Supporting Information

for "CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations" by Giulia Palermo, Yinglong Miao, Ross C. Walker, Martin Jinek and J. Andrew McCammon.

I. Methods

I.1. Structural Models

MD simulations have been based the crystallographic coordinates of the *Streptococcus pyogenes* apo Cas9 (4CMQ.pdb),(1) Cas9:RNA (4ZT0.pdb),(2) Cas9:DNA (4UN3.pdb)(3) and Cas9:pre-cat (5F9R.pdb),(4) solved at 3.09, 2.50, 2.58 and 3.40 Å resolution, respectively (starting structures are shown in **Fig. 1** of the main text and **Fig. S1**). In the Cas9:pre-cat and Cas9:cat systems, the catalytic Mg ions have been included in the RuvC and HNH active sites, as done by Jinek et al.(1) and further reported in our previous study.(5) Missing residues of the 4UN3 and 4CMQ X-ray structures have been added via homology modeling, using SwissModel.(6) Specifically, missing residues of the 4UN3 X-ray structure include: K3, G385, A711-D718, Q766-K775, Y1013-I1029, T1051-R1058, Y1242-N1252, L1365-D1368, while missing residues of the 4CMQ X-ray structure are: K3, Q103-Q114, V308-T313, I448-L502, K528-F539, Y568-R586, Q674-N690, G715-G717, R765-Q774, G792-Q798, S860-D861, A903-G907, Q1028-K1035, G1103-S1136, A1147-K1158, T1187-K1191, E1243-F1258 and Q1364-D1368. The obtained model systems have been embedded in in explicit waters, leading to periodic simulation cells of 107*158*138 $Å^3$ (apo Cas9, for a total of ~220K atoms), ~148*107*140 $Å^3$ (Cas9:RNA, ~210K atoms), ~144*108*146 $Å^3$ (Cas9:DNA, ~216K atoms), and of ~180*116*139 $Å^3$ (Cas9:pre-cat, ~270K atoms). A further model system, constituting a putative catalytic state (Cas9:cat), has been built by engineering the HNH domain based on the homologous structure of the catalytically active T4 endonuclease VII,(7) as done by Sternberg.(8) The T4 endonuclease

VII is a prototypical homing endonuclease enzyme, which is characterized by the presence of a catalytic HNH motif. The domain is constituted by a two antiparallel β-strands flanked by a α-helix, and including a metal binding (**Fig. S2**). As such, the HNH domain of the T4 endonuclease VII constitutes a valuable template for the modeling of the HNH domain of Cas9. In detail, the HNH domain has been docked at the target DNA (*t*-DNA) cleavage site by aligning the scissile phosphate and flanking nucleotides of a DNA in complex with T4 endonuclease VII (2QNC.pdb)(7) with the scissile phosphate and flanking nucleotides of Cas9:pre-cat (**Fig. S2**). Subsequently, the HNH domain of Cas9:pre-cat has been aligned with the T4 endonuclease VI, thus leading to a model of how the HNH domain binds at the cleavage site. The Cas9:cat system has been used as a target structure for Targeted MD simulations (details are reported below).

I.1. Molecular Dynamics (MD) simulations

The above-mentioned model systems have been equilibrated and production runs have been performed using the Amber ff12SB force field, which includes the ff99bsc0 corrections for DNA(9) and the ff99bsc0+χ*OL3* corrections for RNA.(10, 11) The Åqvist(12) force field parameters for the Mg ions has been employed, which favor an octahedral coordination for the Mg ion. These parameters were employed in our previous studies on similar nucleases,(13-15) performing the catalysis of the DNA via a *"two-metal aided"* mechanism,(16) as suggested for Cas9.(1) Moreover, extensive testing in comparison with other four bonded and non-bonded Mg models (i.e., Allnér,(17) Li,(18) Oelschlaeger(19) and Saxena(20)) has been performed.(21) The TIP3P model has been employed for waters.(22) A salt concentration of 0.08 mM of NaCl has been considered, in agreement with the experimental conditions of cleavage assays.(3, 23) MD simulations have been carried out using an integration time step of 2 fs. Hydrogen atoms were added assuming standard bond lengths and were constrained to their equilibrium position with the SHAKE algorithm.

Temperature control (300 K) has been performed via Langevin dynamics,[24] with a collision frequency $\gamma = 1$. Pressure control was accomplished by coupling the system to a Berendsen barostat,[25] at a reference pressure of 1 atm and with a relaxation time of 2 ps. All the simulations were carried out with the following protocol. First, the systems were subjected to energy minimization to relax the water molecules and the counter ions, keeping the protein, as well as the RNA, DNA and Mg^{2+} ions fixed with harmonic position restraints of 300 kcal/mol · $Å^2$. Then, the systems were heated up from 0 to 100 K in the canonical ensemble (NVT), by running two NVT simulations of 5 ps each, imposing position restraints of 100 kcal/mol · $Å^2$ on the above-mentioned elements of the considered systems. The temperature was further increased up to 200 K in ~100 ps of MD in the isothermal-isobaric ensemble (NPT), in which the restraint was reduced to 25 kcal/mol x $Å^2$. Subsequently, all restraints were released and the temperature of the system was ultimately raised up to 300 K in a single NPT simulation of 500 ps. After ~ 1.1 ns of equilibration, ~ 10 ns of NPT production was carried out allowing the density of the system to stabilize around 1.01 g/cm^{-3}. Finally, production runs were carried out in the NVT ensemble, collecting ~120/150 ns for each system. All simulations have been performed with the GPU version of AMBER 16(26, 27) and the SPFP precision model.(28) These well-equilibrated systems have been used as a starting point for biased MD simulations and for Gaussian accelerated MD.

I.2. Gaussian accelerated Molecular Dynamics (GaMD) simulations

aMD is an enhanced sampling method that works by adding a non-negative boost potential to smoothen the system potential energy surface (PES), thus effectively decreasing the energy barriers and accelerating transitions between the low-energy states.(29, 30) Here, aMD simulations has been performed using the novel and more robust Gaussian aMD (or GaMD)(31) implementation, in which the boost potential follows Gaussian distribution, such allowing for accurate reweighting using cumulant expansion to the 2nd order. GaMD extends

the use of aMD to big-size biological systems, for which the standard reweighting procedure has been prohibitive, given the large statistical noise.(32, 33)

Considering a system with N atoms at positions $\vec{r} = \{\overrightarrow{r_1}, \ldots \overrightarrow{r_N}\}$, when the system potential $V(\vec{r})$ is lower than a threshold energy E, the energy surface is modified by adding a boost potential as:

$$V^*(\vec{r}) = V(\vec{r}) + \Delta V(\vec{r}), \qquad V(\vec{r}) < E, \tag{1}$$

$$\Delta V(\vec{r}) = \frac{1}{2}k\big(E - V(\vec{r})\big)^2, \tag{2}$$

where k is the harmonic force constant. The two adjustable parameters E and k are automatically determined by applying the following three criteria. First, for any two arbitrary potential values $V_1(\vec{r})$ and $V_2(\vec{r})$ found on the original energy surface, if $V_1(\vec{r}) < V_2(\vec{r})$, ΔV should be a monotonic function that does not change the relative order of the biased potential values, i.e., $V_1^*(\vec{r}) < V_2^*(\vec{r})$. Secondly, if $V_1(\vec{r}) < V_2(\vec{r})$, the potential difference observed on the smoothened energy surface should be smaller than that of the original, i.e., $V_2^*(\vec{r}) - V_1^*(\vec{r}) < V_2(\vec{r}) - V_1(\vec{r})$. By combining the first two criteria and plugging in the formula of $V^*(\vec{r})$ and ΔV, we obtain:

$$V_{max} \leq E \leq V_{min} + 1/k, \tag{3}$$

where V_{min} and V_{max} are the system minimum and maximum potential energies. To ensure that Eqn. [4] is valid, k has to satisfy $k \leq 1/V_{max} - V_{min}$. By defining $k \equiv k_0\, 1/V_{max} - V_{min}$, then $0 < k \leq 1$. Thirdly, the standard deviation of ΔV needs to be small enough (i.e., narrow distribution) to ensure accurate reweighting using cumulant expansion to the second order: $\sigma_{\Delta V} = k\big(E - V_{avg}\big)\sigma_V \leq \sigma_0$, where V_{avg} and σ_V are the average and standard deviation of the system potential energies, $\sigma_{\Delta V}$ is the standard deviation

of ΔV and σ_0 as a user-specified upper limit (e.g., 10 k_BT) for accurate reweighting. When E is set to the lower bound, $E = V_{max}$, according to Eqn. [4], k_0 can be calculated as:

$$k_0 = \min(1.0, k_0') = \min\left(1.0, \frac{\sigma_0}{\sigma_V} \cdot \frac{V_{max} - V_{min}}{V_{max} - V_{avg}}\right). \quad [4]$$

Alternatively, when the threshold energy E is set to its upper bound $E = V_{min} + 1/k$, k_0 is:

$$k_0 = k_0'' \equiv \left(1 - \frac{\sigma_0}{\sigma_V}\right) \cdot \frac{V_{max} - V_{min}}{V_{avg} - V_{min}}, \quad [5]$$

if k_0'' is calculated between *0* and *1*. Otherwise, k_0 is calculated using Eqn. [4], instead of being set to 1 directly as described in the original paper.(31)

GaMD has been applied on 24 equally distributed states along Cas9 activation trajectory, including the crystallographic structures and intermediate points for which no experimental data is available, which have been obtained by using a Targeted MD (TMD) approach (full details are reported below). In detail, 14 equally distributed frames describe the conformational transition between the X-ray structures of the apo Cas9 (frame #1) and of Cas9:RNA (frame #14), recovering the ~28 Å RMSD between the initial and the final X-ray structure in steps of ~2 Å. 10 additional frames, describe the conformational transition of the HNH domain between the X-ray structures of Cas9:RNA (frame #14), Cas9:DNA (frame #18), Cas9:pre-cat (frame #22) and the model structure of the catalytic Cas9 (frame #24), proposed by Sternberg (**Fig. S2**).(8) The experimental FRET distances between the Cα atoms of the N1054–S867 and S867–S355 couples of residues, which describe the rearrangement of the HNH domain, have been used as criteria for ensuring an equal distribution of the frames.(8)

Based on extensive tests, the system threshold energy has be set to $E = V_{max}$ for all GaMD simulations. The boost potential has been applied in a *dual-boost* scheme, in which two acceleration potentials are applied simultaneously to the system: *(i)* the torsional terms

only and *(ii)* across the entire potential. A timestep of 2 fs has been used. Given an average system size of ~220K atoms, the maximum, minimum, average and standard deviation values of the system potential ($V_{max}, V_{min}, V_{avg}$ and σ_V) has been obtained from an initial ~12 ns NPT simulation with no boost potential. Each GaMD simulation proceeded with a ~80 ns run, in which the boost potential has been updated every 1.6 ns, thus reaching equilibrium values. Finally, ~ 400 ns of GaMD simulations have been carried out in the NVT *ensemble*. For each system, ~400 ns (GaMD production) + 80 ns (GaMD equilibration) + 12 ns (pre-equilibration classical MD) have been carried out. Since GaMD has been applied on 24 equally distributed states along Cas9 activation trajectory, a total of ~11.8/12 μs (i.e., ~492 ns * 24 systems) of molecular simulations has been produced. These simulations have been performed with the GPU version of AMBER 16.(26) Importantly, GaMD simulations have been preceded by the equilibration of the 24 selected TMD structures by means of classical MD simulations. These runs have been performed for ~120/150 ns, for a total of ~2.8/3 μs (i.e., ~120 ns * 24 systems), as described above (paragraph 1.2). Taken together, a total of ~15 μs of molecular simulations has been carried out (including ~11.8/12 μs of GaMD and ~2.8/3 μs of conventional MD).

I.3. Energetic reweighting of GaMD simulations

The potential of mean force (PMF) is calculated upon accurate reweighting of the simulations using cumulant expansion to the 2nd order. For simulations of a biomolecular system, the probability distribution along a selected reaction coordinate $A(r)$ is written as $p^*A(r)$, where r denotes the atomic positions $\{r_1, \dots, r_n\}$. Given the boost potential $\Delta V(r)$ of each frame, $p^*A(r)$ can be reweighted to recover the canonical ensemble distribution, $p(A)$, as:

$$p(A_j) = p^*(A_j) \frac{\langle e^{\beta \Delta V(r)} \rangle_j}{\sum_{j=1}^{M} \langle e^{\beta \Delta V(r)} \rangle_j} \qquad j = 1, \dots, M, \qquad [6]$$

where M is the number of bins, $\beta = 1/k_B T$, and $\langle e^{\beta \Delta V(r)} \rangle_j$ is the ensemble-averaged Boltzmann factor of $\Delta V(r)$ for simulation frames found in the j^{th} bin. To reduce the energetic noise, the ensemble-averaged reweighting factor can be approximated using cumulant expansion:

$$\langle e^{\beta \Delta V} \rangle = \exp\left\{\sum_{k=1}^{\infty} \frac{\beta^k}{k!} C_k\right\}, \quad [7]$$

where the first three cumulants are given by:

$$C_1 = \langle \Delta V \rangle,$$

$$C_2 = \langle \Delta V^2 \rangle - \langle \Delta V \rangle^2 = \sigma_{\Delta V}^2, \quad [8]$$

$$C_3 = \langle \Delta V^3 \rangle - 3\langle \Delta V^2 \rangle \langle \Delta V \rangle + 2\langle \Delta V \rangle^3 .$$

As shown earlier,(31-33) when the boost potential follows near-Gaussian distribution, cumulant expansion to the second order provides the more accurate reweighting compared with the exponential average and Maclaurin series expansion methods. Finally, the reweighted free energy is calculated as:

$$F(A_j) = -1\left(\frac{1}{\beta}\right) \ln p(A_i). \quad [9]$$

PMF calculations have been performed over the aggregate trajectories, employing the above described experimental FRET distances(8) as reaction coordinates, such describing the conformational change of the apo protein up to RNA binding, and the rearrangement of the HNH domain from the RNA-bound state up to the catalytic state. In detail, 14 aggregate trajectories from the apo-form up to the RNA-bound state (including ~5.6 μs) have been considered for deriving the PMF, which characterize the conformational transition of the apo protein toward RNA binding. 10 aggregate trajectories from the RNA-bound state up to the

catalytic state (including ~4 μs) have been employed for deriving the PMF relative to the HNH conformational change. The PMF profiles have been obtained by employing different bin sizes and by varying trajectory lengths, such enabling checking convergence of the GaMD simulations.(34) Error bars have been computed as the standard error of the PMF values calculated by varying trajectory lengths (i.e., over the entire production runs of ~400 ns, over the last ~300 ns and over the second half (i.e., last ~200 ns) of the trajectories).

I.4. Targeted Molecular Dynamics (TMD) simulations

In a TMD approach,(35) restraint forces are applied to reduce the root-mean-square deviation (RMSD) between initial and target conformations, according to the following potential:

$$U = \frac{k}{2N}\left(RMS(t) - RMS^{*}(t)\right)^{2} \quad [10]$$

where k is the collective force constant, N is the number of targeted atoms, $RMS(t)$ is the actual RMSD between the current (at time t) and target configurations, and $RMS^{*}(t)$ evolves from the RMSD between the initial and target configurations toward the desired final RMSD (i.e., zero). Forces have been applied to the crystallized protein heavy atoms, excluding the modeled loops. The motion of the nucleic acids has been unbiased. Each TMD run has been performed using a per-atom force constant (k/N) of 20 kcal mol^{-1} $A^{\circ -2}$. Based on our tests, this setup allowed smoothly dragging the targeted atoms into the final configuration, leading to the reorganization of the structure during the dynamics. By employing force constant of force constant (k/N) of 20 kcal mol^{-1} $A^{\circ -2}$, the RMSD between the simulated system and its targeted structure was decreased at a rate of approximately ~0.6 A° per ns. These simulations have been performed with NAMD 2.10.(36)

TMD simulations have been used to provide an initial pathway between the crystallized states, providing a set of equally distributed frames and starting points for unconstrained MD, which could not be obtained experimentally. TMD is used to recover the backward Cas9 activation trajectory, from the most complete structure of Cas9 in complex with both nucleic acids (Cas9:pre-cat) up to the apo Cas9. The simulation of the backward process has been performed by gradually removing the nucleic acid components, following the structural transitions of the protein that would occur upon dissociation of the nucleic acids (**Fig. S3**). TMD have been also performed to reach a putative catalytic state, by targeting Cas9:pre-cat into a model structure proposed by Sternberg et al.(8) (**Fig. S2**). The choice of simulating the backward process has been motivated by the fact that the characterization of the forward process would require the association the nucleic acid components *on-the-fly* during the simulations, going beyond the capability of MD simulations. However, with the purpose to provide an initial validation of the observed structural transitions occurring between crystallized states, multiple TMD simulations have been performed, for a total of ~220 ns, including the simulation of the forward direction, as detailed below.

Three TMD simulations have been performed: (TMD1) Cas9:pre-cat has been targeted in Cas9:DNA after removing the non-target DNA (*nt*-DNA) strand, (TMD2) Cas9:DNA has been targeted in Cas9:RNA after removing the DNA, (TMD3) Cas9:RNA has been targeted in the apo Cas9 after removing RNA. The forward process has also been simulated via three additional TMD, leading the apo Cas9 into Cas9:RNA (TMD4), Cas9:RNA into Cas9:DNA (TMD5) and Cas9:DNA into Cas9:pre-cat (TMD6). Moreover, a putative catalytic Cas9 state (Cas9:cat), has been built by engineering the HNH domain (**Fig. S2**).(8) This structure has been used as a target for TMD simulations, such allowing the targeting of Cas9:pre-cat into a catalytic state (TMD7). During TMD7, forces have been applied only at the HNH domain. Indeed, the target structure (Cas9:cat) has been obtained by

docking the HNH domain within the X-ray structure of Cas9:pre-cat.(4) This approach allowed the structural adaptation of the remaining residues to the docked HNH domain. In order to investigate the role of the nucleic acids during the critical step of the turn of the HNH domain (which occurs during the transition of Cas9:pre-cat into Cas9:DNA), Cas9:pre-cat has been targeted in Cas9:DNA after removing the entire DNA (TMD8) and all nucleic acids (TMD9). Finally, Cas9:pre-cat has been targeted in Cas9:RNA including all nucleic acid components (i.e., RNA, *t*-DNA and *nt*-DNA), in order to further understand the role of the *nt*-DNA strand in the HNH conformational change (TMD10). Overall, multiple TMD simulations, for a total of ~220 ns, have been performed, recovering the conformational changes between crystallized states back and forth.

I.5. Ensemble-Averaged Electrostatic Calculations

Ensemble-averaged electrostatics calculations were performed in the Visual Molecular Dynamics (VMD) package with the PME Electrostatics Plugin.(37) Snapshots were taken every 5 ps over the course of the ~400 ns GaMD simulation (40,000 snapshots) of the frames # 1 (apo Cas9), # 3, # 6 (R-exp state), #12 (pre-RNA bound) and #14 (Cas9:RNA) (**Fig. 3c** of the main text).

I.6. Analysis of the results

Analysis of TMD simulations. In order to track the conformational changes of the protein domains observed during TMD simulations, we have calculated the rotation angles θ and ϕ around the z component of the principal axes of inertia of the protein and of each domain. In detail, for each domain of Cas9, two reference systems are defined at time 0 (t_0) of MD (**Fig. S4a**), as composed by *(i)* the planes identified by the x and y components of the principal axes of inertia of the protein and by *(ii)* the planes identified by the x' and y' components of the principal axes of inertia of the protein domain. θ is calculated as the angle

between the z axis and the vector passing through the center of masses of the protein at t_0 (blue dot) and of the single domain along MD (i.e., t_1, t_2, …t_n; magenta dot), while ϕ is the angle between the z' axis and the vector passing through the center of masses of the protein domain at t_0 (green dot) and along MD (i.e., t_1, t_2, …t_n; magenta dot). θ is a measure of the rotation of the domain with respect to the protein, while ϕ is a measure of the rotation of the domain around itself (**Fig. S4b**). The displacement of each single domain with respect to the protein is calculated as the distance (d) between the center of masses of the protein and of the domain.

II. Additional Results

II.1. Additional results from Targeted MD (TMD)

TMD has been used to recover the backward Cas9 activation trajectory, from the most complete structure of Cas9 in complex with both nucleic acids (Cas9:pre-cat) up to the apo Cas9 (**Fig. S3**). The simulation of the backward process has been performed by gradually removing the nucleic acid components, following the structural transitions of the protein that would occur upon dissociation of the nucleic acids (**Fig. S3**). TMD have been also performed to reach a putative catalytic state, by targeting Cas9:pre-cat into a model structure proposed by Sternberg et al.(8) (**Fig. S2**). The choice of simulating the backward process has been motivated by the fact that the characterization of the forward process would require the association the nucleic acid components *on-the-fly* during the simulations, going beyond the capability of MD simulations. However, with the purpose to provide an initial validation and the reproducibility of the structural transitions observed between crystallized states, we performed multiple TMD rung, including simulations in the forward direction, for a total of ~220 ns (as detailed in the Methods section). Complete analysis and comparison of the forward and backward trajectories are reported below.

Targeted MD of Cas9:RNA in apo Cas9. Here, we report additional results from TMD simulations performed by targeting Cas9:RNA in the apo Cas9, after removing the RNA. TMD captures this conformational transition of Cas9 in ~50 ns, with results reported in **Figs. 2** (main text) and **S5**. Both backward and forward trajectories consistently reveal a large rotation of REC*III* region with respect to the protein (with a change of the θ angle of ~60°), as well as with respect to itself (ϕ ~90°), while the REC*II* region rotates with a smaller extent. The REC*I* region moves in opposite direction with respect to REC*II-III* and together with the R-rich helix, confirming previous observations indicating the concerted movement of these protein domains.(5) The nuclease lobe remains stable, as shown by the time evolution of the θ and ϕ angles (**Fig. S5**). As an important note, we remark that in the backward and forward simulations trajectories, the conformational changes occur in opposite directions, leading the values of the θ in the final state (time t = ~50 ns) of the backward trajectory to be coincident to the values of the θ in the initial state (t = 0 ns) of the forward trajectory. The ϕ angle assumes an initial value of zero in both backward and forward simulations, since it is a measure of the rotation of the protein domain around itself (details are in the Methods).

Targeted MD of Cas9:DNA in Cas9:RNA. Here, we report additional results from TMD simulations performed by targeting Cas9:DNA into Cas9:RNA, after removing the DNA. The initial RMSD between the X-ray structures is 12.2 Å. TMD recovers the process in ~18 ns. TMD well describes the structural transition of the HNH domain, with a modest shift of the regions REC*II-III* (**Fig. S6**). Both backward and forward simulation trajectories consistently depict the displacement of the HNH domain toward the target DNA strand, which is accompanied by a modest shift of the REC*II-III* regions.(2) It is important to note that we currently lack structural information of the binding of a double-stranded DNA within Cas9. As such, these simulations are only informative of the structural transitions occurring between the two crystal structures.

Targeted MD of Cas9:pre-cat in Cas9:DNA. Here, we report additional results from TMD performed by targeting Cas9:pre-cat in Cas9:DNA, after removing the *nt*-DNA strand. Results are reported in **Fig. 2** of the main text and in **Fig. S7**, showing the H-bond interactions established by the L2 loop (residues 906-918) and the RNA:*t*-DNA hybrid during the simulations. With the aim of understanding the role of the nucleic acids during the HNH repositioning, we also performed TMD simulations removing both DNA strands (w/o DNA, i.e., in the presence of the RNA only), and removing entirely the nucleic acids (w/o nucleic). TMD simulations have been also performed including all nucleic acids chains (i.e., RNA, *t*-DNA and *nt*-DNA). This latter simulation has not been performed in the forward direction, given that the *nt*-DNA does not fit in the X-ray structure of Cas9:DNA. **Fig. S8a** reports the time evolution of the ϕ angle (i.e., *"moon"*), calculated for the HNH domain, during the above-mentioned simulations. We found that in the absence of interactions between the L2 linker and the RNA:*t*-DNA hybrid, the L1 and L2 linkers engage in an unphysical overlap that precludes the complete turn of the HNH domain (**Fig. S8b**). The latter moves in the opposite direction with respect to what was observed in the presence of the RNA:*t*-DNA hybrid and with a ϕ angle of only ~50°. Finally, we also performed TMD simulations of Cas9:pre-cat targeted in Cas9:DNA including all nucleic acid chains (i.e., RNA, *t*-DNA and *nt*-DNA), revealing that the presence of the *nt*-DNA within the RuvC groove exerts a steric constraint that hampers the structural transition of HNH (ϕ does not change phase). These observations hint at a critical role of the L2 loop, which initiates proper conformational transition of the HNH domain by interacting with the RNA:*t*-DNA hybrid.(4) As a final note, during the simulations, the observed structural transitions mainly involve the HNH domain, given that in the X-ray structures of Cas9:pre-cat and Cas9:DNA the remaining protein domains show slight differences.

II.2. Additional results from Gaussian accelerated MD (GaMD)

Based on TMD simulations, a pathway for Cas9 conformational activation, among the crystallized states, has been identified. Along this pathway, 24 equally distributed frames have been selected and subjected to conventional and GaMD for a total of ~12 μs of simulation. Intermolecular FRET distances have been used to identify the conformational states of Cas9 along its activation trajectory (**Fig. 3a** of the main text).(8) The D435–E945 distance (calculated between the Cα atoms) is descriptive of the conformational transition from the apo form up to the RNA bound form, while the N1054–S867 and S867–S355 distances (calculated between the Cα atoms) describe the conformational transition of the HNH domain from the RNA-bound state up to the catalytic state. **Fig. S9** reports the probability distributions of the D435–E945 distance during conventional MD simulations (~100 ns) and GaMD (~400 ns) of the frames #1–14, which describe the conformational transition from the apo Cas9 up to the RNA bound state. **Fig. S10** shows the probability distributions of the N1054–S867 and S867–S355 distances during conventional MD simulations (~100 ns) and GaMD (~400 ns) of the frames #14–24 describing the conformational rearrangement of the HNH domain from the RNA-bound state up to the catalytic state. It is remarkable that during the simulations, the distribution of the FRET distances in the simulations of the apo Cas9, Cas9:RNA, Cas9:DNA are centered on the experimental values, while also being in agreement with the catalytic state proposed by Sternberg et al.(8) Moreover, the distribution of the FRET distances indicates the stability of the intermediate states identified via TMD. Indeed, unbiased MD starting from the intermediate states sample a nearby conformational subspace, with partial overlap of the FRET distances distribution, indicating the stability of the identified paths and providing a validation *a posteriori* of the TMD simulations. This is an important point, which adds to the initial validation done by performing multiple TMD simulations and supports the reliability of the biasing simulation method, as employed here. As mentioned above, GaMD enabled the

characterization of a catalytic Cas9, which well agrees with the state proposed by Sternberg et al. based on FRET experiments.(8) In this state, the catalytic H840 locates at a ~4-6 Å distance from the scissile phosphate on the *t*-DNA strand (**Fig. S12**), allowing water molecules to bridge the scissile phosphate and H840, in line with a *"one-metal ion"* catalytic mechanism.(1, 7)

Energetic reweighting of the GaMD simulations describing the conformational transition from the apo Cas9 to Cas9:RNA (frames #1–14) has been done using as reaction coordinates (RC) the D435–E945 distance and the RMSD with respect to the apo Cas9, describing the energetics in terms of Potential of Mean Force (PMF). The N1054–S867 and S867–S355 distances are used as RC for the reweighting of GaMD simulations and obtain the PMF for the conformational transitions of the HNH domain from the RNA-bound state up to the catalytic state (frames #14–24).

In order to obtain an overall description of the conformational transition of the apo Cas9 into the RNA-bound form, as well as of the HNH domain from the RNA-bound state up to the catalytic Cas9, the PMF have been computed over the aggregate trajectories of the #1–14 (from apo to RNA-bound) and the #14–24 (from RNA-bound to catalytic Cas9) frames, considering the production runs of ~400 ns for each trajectory. Results are reported in the main text (**Figs. 3** and **4**), as well as in **Figs. S13-S14**. Specifically, **Fig. S13** reports the PMFs computed by varying trajectory lengths (i.e., over the entire production runs of ~400 ns, over the last ~300 ns and over the second half (i.e., last ~200 ns) of the trajectories), such enabling the calculation of the error associated to the PMF (**Figs. 3** and **4** of the main text). **Fig. S14** reports the PMFs computed by employing different bin sizes, suggesting convergence of the presented GaMD simulations.(34)

III. References

1. Jinek M*, et al.* (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* 343(6176):1247997.
2. Jiang F, Zhou K, Ma L, Gressel S, Doudna JA (2015) STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* 348(6242):1477-1481.
3. Anders C, Niewoehner O, Duerst A, Jinek M (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* 513(7519):569-573.
4. Jiang FG*, et al.* (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* 351(6275):867-871.
5. Palermo G, Miao Y, Walker RC, Jinek M, McCammon JA (2016) Striking Plasticity of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular Simulations. *ACS Cent Sci* 2(10):756−763.
6. Biasini M*, et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* 42(W1):W252-W258.
7. Biertumpfel C, Yang W, Suck D (2007) Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature* 449(7162):616-U614.
8. Sternberg SH, LaFrance B, Kaplan M, Doudna JA (2015) Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* 527(7576):110-113.
9. Perez A*, et al.* (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* 92(11):3817-3829.
10. Banas P*, et al.* (2010) Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J Chem Theory Comput* 6(12):3836-3849.
11. Zgarbova M*, et al.* (2011) Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput* 7(9):2886-2902.
12. Aqvist J (1990) Ion-Water interaction Potentials Derived from Free Energy Perturbation Simulations. *J Phys Chem* 94(21):8021-8024.
13. L. Casalino, G. Palermo, U. Rothlisberger, Magistrato A (2016) Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns. *J Am Chem Soc* 138:10374–10377.
14. Palermo G*, et al.* (2015) Catalytic metal ions and enzymatic processing of DNA and RNA. *Acc Chem Res* 48(2):220-228.
15. Palermo G, Stenta M, Cavalli A, Dal Peraro M, De Vivo M (2013) Molecular Simulations Highlight the Role of Metals in Catalysis and Inhibition of Type II Topoisomerase. *J Chem Theory Comput* 9(2):857-862.
16. Steitz TA, Steitz JA (1993) A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci U S A* 90(14):6498-6502.
17. Allner O, Nilsson L, Villa A (2012) Magnesium Ion-Water Coordination and Exchange in Biomolecular Simulations. *J Chem Theory Comput* 8(4):1493-1502.
18. Li PF, Roberts BP, Chakravorty DK, Merz KM (2013) Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J Chem Theory Comput* 9(6):2733-2748.
19. Oelschlaeger P, Klahn M, Beard WA, Wilson SH, Warshel A (2007) Magnesium-cationic dummy atom molecules enhance representation of DNA polymerase beta in molecular dynamics simulations: Improved accuracy in studies of structural features and mutational effects. *J Mol Biol* 366(2):687-701.

20. Saxena A, Sept D (2013) Multisite Ion Models That Improve Coordination and Free Energy Calculations in Molecular Dynamics Simulations. *J Chem Theory Comput* 9(8):3538-3542.
21. Casalino L, Palermo G, Abdurakhmonova N, Rothlisberger U, Magistrato A (2017) Development of Site-Specific Mg2+-RNA Force Field Parameters: A Dream or Reality? Guidelines from Combined Molecular Dynamics and Quantum Mechanics Simulations. *J Chem Theory Comput* 13(1):340-352.
22. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* 79(2):926-935.
23. O'Connell MR*, et al.* (2014) Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature* 516(7530):263-266.
24. Turq P, Lantelme F, Friedman HL (1977) Brownian Dynamics - Its Application To Ionic-Solutions. *J Chem Phys* 66(7):3039-3044.
25. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81(8):3684-3690.
26. Case DA*, et al.* (2016) AMBER 2016. *University of California, San Francisco*.
27. Salomon-Ferrer R, Gotz AW, Poole D, Le Grand S, Walker RC (2013) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* 9(9):3878-3888.
28. Le Grand S, Goetz AW, Walker RC (SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput Phys Commun* 148(2):374–380.
29. Hamelberg D, de Oliveira CAF, McCammon JA (2007) Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J Chem Phys* 127(15): 155102.
30. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J Chem Phys* 120(24):11919-11929.
31. Miao Y, Feher VA, McCammon JA (2015) Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J Chem Theory Comput* 11(8):3584-3595.
32. Miao Y*, et al.* (2014) Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *J Chem Theory Comput* 10(7):2677-2689.
33. Miao Y, McCammon JA (2016) Graded activation and free energy landscapes of a muscarinic G protein-coupled receptor. *Proc Natl Acad Sci U S A* 113(43):12162–12167.
34. Trzesniak D, Kunz AP, van Gunsteren WF (2007) A comparison of methods to compute the potential of mean force. *ChemPhysChem* 8(1):162-169.
35. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nature Struct Biol* 9(9):646-652.
36. Phillips JC*, et al.* (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26(16):1781-1802.
37. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual Molecular Dynamics. *J Mol Graph* 14(1):33-38, 27-38.

IV. Supplementary Tables

Table S1. Experimental values of the intermolecular FRET distances(8) used to identify the conformational states of Cas9 in this work. Experimental values are measured between Cα atoms of the D435–E945, N1054–S867 and S867–S355 couples of residues of the X-ray structures,(8) relative to the apo Cas9 (4CMP.pdb), Cas9:RNA (4ZT0.pdb), Cas9:DNA (4UN3.pdb) and Cas9:pre-cat (5F9R.pdb). For Cas9:cat, the *d2–d3* distances are reported as by Sternberg et al.(8)

FRET pair	**apoCas9** (4CMP)	**Cas9:RNA** (4ZT0)	**Cas9: RNA:DNA** (4UN3)	**Cas9: pre-cat** (5F9R)	**Cas9:cat**
					Sternberg et al. Nature (2015)
D435–E945	21 Å	78 Å	83 Å	81 Å	–
S867–N1054	6 Å	7 Å	28 Å	45 Å	57 Å
S355–S867	79 Å	81 Å	59 Å	29 Å	21 Å

V. Supplementary Figures

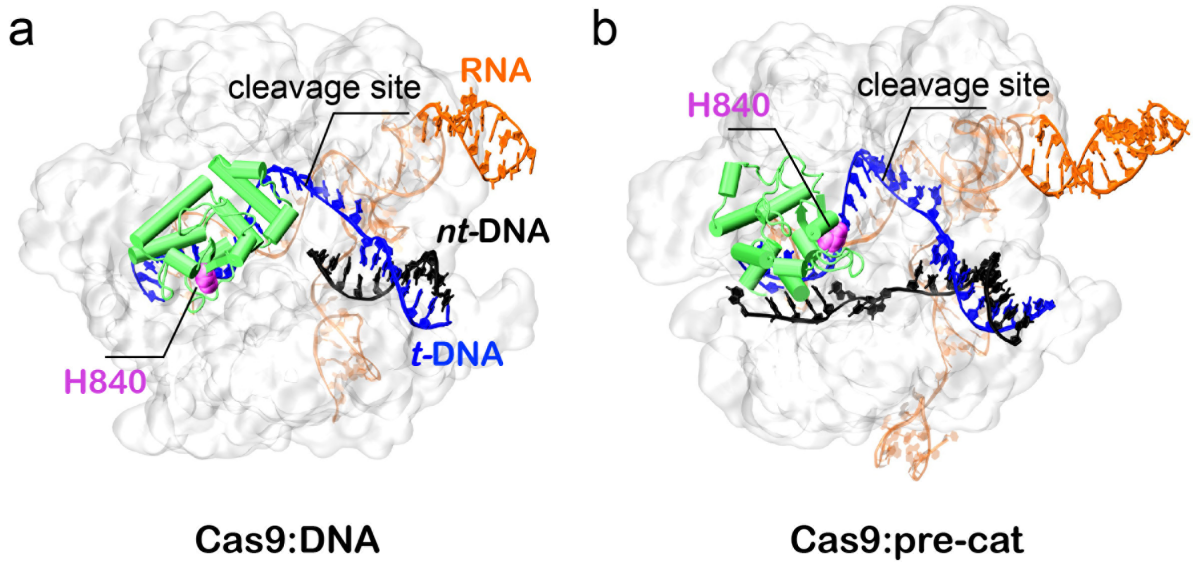

Fig. S1. X-ray structures of the *Streptococcus pyogenes* Cas9:DNA (4UN3.pdb)(3) **(a)** and Cas9:pre-cat (5F9R.pdb)(4) **(b)**, revealing different orientations of the HNH domain. Cas9 is shown in molecular surface, highlighting the HNH domain (green) using a cartoon representation. The RNA (orange), target DNA (*t*-DNA, blue) and non-target DNA (*nt*-DNA, black) are shown as ribbons. The catalytic H840 (magenta) is shown in space-filling representation. In Cas9:DNA, the catalytic H840 points in different direction with respect to the cleavage site on the *t*-DNA. In Cas9:pre-cat, a ~180° turn of the HNH domain is observed, such directing the catalytic His840 toward the scissile phosphate, from which it remains separated by a ~15 Å distance.

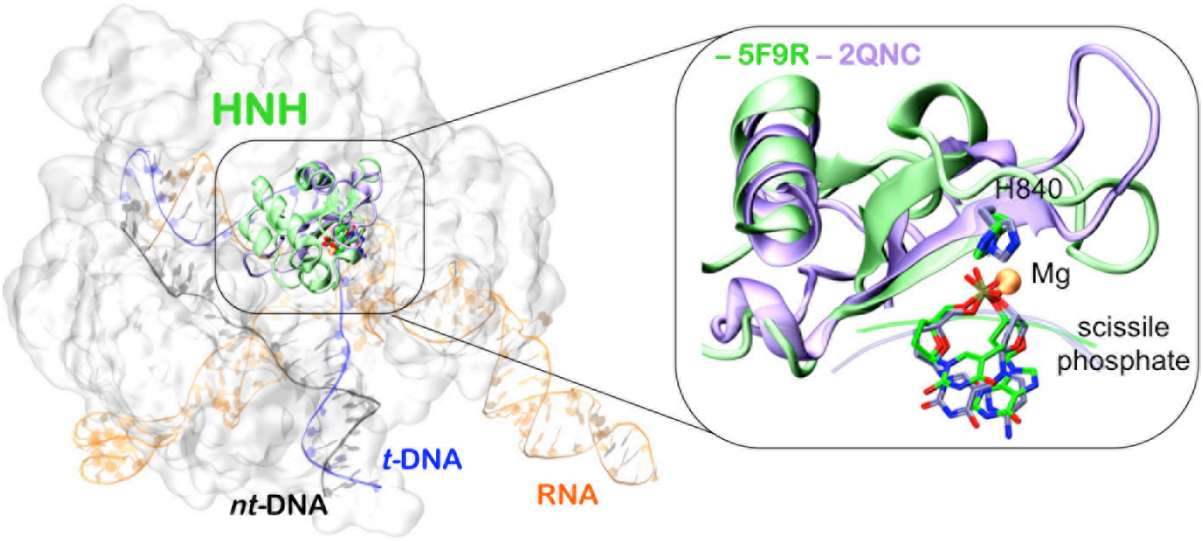

Fig. S2. Model of a putative catalytic state (Cas9:cat), built by docking the HNH catalytic site at the target DNA (*t*-DNA) cleavage site. Cas9 is shown in molecular surface, while the HNH domain (green) is shown as ribbons and is superposed to the catalytic domain of T4 endonuclease VII.(7) The box on the right details the superimposition of the aligned scissile phosphate and flanking nucleotides of a DNA in complex with T4 endonuclease VII (2QNC.pdb, violet)(7) with the scissile phosphate and flanking nucleotides of Cas9:pre-cat (59FR.pdb, green).

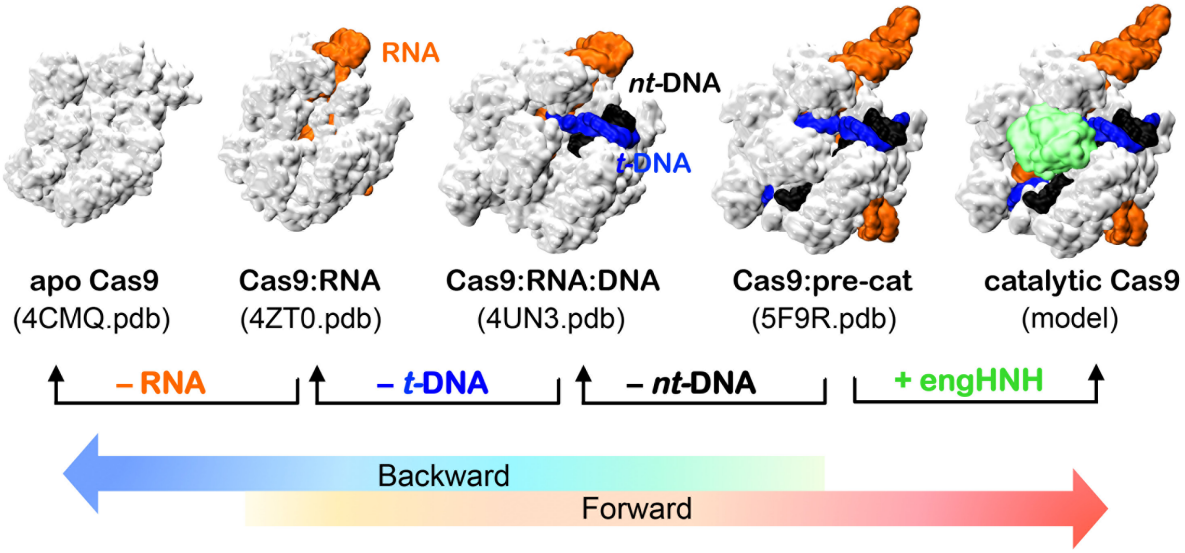

Fig. S3. Targeted MD (TMD) approach employed in this study. The backward activation trajectory (blue arrow) has been recovered from the most compete X-ray structure of Cas9 in complex with both nucleic acids (Cas9:pre-cat) up to the apo Cas9. Three TMD have been performed: Cas9:pre-cat (5F9R.pdb[4]) has been targeted in Cas9:DNA (4UN3.pdb[3]) after removing the non-target DNA strand (– *nt*-DNA), Cas9:DNA has been targeted in Cas9:RNA (4ZT0.pdb[2]) after removing the target DNA strand (– *t*-DNA), Cas9:RNA has been targeted in the apo Cas9 (4CMQ.pdb[1]) after removing RNA (– RNA). The forward process (red arrow) has been simulated too, leading the system from the apo state into a putative catalytic state in which the HNH domain has been engineered (+ engHNH). For this purpose, a putative catalytic Cas9 state (Cas9:cat) has been built by engineering the HNH domain, as done by Sternberg(8) (**Fig. S2**) and used as a target structure leading the pre-catalytic Cas9 into its catalytic state. The protein (silver), as well as the RNA (orange), *t*-DNA (blue) and *nt*-DNA (black) are shown as molecular surface. In the catalytic Cas9, the engineered HNH domain is highlighted in green.

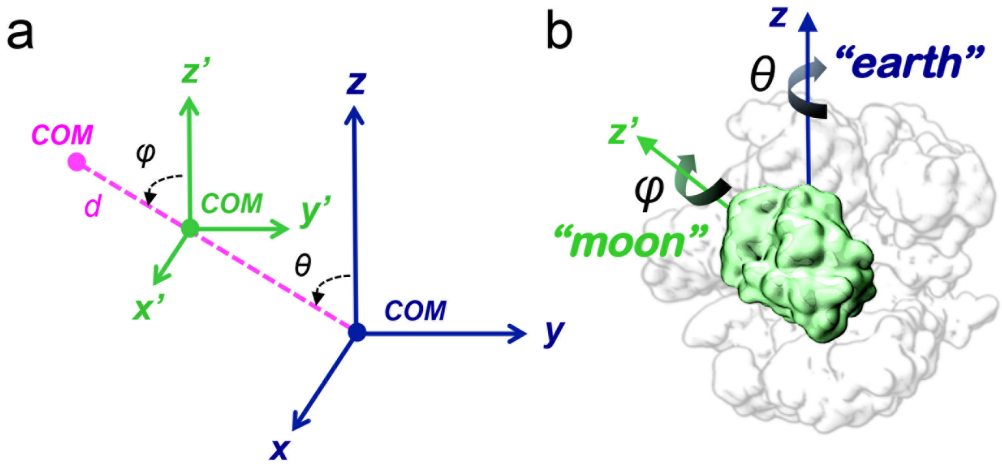

Fig. S4. Schematic representation of the θ and ϕ angles, which characterize the rotation of each individual protein domain with respect to the protein (θ) and itself (ϕ). The θ and ϕ angles define an *"earth & moon"* model **(b)**, in which each individual protein domain (here shown for HNH, green) can rotate around the main protein axis, like a satellite rotation around the earth (θ angle), and around itself (ϕ angle).

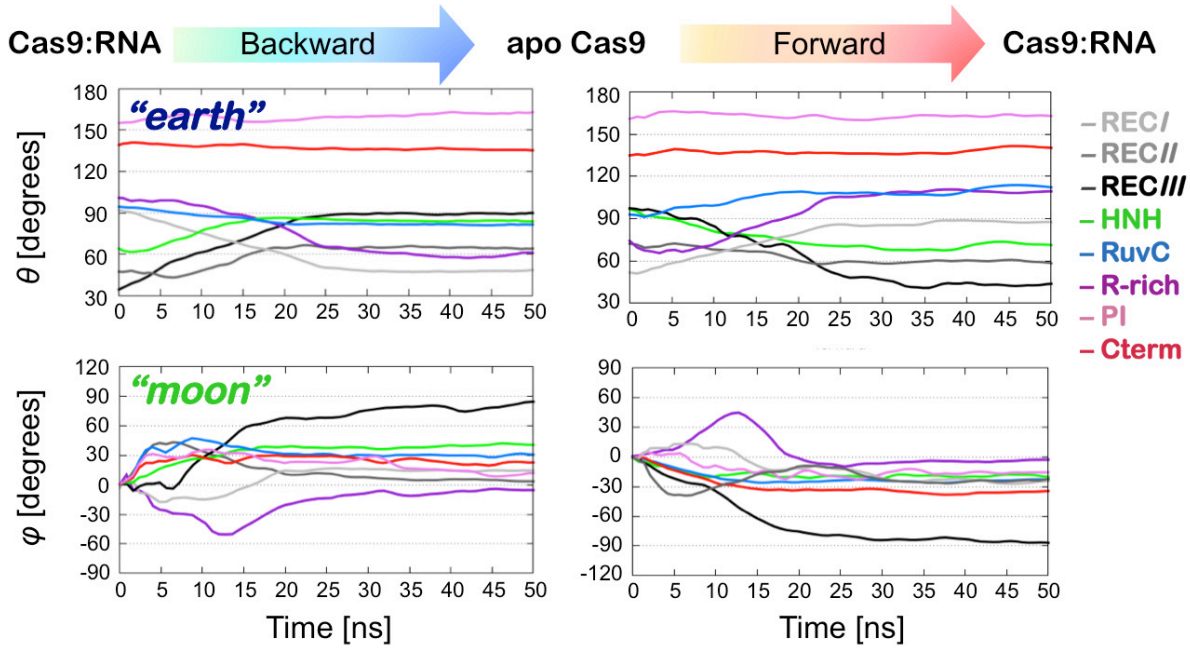

Fig. S5. Time evolution of the *"earth & moon"* angles θ (top graphs) and ϕ (bottom graphs), calculated for the individual domaisn of Cas9, during the process of RNA association, as simulated by targeting Cas9:RNA in the apo Cas9 (backward trajectory, left) and vice versa (foreward trajectory, right). For the sake of the clarity, data are smoothen using a Bézier function.

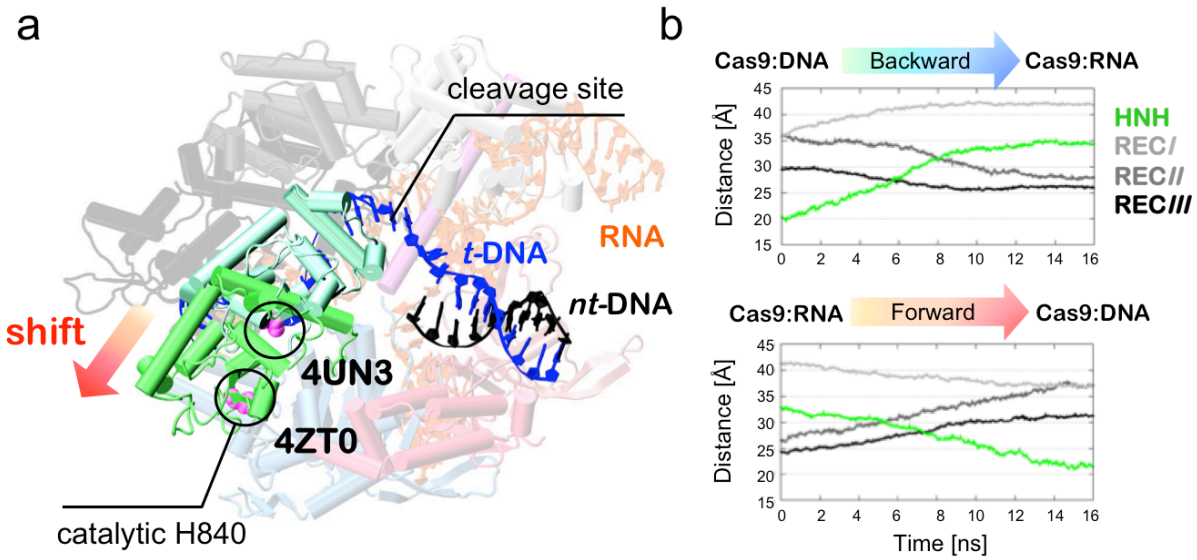

Fig. S6. (a) Shift of the HNH domain occurring upon DNA binding, shown as a superposition of the Xray structures of Cas9:RNA (4ZT0.pdb) and Cas9:DNA (4UN3.pdb). The protein is shown as cartoon, highligting the different configurations assumed by the HNH domain in dark (Cas9:RNA) and light (Cas9:DNA) green. The RNA (orange), target DNA (*t*-DNA, blue) and non-target DNA (*nt*-DNA, black) are shown as ribbons. The catalytic H840 (magenta) is shown in space filling representation. Targeted MD recovers this structural transition in ~16 ns (full details in the main text). **(b)** Time evolution of the distances between the Center Of Mass (COM) of the Cas9 protein and its individual domains (HNH, REC*I-III*), along Targeted MD simulations of Cas9:DNA targeted in Cas9:RNA, upon deletion of the DNA (backward trajectory) and Cas9:RNA targeted in Cas9:DNA (forward trajectory).

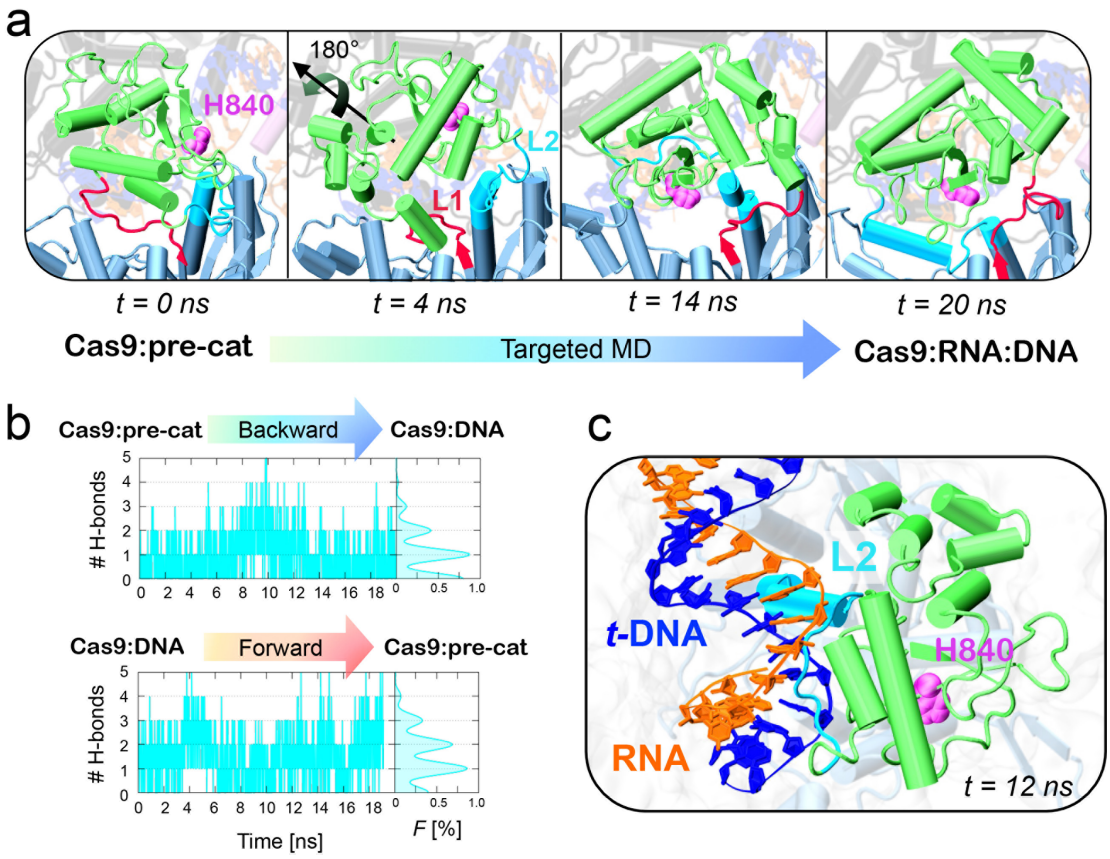

Fig. S7. (a) Representative snapshots along Targeted MD (TMD) of Cas9:pre-cat targeted to Cas9:DNA. The conformational changes of the protein are shown at time t = ~0, ~4, ~14 and ~20 ns. During the simulation, the HNH domain (green) turns ~180° around itself, while the catalytic H840 (magenta) changes configuration facing the RNA:t-DNA hybrid. The L1 (red) and L2 (cyan) loops also show remarkable structural transitions. **(b)** Occurrence (i.e., number) and relative probability distribution of the H-bonds established by the L2 loop (residues 906-918) and the RNA:*t*-DNA hybrid during Targeted MD simulations of Cas9:pre-cat targeted in Cas9:DNA (backward trajectory, top) and of Cas9:DNA targeted in Cas9:pre-cat (forward trajectory, bottom). **(c)** Representative snapshot from the simualtions showing the interactions between the L2 loop (cyan) and the RNA:*t*-DNA hybrid (shown at time t = 12 ns). Cas9 is shown as cartoon, highlighting the HNH domain in green. The catalytic H840 (magenta) is shown in space filling representation. The RNA (orange) and *t*-DNA (blue) are shown as ribbons.

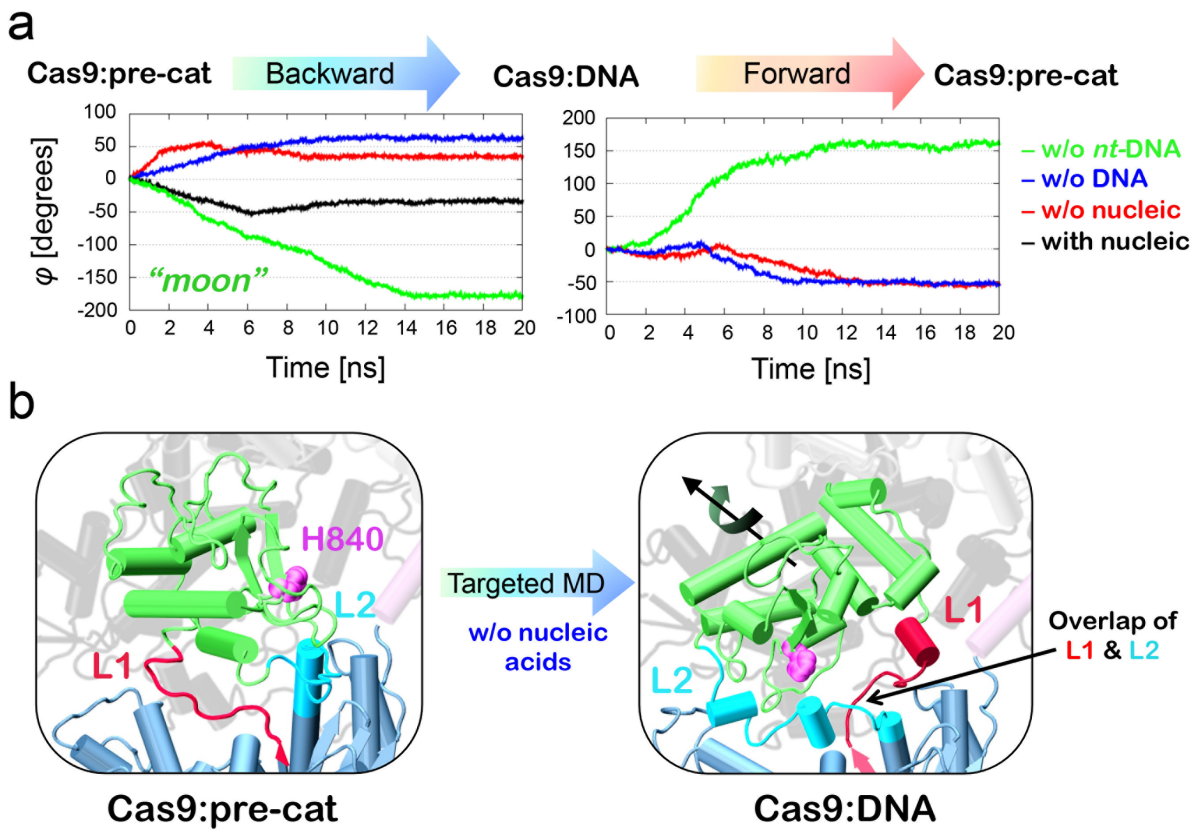

Fig. S8. (a) Time evolution of the ϕ angle (i.e., *"moon"*), calculated for the HNH domain, along Targeted MD simulations of Cas9:pre-cat targeted in Cas9:DNA (backward trajectory, left) and of Cas9:DNA targeted in Cas9:pre-cat (forward trajectory, right). Data are reported for TMD simulations performed witout the non-target DNA (w/o *nt*-DNA, green), without both DNA strands (w/o DNA, blue), without all nucleic acids (w/o nucleic, red) and including all nulcleic acids (with nucleic, black). This latter simulation has been performed only in the backward direction. **(b)** Representative snapshot from TMD simualtions of the backward process performed in the absence of the nucleic acids. During the simulation, the HNH domain turns with an extent of ϕ ~50° and in opposite direction, with respect to the simulation that includes the RNA and *t*-DNA strand (**Fig. 2** of the main text). This leads to an unphysical overlap of the L1 (red) and L2 (cyan) loops. Cas9 is shown as cartoon, highlighting the HNH domain in green. The catalytic H840 (magenta) is shown in space filling representation.

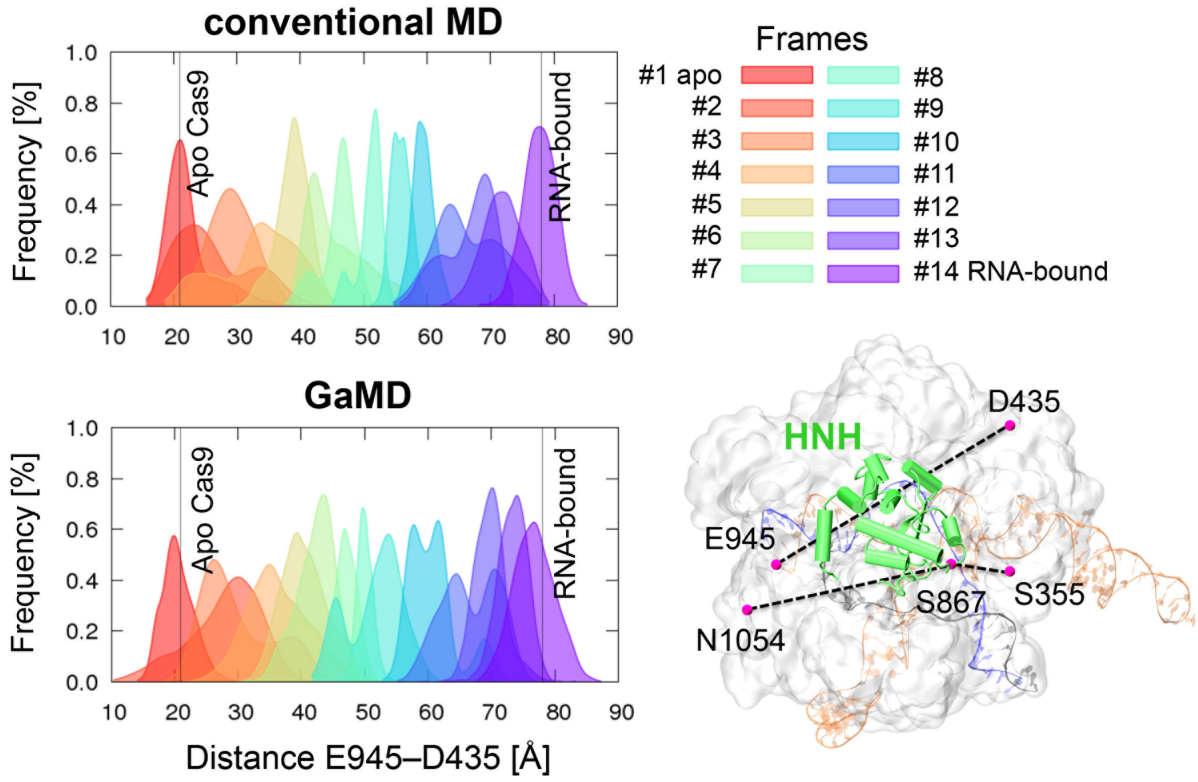

Fig. S9. Probability distributions of the D435–E945 distance (calculated between Cα atoms) during conventional MD simulations (~100 ns, top graph) and GaMD (~400 ns, bottom graph) of the frames #1–14, which describe the conformational transition from the apo Cas9 up to the RNA bound state. Experimental values of the D435–E945 distance in the apo Cas0 and RNA-bound states are indicated using vertical bars. A cartoon of Cas9 shows the three E945–D435, N1054–S867 and S867–S355 distances, which are representative of Cas9 conformational states.

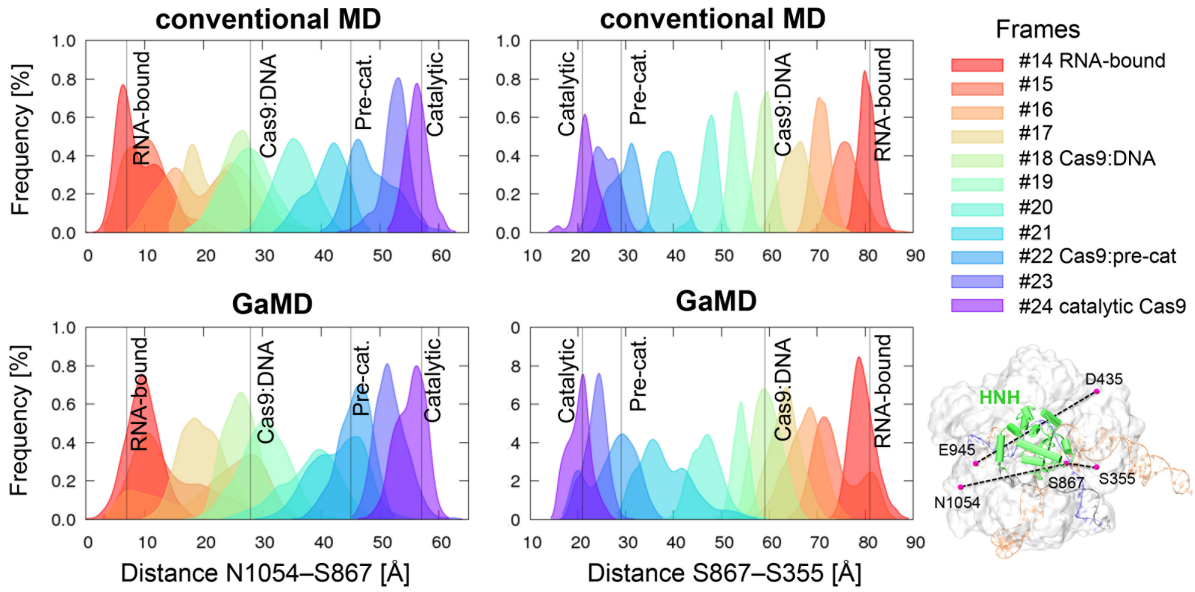

Fig. S10. Probability distributions of the N1054–S867 (left panel) and S867–S355 (right panel) distances during conventional MD simulations (~100 ns, top graphs) and GaMD (~400 ns, bottom graphs) of the frames #14–44, which describe the rearrangement of the HNH domain from the RNA-bound form up to the catalytic state. Distances are calculated between Cα atoms. Experimental values of the N1054–S867 and S867–S355 distances in the RNA-bound state, Cas9:DNA, Cas9:pre-cat, as well as in the putative catalytic state, as reported by Sternberg et al.(8) are indicated using vertical bars. A cartoon of Cas9 shows the three E945–D435, N1054–S867 and S867–S355 distances, which are representative of Cas9 conformational states.

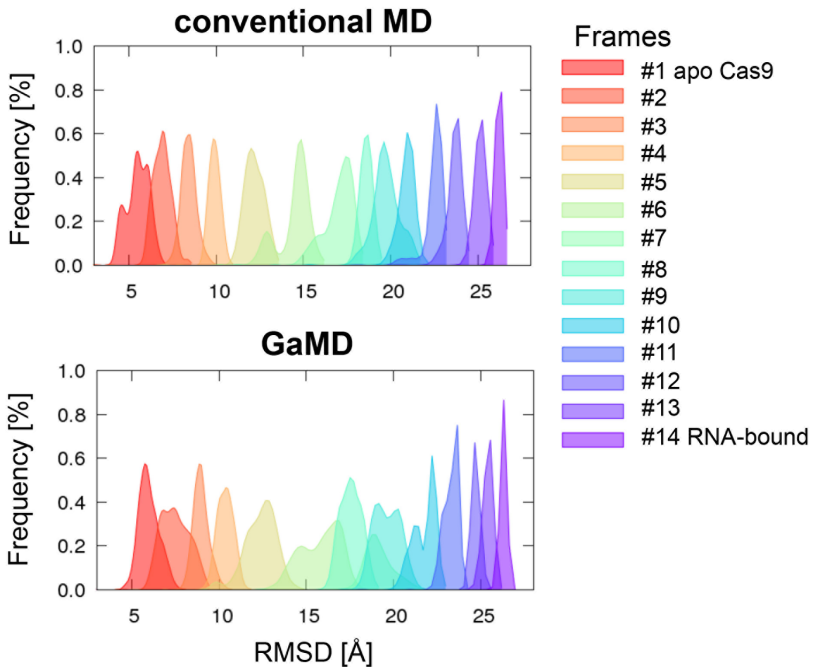

Fig. S11. Probability distributions of the RMSD calculated with respect to the apo Cas9 X-ray structure (4CMP.pdb)(1) during conventional MD simulations (~100 ns, top graph) and GaMD (~400 ns, bottom graph) of the frames #1–14, which describe the conformational transition from the apo Cas9 up to the RNA bound state. The RMSD distribution in the RNA-bound state results more peaked, due to the binding stabilization of the system.

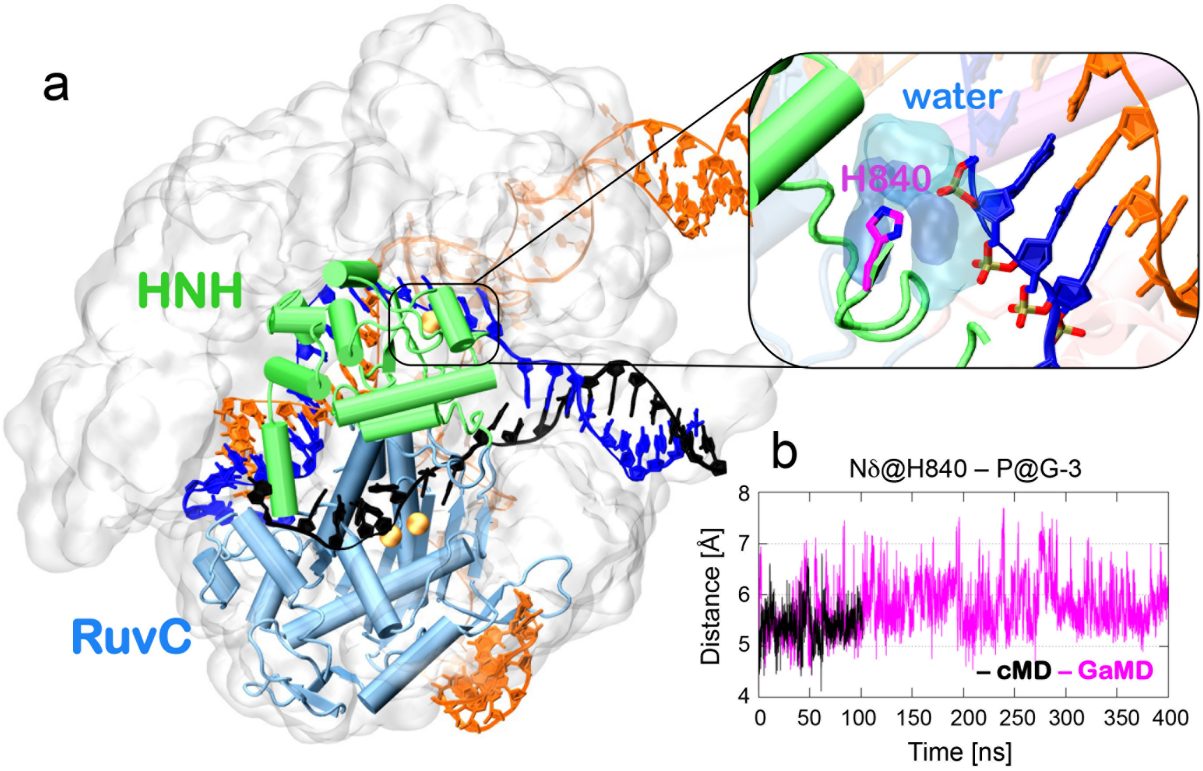

Fig. S12. Catalytic state of CRISPR-Cas9, as from GaMD simulations. The protein is shown in molecular surface, highlighting the HNH (green) and RuvC (blue) domains in cartoon view. The RNA (orange), target DNA (*t*-DNA) and non-target DNA (*nt*-DNA) strands are shown as ribbons. Catalytic Mg ions are shown as gold spheres. A close view of the HNH active site, showing a shell of water moleucles bridging H840 and the scissile phosphate, in agreement with the mechanistic hypothesis of a *"one-metal ion"* mechanism. Accordingly, H840 would act as a general base deprotonating the water nucleophile for the attack to the scissile phosphate.(1, 7) **(b)** Time evolution of the distance between Nδ of H840 and P@G-3 (scissile phosphate of the non-target DNA) along ~100 ns of conventional MD (cMD, black) and ~400 ns of GaMD (magenta) simulations of the catalytic state of the CRISPR-Cas9 system. Nδ@H840 stably loctes at ~5-6 Å from the scissile phosphate (P@G-3), allowing space for water molecules, as in a *"one-metal aided"* architecture.(1, 7)

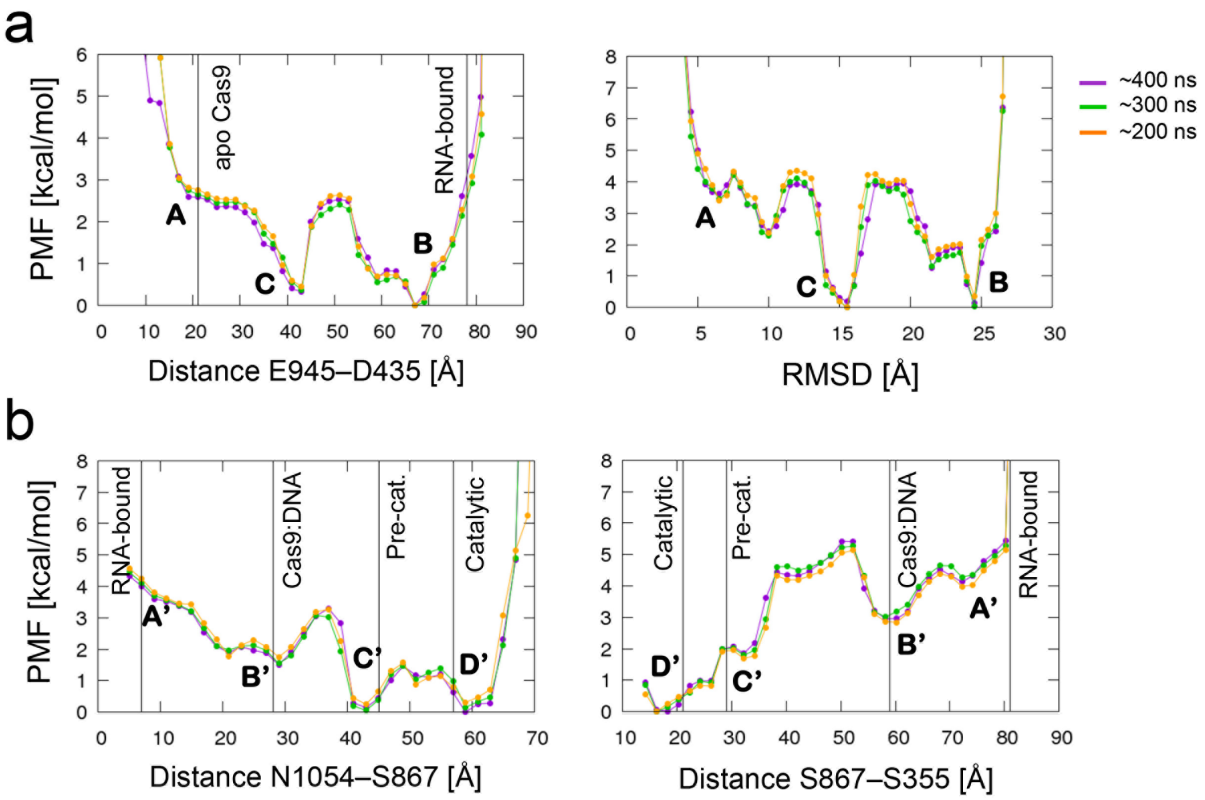

Fig. S13. (a) One-dimensional Potential of Mean Force (PMF) of the D435–E945 (left panel) and of the RMSD with respect to the apo Cas9 X-ray structure (4CMP.pdb, right panel),(1) calculated by varying trajectory lengths (i.e., over the entire production runs of ~400 ns, over the last ~300 ns and over the second half (i.e., last ~200 ns) of the trajectories). By varying trajectory length, both PMFs reveal three local minima, corresponding to the crystallographic apo (A) and RNA-bound (B) states, and to an intermediate state (C), characterized by the exposure toward the solvent of the arginine residues of the R-rich helix (i.e., R-exp state). **(b)** One-dimensional PMF of the N1054–S867 (left panel) and S867–S355 (right panel) distances, calculated by varying trajectory lengths. The PMFs reveal four local minima, corresponding to the crystallographic states of Cas9:RNA (A'), Cas9:DNA (B') Cas9:pre-cat (C') and to the putative catalytic state (D'). Experimental values of the E945–D435, N1054–S867 and S867–S355 distances are indicated using vertical bars. Distances are calculated between Cα atoms.

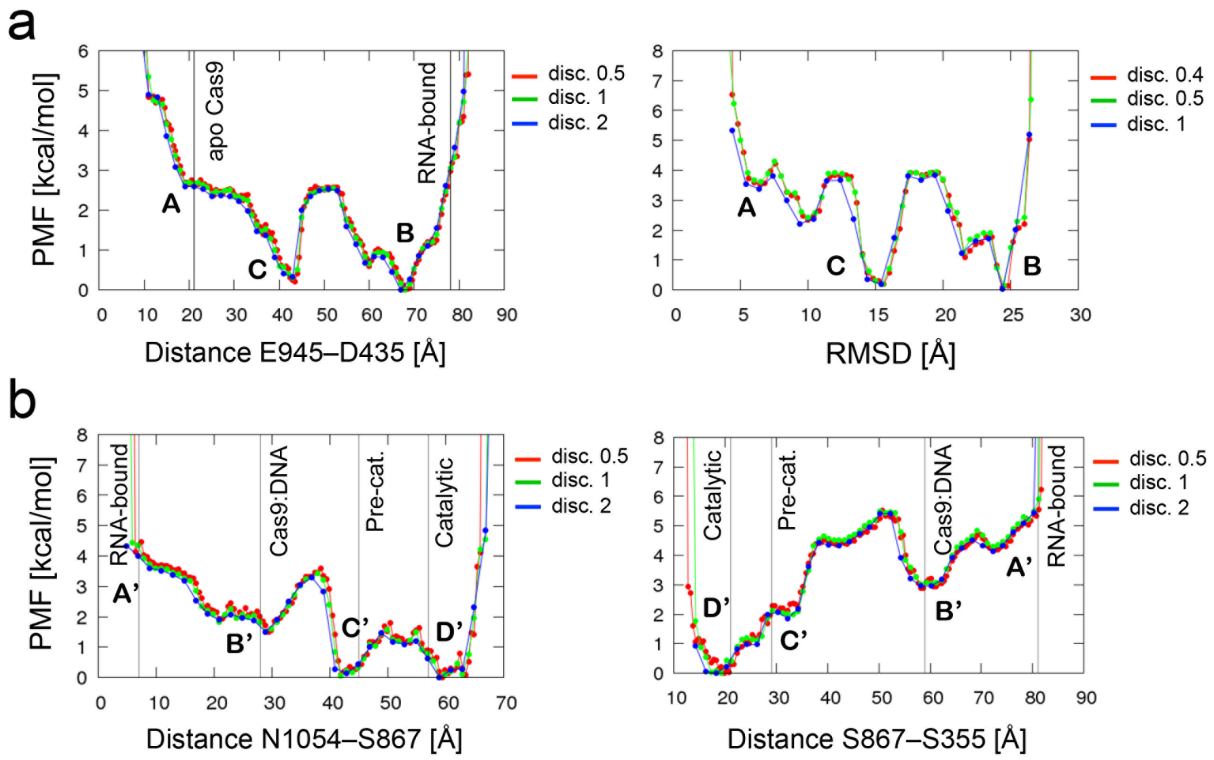

Fig. S14. (a) One-dimensional Potential of Mean Force (PMF) of the D435–E945 (left panel) and of the RMSD with respect to the apo Cas9 X-ray structure (4CMP.pdb, right panel),(1) calculated using different discretization values over the whole trajectories (i.e., ~400 ns per frame), as in the legend. Three local minima are identified, corresponding to the crystallographic apo (A) and RNA-bound (B) states, and to an intermediate state (C), characterized by the exposure toward the solvent of the arginine residues of the R-rich helix (i.e., R-exp state). **(b)** One-dimensional PMF of the N1054–S867 (left panel) and S867–S355 (right panel) distances, calculated using different discretization values over the whole trajectories (i.e., ~400 ns per frame). The PMFs reveal four local minima, corresponding to the crystallographic states of Cas9:RNA (A'), Cas9:DNA (B') Cas9:pre-cat (C') and to the putative catalytic state (D'). Experimental values of the E945–D435, N1054–S867 and S867–S355 distances are indicated using vertical bars. Distances are calculated between Cα atoms.