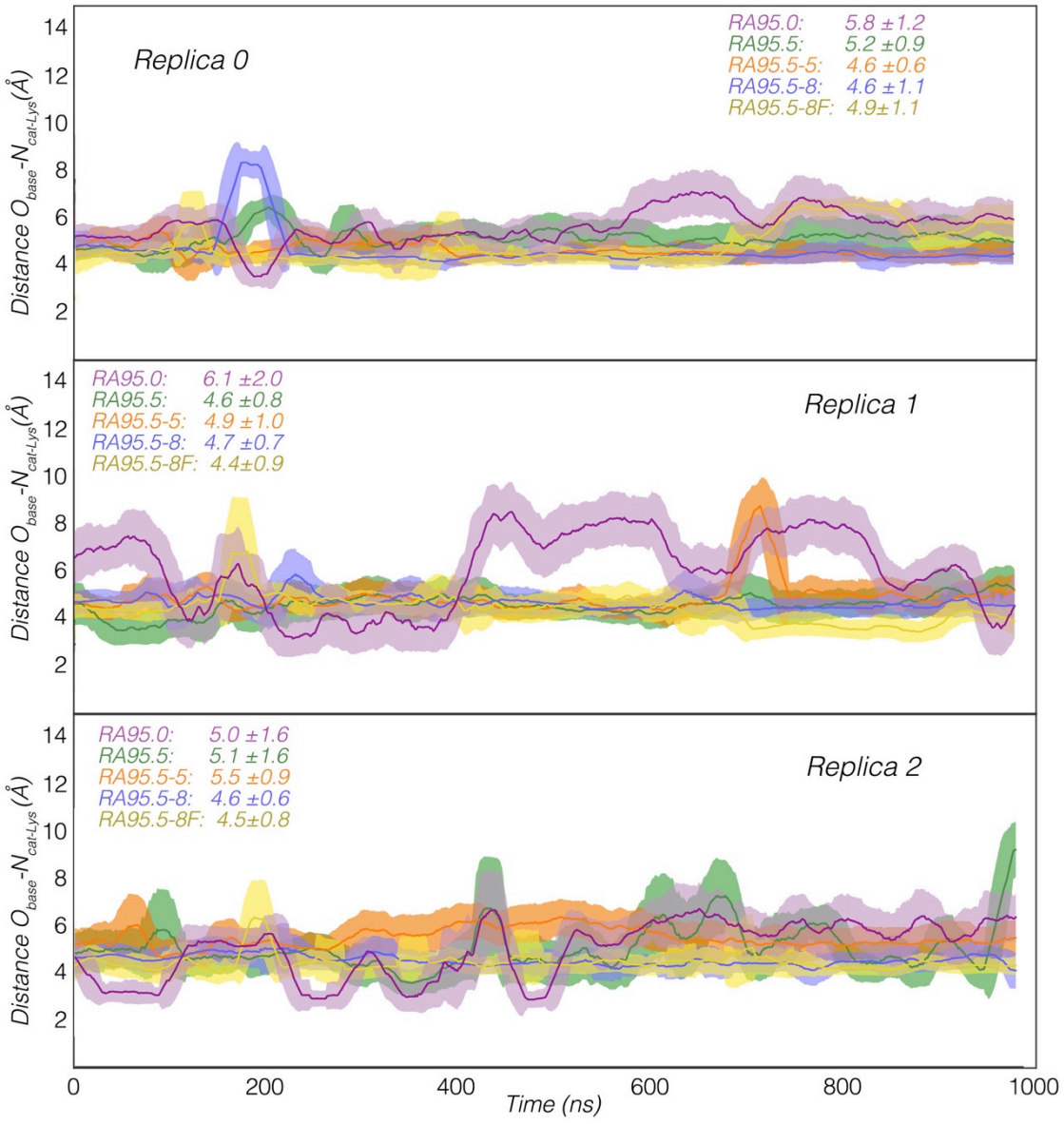

Figure S1 | Representation of the key distances of the enzyme active site residues in the *apo* state. Plot of the distance between the base and the catalytic lysine that will be involved in the Schiff base intermediate formation along the three 1 microsecond MD trajectories for RA95.0 (purple), RA95.5 (green), RA95.5-5 (orange), RA95.5-8 (blue), and RA95.5-8F (yellow). All distances are in Å.

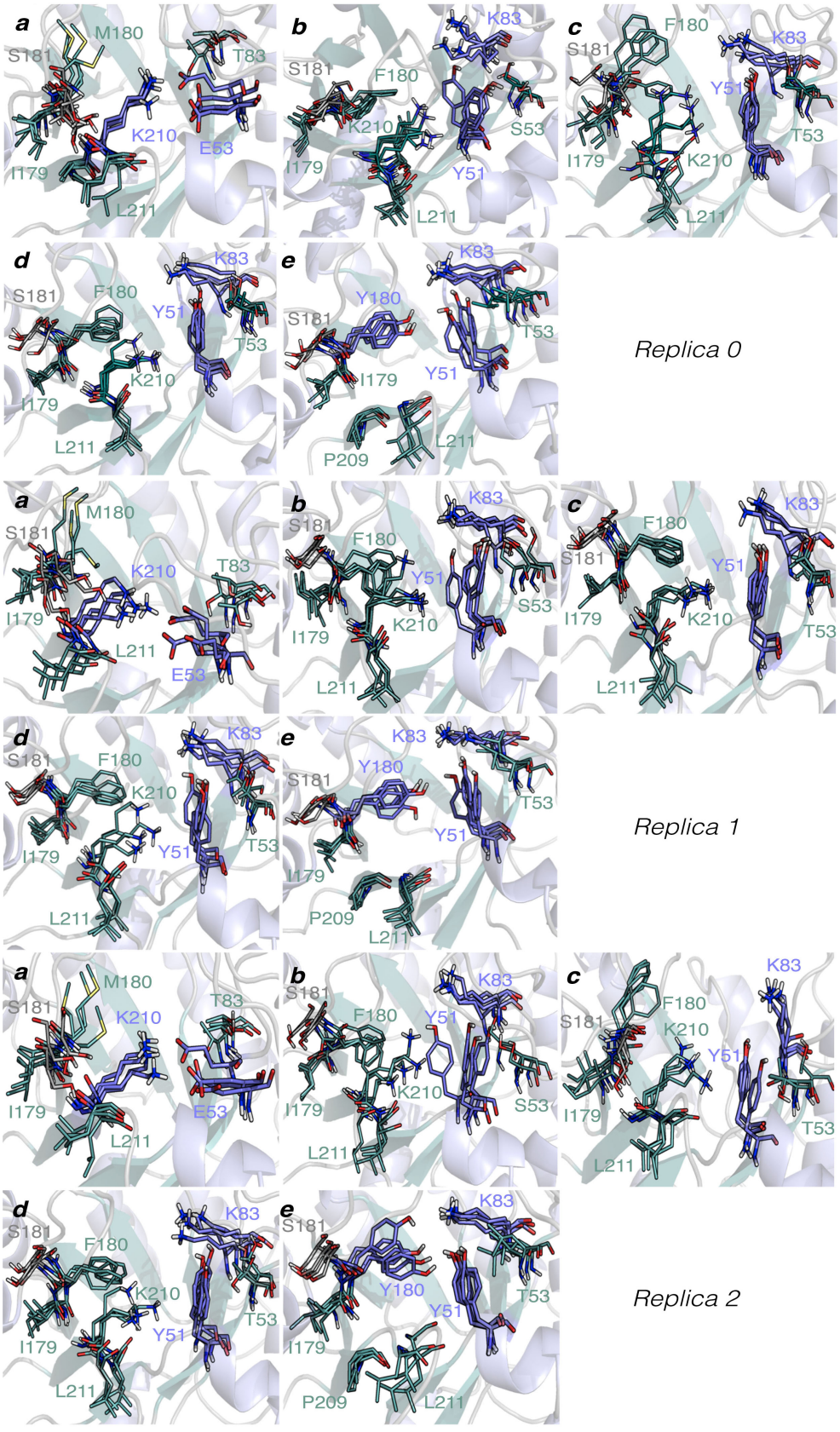

Figure S2 | Representation of the enzyme active site conformational dynamics. Overlay of representative snapshots obtained along the the three 1 microsecond MD trajectories for the *apo* states of: (**a**) RA95.0, (**b**) RA95.5, (**c**) RA95.5-5, (**d**) RA95.5-8, and (**e**) RA95.5-8F. Catalytic residues are represented in blue and sticks (for visualization purposes Asn109 has not been included).

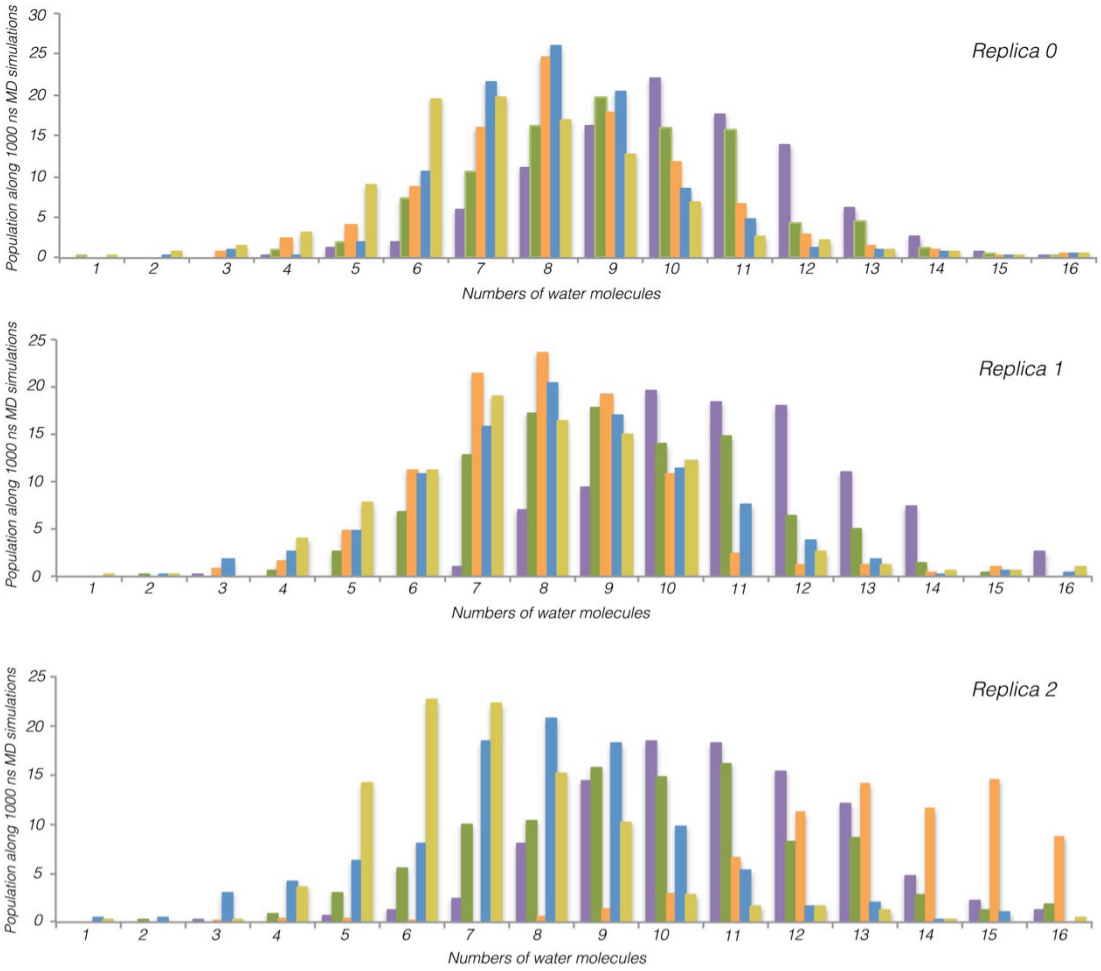

Figure S3 | Watershell estimation along the MD trajectories for the different RA variants in their *apo* state. Number of water molecules surrounding the catalytic lysine, which suggest a change of acidity from RA95.0 (purple), RA95.5 (green), RA95.5-5 (orange), RA95.5-8 (blue), to RA95.5-8F (yellow).

Table S2 | pKa estimation for the different RA variants in their *apo* state. pKa calculations using Propka3.0[25] for the catalytic lysine in the *apo* state were done using the most populated cluster obtained from the 1 microsecond MD simulations.

	RA95.0	RA95.5	RA95.5-5	RA95.5-8	RA95.5-8F
	Lys210	**Lys 83**	**Lys 83**	**Lys 83**	**Lys 83**
Replica 0	10.7	7.7	8.1	9.3	8.0
Replica 1	10.2	8.7	11.5	10.8	7.8
Replica 2	9.7	10.0	8.5	8.5	9.0

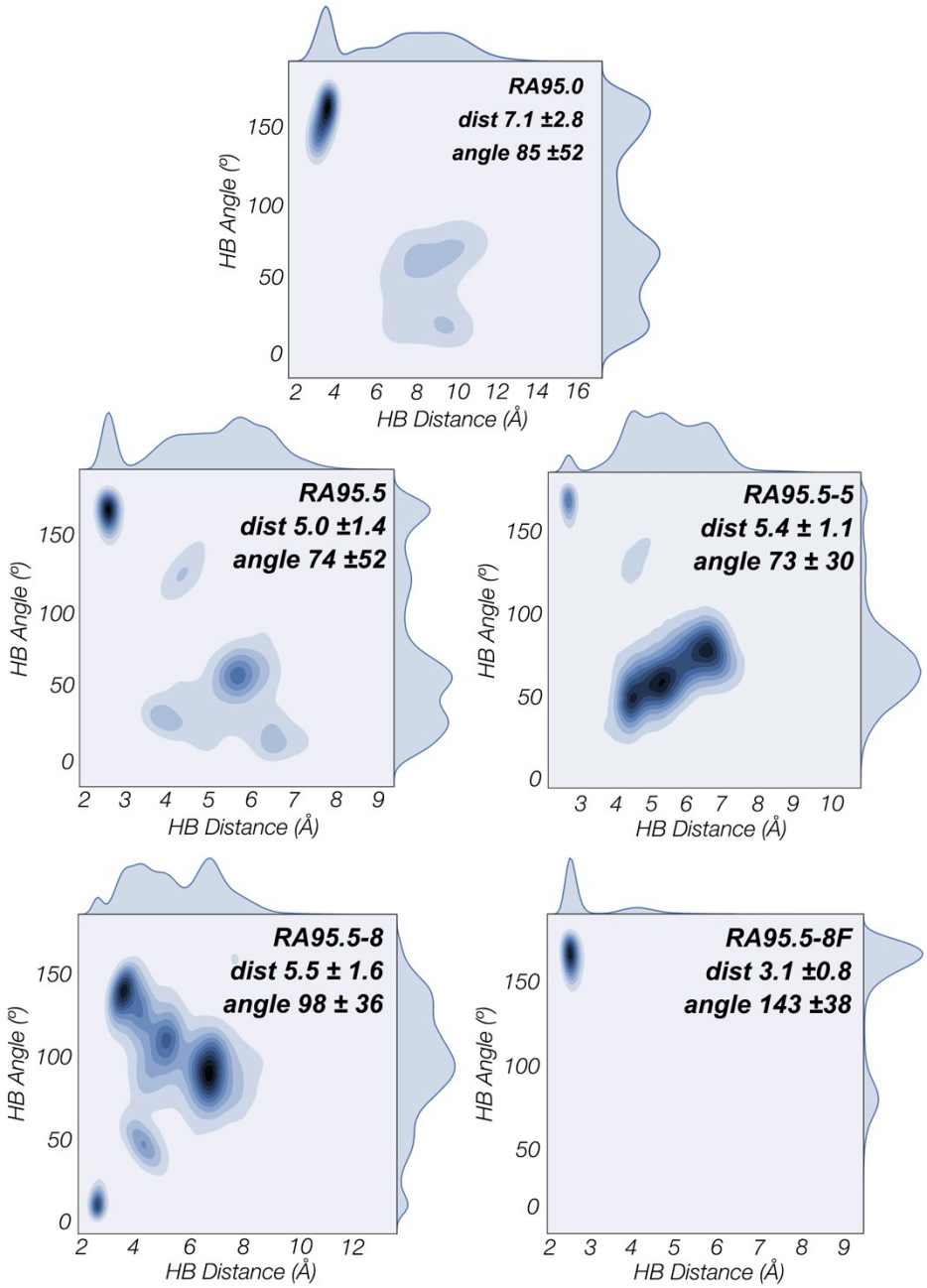

Figure S4 | Schiff base intermediate distances and angles. Representation of the distance between the deprotonated base and the oxygen of the Schiff base β-alcohol is represented (in Å) and the angle (X···H-O in º) of the hydrogen bond from the oxygen of the Schiff base β-alcohol along the MD trajectories. The representations show the different Hbond states explored for the deprotonation step of the reaction.

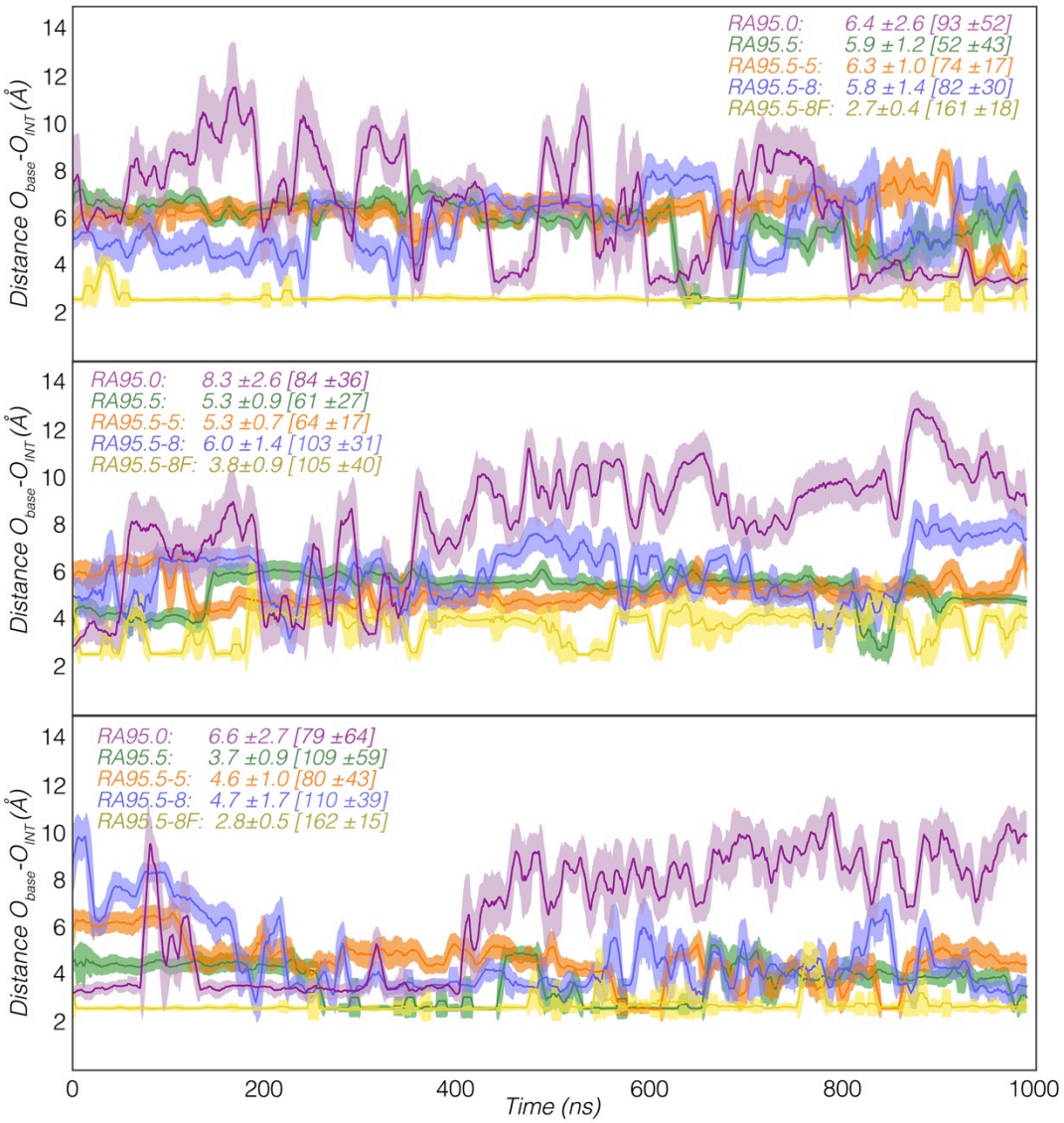

Figure S5 | Schiff base intermediate distances. Plot of the distance between the base and the β-alcohol ($X^{-}\cdots O$) that will be deprotonated and the $X^{-}\cdots H$-O angle (in º, square parenthesis) of the hydrogen from the oxygen of the Schiff base β-alcohol along the three 1 microsecond MD trajectories for RA95.0 (purple), RA95.5 (green), RA95.5-5 (orange), RA95.5-8 (blue), and RA95.5-8F with Tyr180 acting as the base (yellow). All distances are in Å.

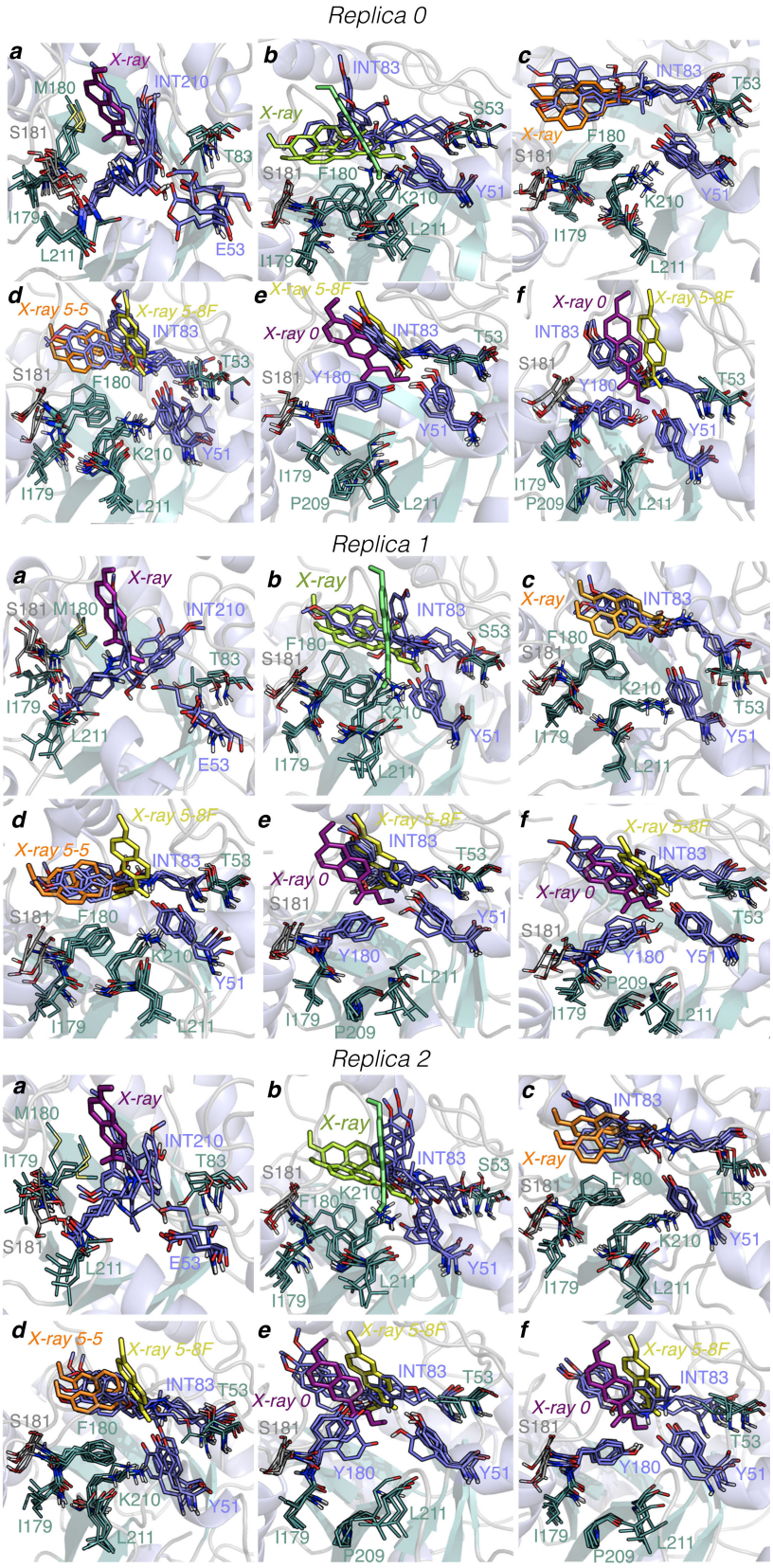

Figure S6 | Schiff base intermediate overlays of the three replicas. (**a**) RA95.0, (**b**) RA95.5, (**c**) RA95.5-5, (**d**) RA95.5-8, (**e**) RA95.5-8F with Tyr180 acting as the base, (**f**) RA95.5-8F with Tyr51 deprotonated. Catalytic residues are represented in blue and sticks (for visualization purposes Asn109 has not been included). X-ray structures with the diketone inhibitor bound are displayed for: RA95.0 (PDB: 4A29, in purple), RA95.5 (4A2S, in two types of green, lime green for the inhibitor bound to position 83, and light green for position 210), RA95.5-5 (4A2R, orange), RA95.5-8F (5AN7, in yellow).

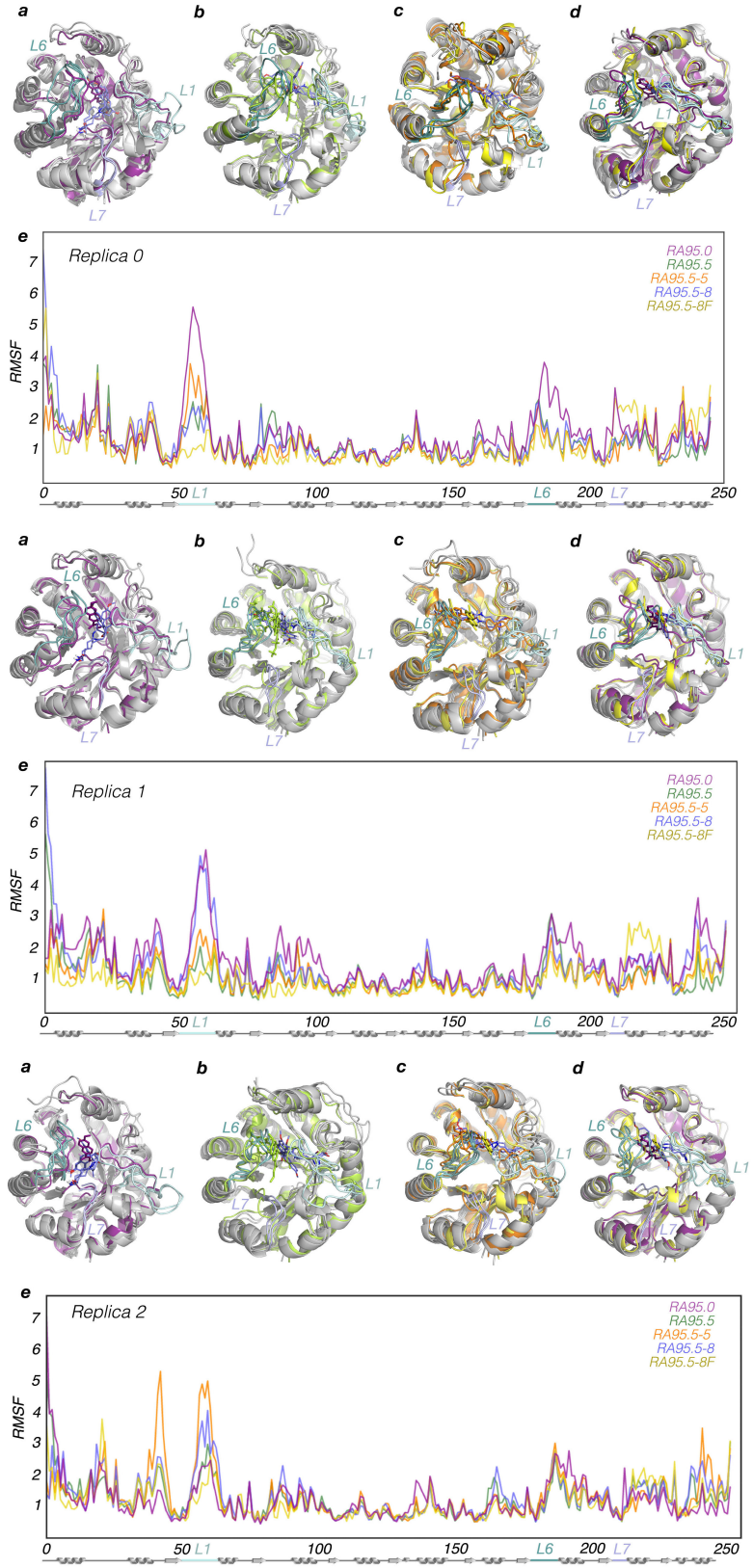

Figure S7 | RA Schiff base intermediate conformational dynamics for all replicas. Overlay of 10 conformational states sampled along the MD trajectories for the RA designs: (**a**) RA95.0, (**b**) RA95.5, (**c**) RA95.5-8, (**d**) RA95.5-8F. The X-ray structures are also displayed in purple (RA95.0), green (RA95.5), orange (RA95.5-5), and yellow (RA95.5-8F). The location of the most mobile loops L1 (residues 52-66), L6 (residues 180-190), and L7 (residues 211-215) is marked. (**e**) Root Mean Square Fluctuation (RMSF, in Å) for all RA variants along the microsecond timescale MD simulations for all replicas.

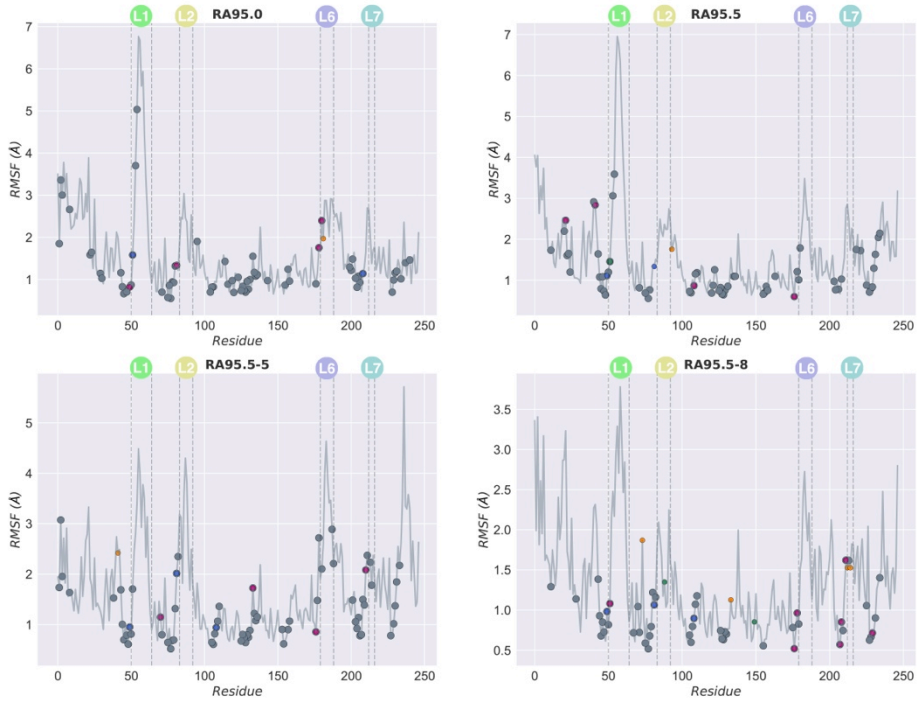

Figure S8 | Root Mean Square Fluctuation (RMSF) of all residues along the MD simulations in the apo state. The residues included in the SPM prediction have been marked using gray spheres. Those SPM positions that have been mutated in Directed Evolution (DE) experiments are colored in purple (if included in the SPM), orange (if displaced by a few residues from the path), or green (if displaced 6 positions from SPM). Most DE mutations and SPM predictions are located in regions that exhibit moderate flexibility.

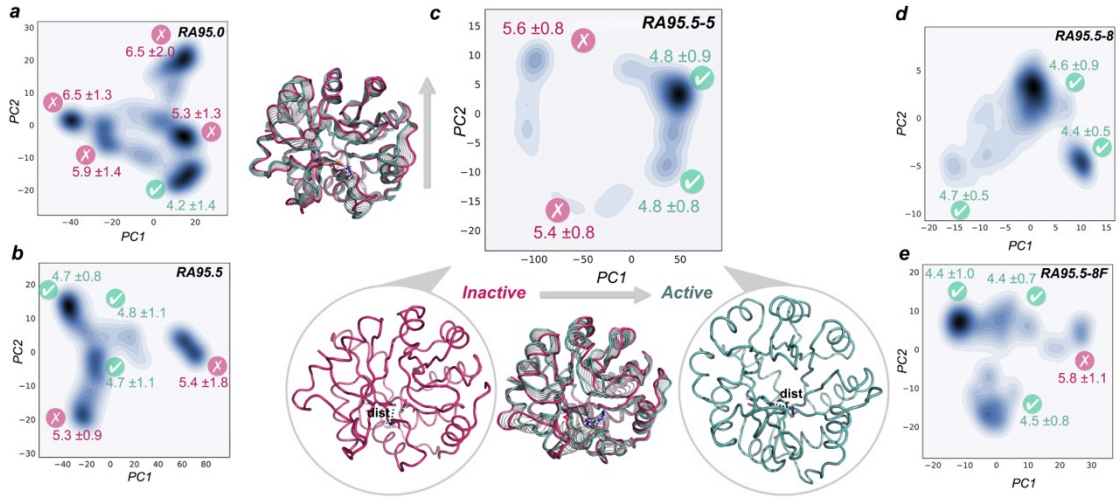

Figure S9 | Representation of the MD trajectories in the *apo* state projected into the two most important principal components (PC1, PC2) based on C_{alpha} contacts for (a) RA95.0, (b) RA95.5, (c) RA95.5-5, (d) RA95.5-8, and (e) RA95.5-8F *apo* states. For each sub-state, the distance between the heteroatom of the base and the nitrogen of the catalytic lysine is represented (in Å). Those states exploring distances in the 2.0-4.0 Å range were colored green, i.e. they are catalytically competent, represented with (✔); otherwise in red (✗). PC1 (x axis) differentiates inactive states (low PC1 values, pink structure in **b**) that present long catalytic distances from those properly oriented for the catalysis (high PC1 values, green structure in **b**). An overlay of the interpolated structures along PC1, and PC2 is also displayed for RA95.5.

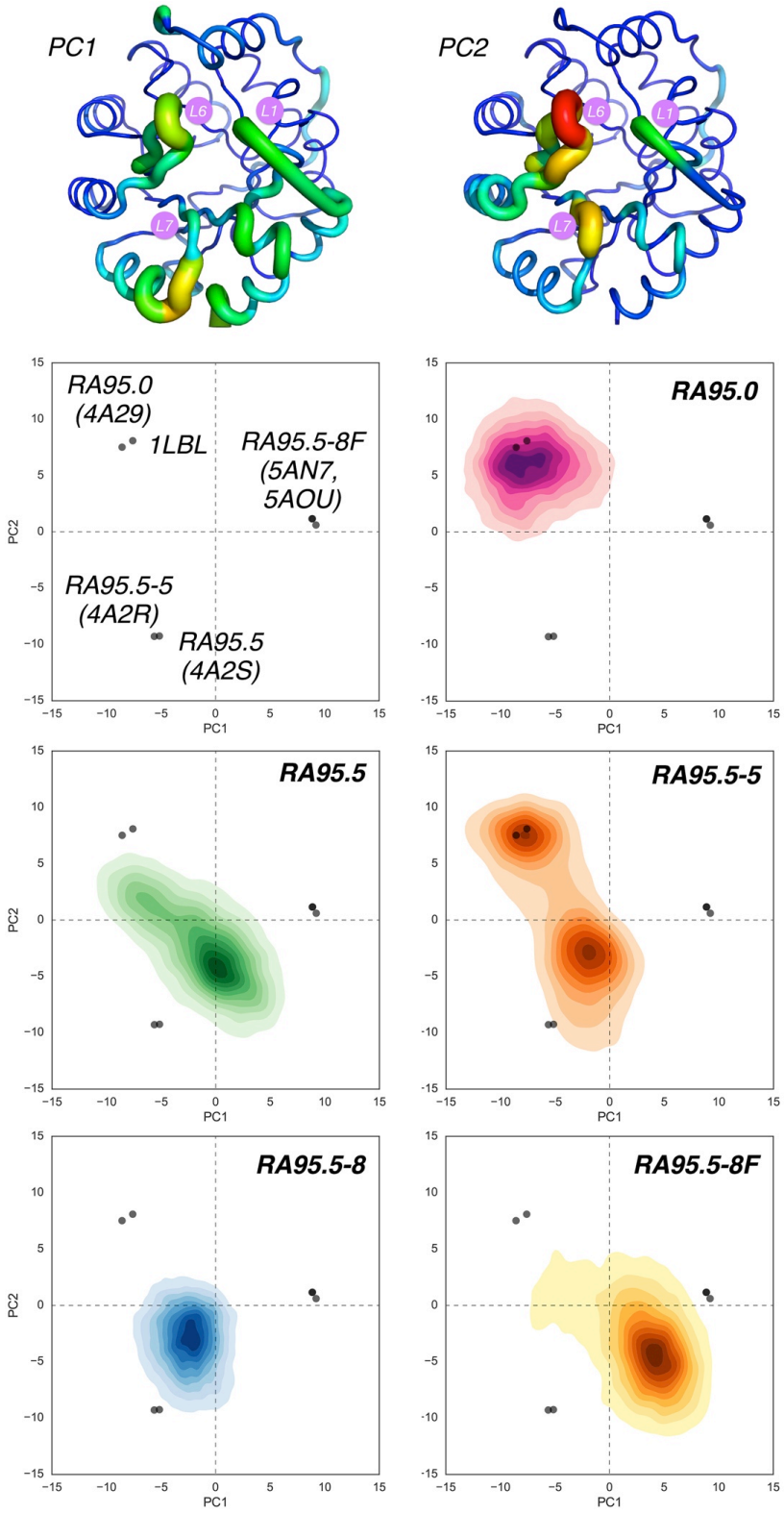

Figure S10 | Principal component analysis (PCA) based on available X-ray structures of RA variants (4A29, 4A2S, 4A2R, 5AN7, 5AOU). PCA compares the available structures and finds which regions of the protein present the major structural differences. The most flexible regions are represented in red and thicker loops, and the least flexible in blue (and thinner loops). The projection of the MD trajectories on the generated X-ray based PC space for RA95.0 (purple), RA95.5 (green), RA95.5-5 (orange), RA95.5-8 (blue), and RA95.5-8F (yellow) is also shown. In contrast to Figure 3 in main text, these projections are not able to discriminate between active and inactive conformational states.

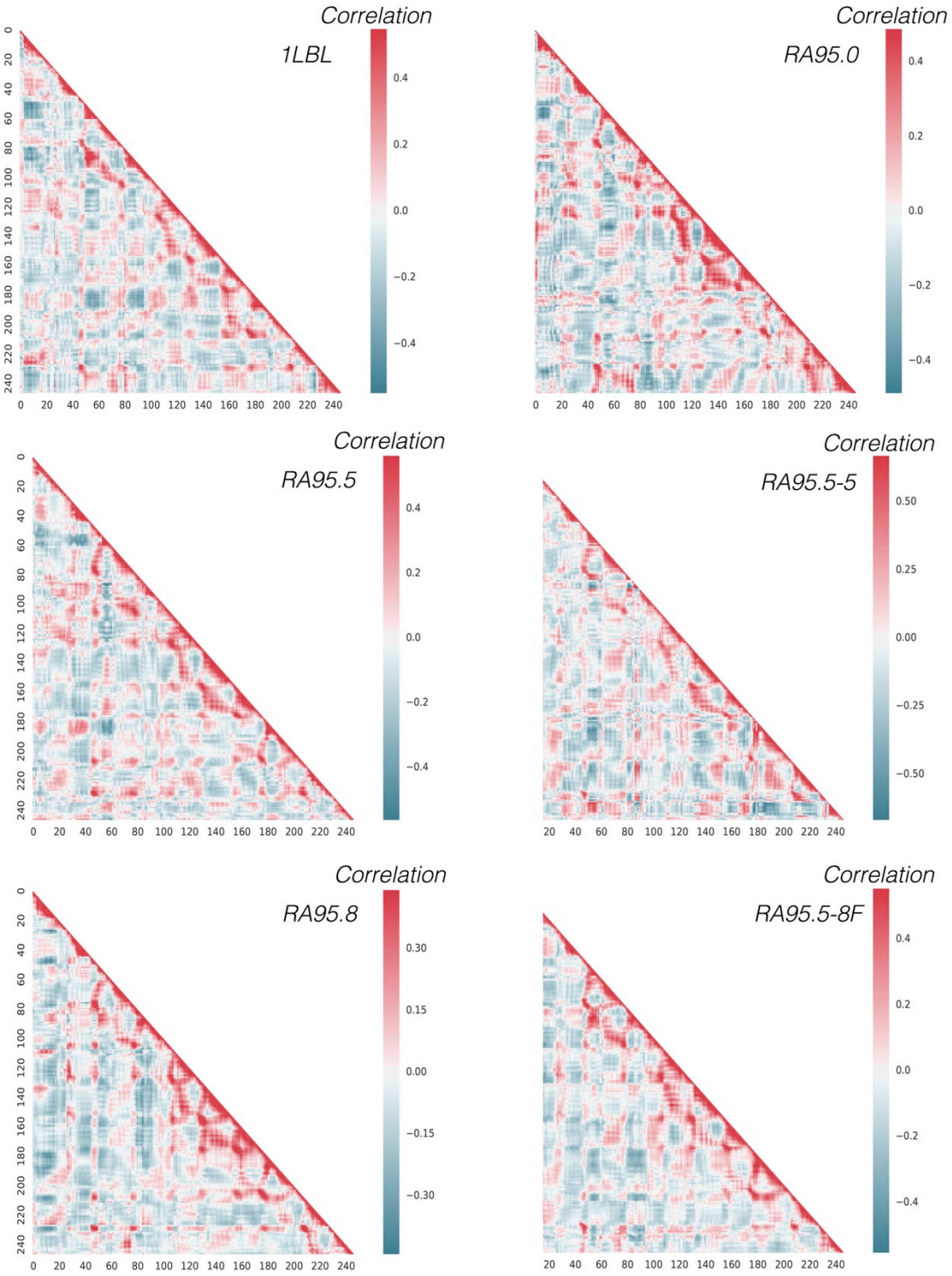

Figure S11 | Representation of the correlation matrices. Along the evolutionary pathway, the correlation matrix has been computed for all replicas, where the X- and Y-axis represent the residues of the enzyme.

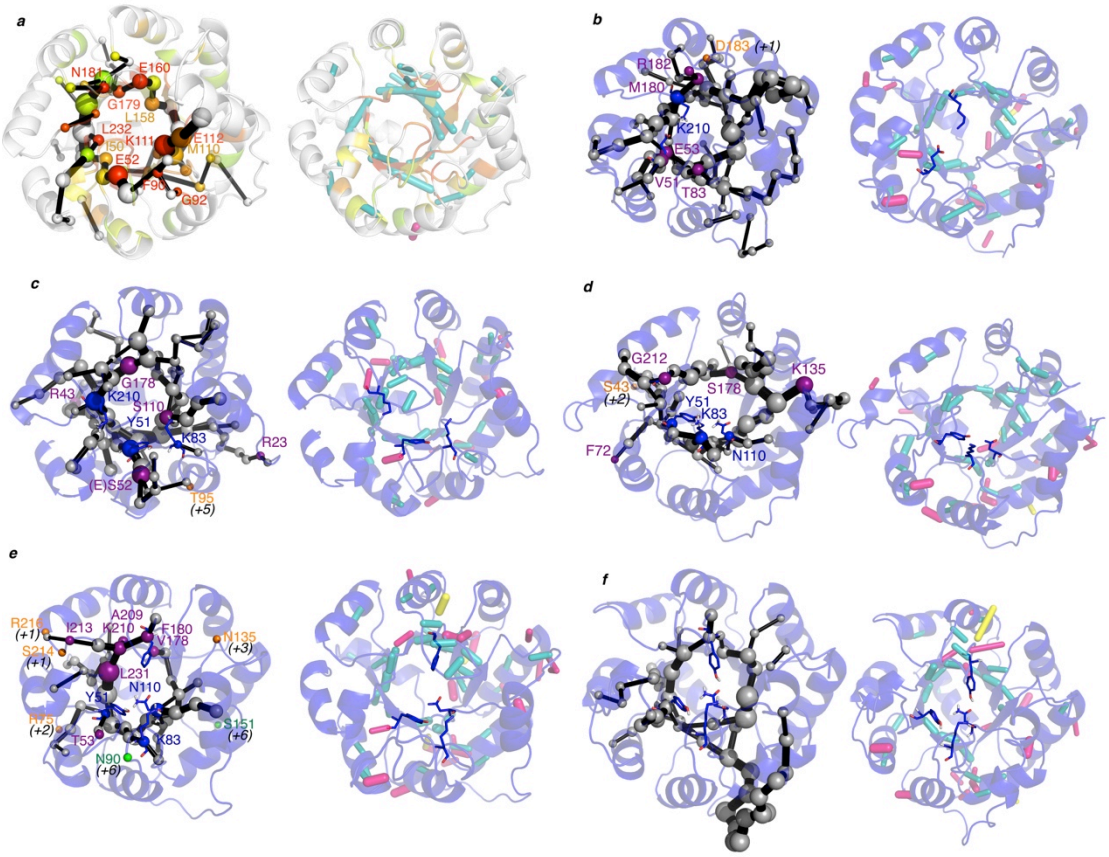

Figure S12 | Representation of the SPM and H-bond analysis network along the evolutionary pathway for: (**a**) 1LBL, (**b**) RA95.0, (**c**) RA95.5, (**d**) RA95.5-5, (**e**) RA95.5-8, and (**f**) RA95.5-8F. The size of the sphere is indicative of the importance of the position, and black edges represent the communication path, i.e. how the different residues are connected. Points mutated via DE are marked in purple (if they are included in the SPM), in orange if they are located in adjacent positions of the SPM (in parenthesis how far in sequence from the closest residue included in SPM), and in green if the mutation is located at more than 5 positions far away in sequence from the SPM. Those hydrogen bonds that have been observed at least half of the simulation time are represented in sticks: in blue those hydrogen bonds that occur between backbone atoms, in red those contacts between backbone and side-chain positions, and finally in yellow hydrogen bonds between side-chains. The weight of the stick indicates how frequent the hydrogen bond is observed.

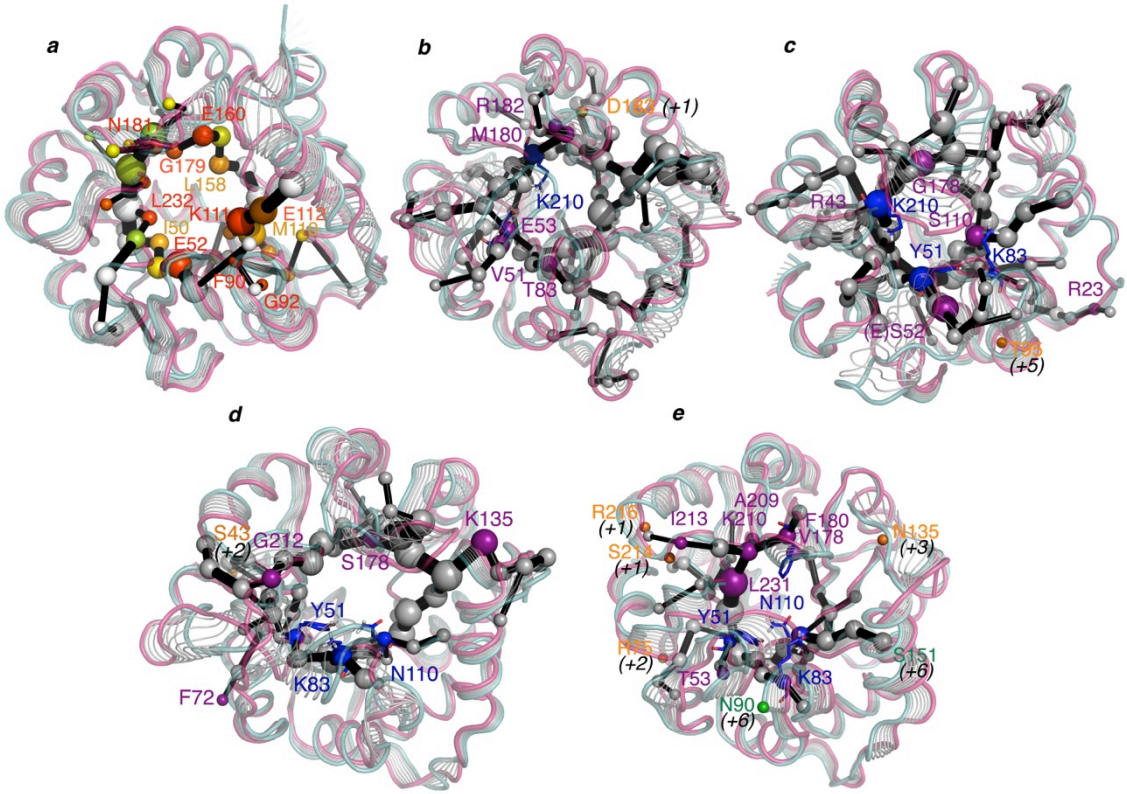

Figure S13 | Representation of the SPM and PCA along the evolutionary pathway for: (**a**) 1LBL, (**b**) RA95.0, (**c**) RA95.5, (**d**) RA95.5-5, and (**e**) RA95.5-8. The size of the sphere is indicative of the importance of the position, and black edges represent the communication path, i.e. how the different residues are connected. Thus point mutated via DE are marked in purple (if they are included in the SPM), in orange if they are located in adjacent positions of the SPM (in parenthesis how far in sequence from the closest residue included in SPM), and in green if the mutation is located at more than 5 positions far away in sequence from the SPM. An overlay of the interpolated structures along PC1 differentiates inactive states (low PC1 values, pink structure that present long catalytic distances from those properly oriented for the catalysis (high PC1 values, light blue structure).

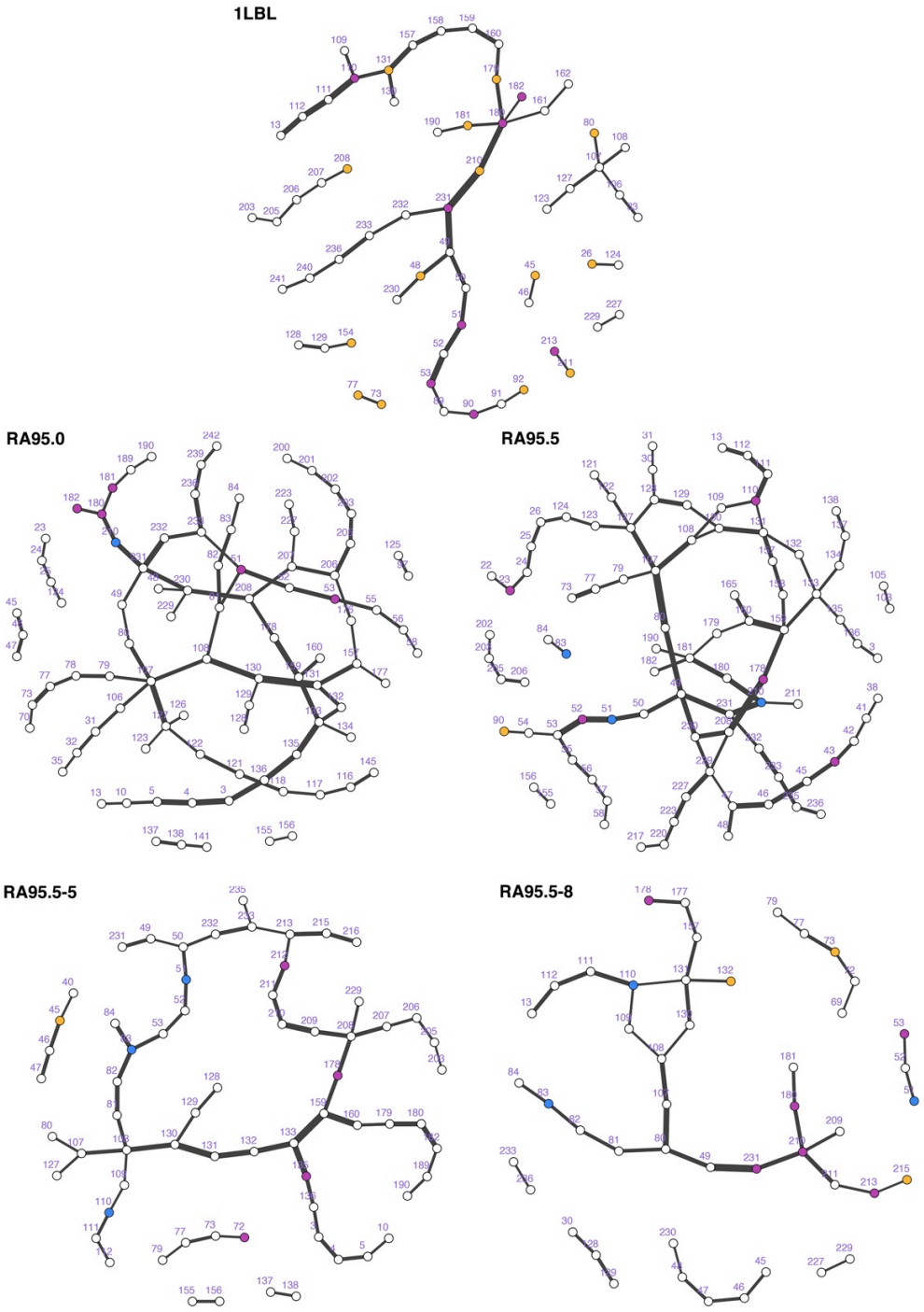

Figure S14 | Representation of the Shortest Path Map residue connections along the evolutionary pathway, generated from the 3,000 ns of MD trajectories accumulated for each enzyme in the *apo* state. Directed evolution (DE) mutations have been marked in purple (if included in the path), and orange (if displaced by less than 4 positions from the path). Interestingly, the SPM of the original scaffold 1LBL already contains all 23 DE mutations, highlighting that the highly active RA95.5-8F variant could have been generated much more efficiently had the SPM tool been applied. Catalytic residues, which are always included in the SPM, are marked in blue.

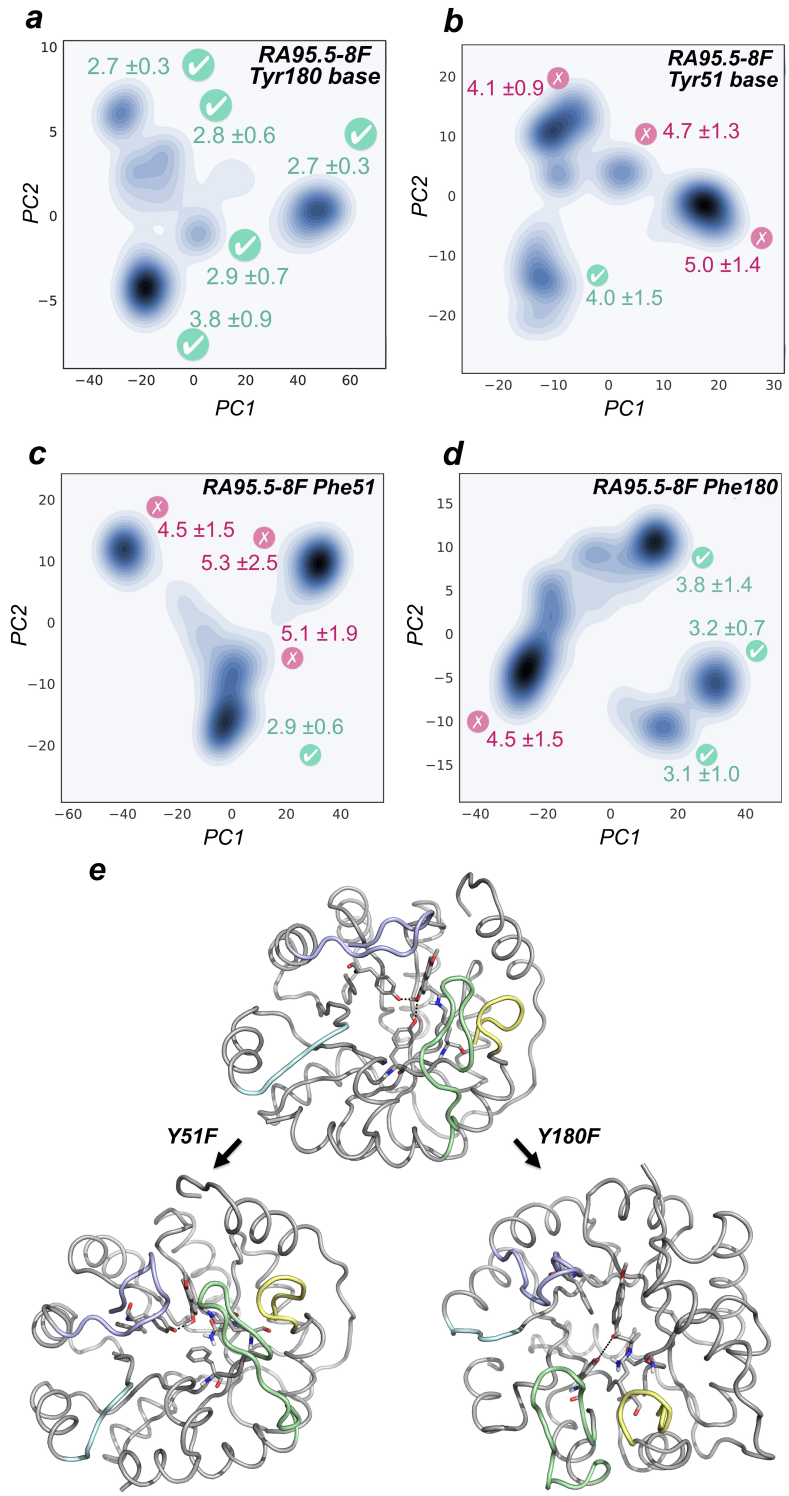

Figure S15 | Projection of the MD trajectories in the Schiff base intermediate into the two most important principal components (PC1, PC2) based on C_{alpha} contacts for: a) RA95.5-8F with Tyr180 deprotonated, b) RA95.5-8F with Tyr51 deprotonated, c) RA95.5-8F variant with position 180 mutated back to Phenylalanine (Tyr180Phe, i.e. Tyr51 is the base), d) RA95.5-8F variant with position 51 mutated to Phenylalanine (Tyr51Phe, i.e. Tyr180 is the base). For each sub-state, the mean distance between the heteroatom of the base and the oxygen of the Schiff base β-alcohol is represented together with the standard deviation (in Å). Those states exploring distances in the 2.0-4.0 Å range were colored green, i.e. they are catalytically competent, represented with (✔); otherwise in red (✗). The catalytically active conformation observed for RA95.5-8F and Y51F, Y180F variants is displayed in e).

REFERENCES

1. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A., *AMBER 16, University of California, San Francisco, 2016*.
2. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., *J. Comp. Chem.* **2004,** 25 (9), 1157-1174.
3. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., *J. Phys. Chem.* **1993,** 97 (40), 10269-10280.
4. Besler, B. H.; Merz, K. M.; Kollman, P. A., *J. Comp. Chem.* **1990,** 11 (4), 431-439.
5. Singh, U. C.; Kollman, P. A., *J. Comp. Chem.* **1984,** 5 (2), 129-145.
6. M. J. Frisch, G. W. T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, *Inc.: Wallingford, CT* **2009**.
7. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V., *Nucleic Acids Res.* **2012,** 40 (W1), W537-W541.
8. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., *J. Chem. Phys.* **1983,** 79 (2), 926-935.
9. Wang, J.; Cieplak, P.; Kollman, P. A., *J. Comp. Chem.* **2000,** 21 (12), 1049-1074.
10. Davis, I. W.; Arendall Iii, W. B.; Richardson, D. C.; Richardson, J. S., *Structure* **2006,** 14 (2), 265-274.
11. Humphris, E. L.; Kortemme, T., *Structure* **2008,** 16 (12), 1777-1788.
12. Smith, C. A.; Kortemme, T., *J. Mol. Biol.* **2008,** 380 (4), 742-756.
13. Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T., *Bioinformatics* **2005,** 22 (2), 195-201.
14. Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Cassarino, T. G.; Bertoni, M.; Bordoli, L.; Schwede, T., *Nucleic Acids Res.* **2014,** 42 (W1), W252-W258.
15. Bordoli, L.; Kiefer, F.; Arnold, K.; Benkert, P.; Battey, J.; Schwede, T., *Nat. Protocols* **2008,** 4 (1), 1-13.
16. Benkert, P.; Biasini, M.; Schwede, T., *Bioinformatics* **2010,** 27 (3), 343-350.
17. Darden, T.; York, D.; Pedersen, L., *J. Chem. Phys.* **1993,** 98 (12), 10089-10092.
18. Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z., *Proc. Nat. Acad. Sci.* **2009,** 106 (16), 6620-6625.
19. Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S., *Proc. Nat. Acad. Sci.* **2012,** 109 (22), E1428-E1436.
20. Csárdi, G.; Nepusz, T., *InterJournal* **2006,** Complex Systems, 1695-1704.
21. Schrödinger, L. *The PyMOL Molecular Graphics System, Version v1.8.2.3*, 2016.
22. Althoff, E. A.; Wang, L.; Jiang, L.; Giger, L.; Lassila, J. K.; Wang, Z.; Smith, M.; Hari, S.; Kast, P.; Herschlag, D.; Hilvert, D.; Baker, D., *Protein Sci.* **2012,** 21 (5), 717-726.
23. Obexer, R.; Godina, A.; Garrabou, X.; Mittl, P. R. E.; Baker, D.; Griffiths, A. D.; Hilvert, D., *Nat. Chem.* **2017,** 9 (1), 50-56.
24. Giger, L.; Caner, S.; Obexer, R.; Kast, P.; Baker, D.; Ban, N.; Hilvert, D., *Nat. Chem. Biol.* **2013,** 9 (8), 494-498.
25. Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A., *Nucleic Acids Res.* **2004,** 32 (suppl_2), W665-W667.