Supporting Information

Striking plasticity of CRISPR-Cas9 and key role of non-target DNA, as revealed by molecular simulations

Giulia Palermo, Yinglong Miao, Ross C. Walker, Martin Jinek and J. Andrew McCammon

Table of contents

1. Supplementary Materials and Methods
   1.1 Structural models
   1.2 Classical MD simulations
   1.3 Analysis of the results

2. Supplementary Results
   2.1. Analysis of RMSF and RMSD
   2.2 Correlation analyses
   2.3 MD simulations of Cas9:pre-cat in the absence of the nt-DNA
   2.4. H-bond network established by the nt-DNA in Cas9:pre-cat
   2.5. MD simulations of Cas9:RNA:DNA and Cas9:pre-cat in absence of the nucleic acids
   2.6. Convergence tests

3. Supplementary References

4. Supplementary Figures

5. Supplementary Movies

Corresponding authors:
Dr. Giulia Palermo (gpalermo@ucsd.edu)
1. Supplementary Materials and Methods

1.1. Structural models. Classical MD simulations have been performed on four model systems of the Cas9 in apo form (apo Cas9)\(^1\) and in complex with RNA (Cas9:RNA),\(^2\) with an incomplete DNA (Cas9:RNA:DNA)\(^3\) and in a pre-catalytic state (Cas9:pre-cat)\(^4\) including both DNA strands. These model systems have been prepared using the crystallographic coordinates of the *Streptococcus pyogenes* apo Cas9 (4CMQ.pdb),\(^1\) Cas9:RNA (4ZT0.pdb),\(^2\) Cas9:RNA: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. In order to study the effect of the non-target DNA (nt-DNA) on the dynamics of the pre-catalytic state, a fifth model system has been built, deleting the nt-DNA strand from Cas9:pre-cat (Cas9:pre-cat w/o nt-DNA). Moreover, with the purpose of studying the conformational dynamics of HNH in the absence of the nucleic acids, two additional model systems have been built deleting the nucleic acids from the DNA bound states (i.e., Cas9:RNA:DNA and Cas9:pre-cat, corresponding to the PDB codes 4UN3 and 5F9R), which differ in the orientation of the HNH domain. Missing residues of the 4UN3 and 4CMQ X-ray structures have been added via homology modeling, using SwissModel by Schwede.\(^5\) A total of 7 model systems has been embedded 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:RNA:DNA, ~216K atoms), ~180*116*139 Å\(^3\) (Cas9:pre-cat with and w/o nt-DNA, ~270K atoms each) and of ~136*103*144 Å\(^3\) (Cas9:RNA:DNA and Cas9:pre-cat without nucleic acids, ~190K atoms each).

The pre-catalytic state (Cas9:pre-cat, 5F9R.pdb)\(^4\) has been crystallized in the absence of the catalytic Mg metal ions. However, given the importance of Mg ions during the catalytic process, catalytic Mg ions have been included in the RuvC and HNH active sites, as done in Jinek et al.\(^1\) In detail, the Mg-bound RuvC domain of Cas9:pre-cat has been reconstructed by superposing the X-ray structure of the apo Cas9 (4CMQ.pdb),\(^1\) in which the co-crystallized Mn ions have been substituted with Mg. In the case of the HNH domain, the Mg-bound catalytic site has been obtained by superposing the domain to the apo structure of *Actinomyces naeslundii* Cas9 (4OGC.pdb),\(^1\) which includes Mg. All model systems based on the 5F9R.pdb include Mg ions within the catalytic sites.
1.2. Classical 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 and the ff99bsc0+χOL3 corrections for RNA. The Åqvist force field parameters for the Mg ions been employed, which favor an octahedral coordination for the Mg ion. These parameters were employed in our previous studies on similar nucleases, performing the catalysis of the DNA via a “two-metal aided” mechanism, as suggested for Cas9. Moreover, extensive testing in comparison with other four bonded and non-bonded Mg models (i.e., Allnér, Li, Oelschlaeger and Saxena) has been performed. The TIP3P model has been employed for waters. Na+ counter-ions have been used to neutralize the total charge. Hydrogen atoms were added assuming standard bond lengths and were constrained to their equilibrium position with the RATTLE algorithm implemented in NAMD. MD simulations have been performed in the isothermal–isobaric (NPT) ensemble by using a time step of 2 fs. The systems have been coupled to a Langevin thermostat at 298 K and barostat at 1 atm. Periodic boundary conditions were applied. PME method was used to evaluate long-range electrostatic interactions, and a cutoff of 12 Å was used to account for the van der Waals interactions. All the simulations were carried out with the following protocol. First, the systems were subjected to energy minimization by using the Steepest Descent algorithm. Then, the systems were thermalized up to physiological temperature in the canonical ensemble (NVT) using a Langevin bath in three consecutive steps: (1) the solvent was first equilibrated in ~10 ps of MD, slowly increasing the temperature from 0 to 100 K and maintaining both the protein and the nucleic acids fixed, (2) the temperature was further increased up to 200 K in ~10 ps of MD, while keeping fixed only the coordinates of backbone atoms of the protein/nucleic acids complex, (3) constraints were released and the systems were simulated for ~25 ps of MD to reach the temperature of 298 K. Then, we switched to the NPT statistical ensemble, performing ~100 ps of MD at 298 K. After this initial phase, equilibration runs were carried out in the NPT statistical ensemble, obtaining ~40 ns of MD at 298 K. Production runs have been carried out reaching ~1.5 µs for each system, for a total of > 10 µs of classical MD (i.e., ~1.5 µs x 7 systems). Coordinates of the systems were collected every 10 ps for a total of ~150,000/160,000 frames for each run.
1.3. Analysis of the results

Principal Component Analysis (PCA). PCA has been employed to capture the essential motions of the simulated systems. In PCA, the covariance matrix of the protein Cα atoms is calculated and diagonalized to obtain a new set of coordinates (eigenvectors) to describe the system motions. Each eigenvector – also called Principal Component (PC) – is associated to an eigenvalue corresponding to the mean square fluctuation contained in the system’s trajectory projected along that eigenvector. By sorting the eigenvectors according to their eigenvalues, the first Principal Component (PC1) corresponds to the system’s largest amplitude motion, and the dynamics of the system along PC1 is usually referred as “essential dynamics”. In this work, each structure arising from the MD trajectories is projected into the collective coordinate space defined by the first two eigenvectors (PC1 and PC2), such allowing the characterization the conformational space sampled by Cas9 during MD. Importantly, in order to identify differences in the essential structural-dynamic properties of Cas9, each simulated system has been superposed onto the same reference structure (i.e., considering as a reference the RuvC and Cterm domains that do not show relevant conformational differences among the crystallized states) and aligned, such allowing the projection into the same collective coordinate space. PCA has been performed using the GROMACS 4.4.5 suite of analysis codes, which have also been used for the analysis of RNSD and RMSF.21 In detail, the g_covar program has been employed for the construction and diagonalization of the covariance matrix. Subsequently, the program g_anaeig has been used to analyze and visualize the eigenvectors. The two codes have been employed as detailed in the following script (Full details are reported in the GROMACS 4.4.5 manual):

---- Script for PCA using GROMACS 4.4.5. ----
## Construction and diagonalization of the covariance matrix:
g_covar -s structure.tpr -f trajectory.trr -o eigenvalues.xvg -v eigenvectors.trr -last 10 -ascii covar.dat -xpm covar.xpm -av average.pdb
## Analysis of the eigenvectors
g_anaeig -s structure.tpr -v eigenvectors.trr -f trajectory.trr -proj proj.xvg -last 10
---- End Of File ---
Figure 2 in the main text has been produced using the Normal Mode Wizard (NMWiz) plugin of the Visual Molecular Dynamics (VMD) molecular visualization program.\textsuperscript{22}

**Cross-Correlation (CC\textsubscript{ij}) and Generalized-Correlation (GC\textsubscript{ij}) analysis.** The Cross-Correlation matrix $CC\textsubscript{ij}$ – based on Pearson coefficients – between the fluctuations of the C\textalpha{} atoms relative to their average positions has been used in order to identify the coupling of the motions between the protein residues. $CC\textsubscript{ij}$ has been calculated using equation [1],

$$CC\textsubscript{ij} = \frac{\langle \Delta r_i(t) \Delta r_j(t) \rangle}{\langle (\Delta r_i(t)^2) (\Delta r_j(t)^2) \rangle^{1/2}}$$ \hspace{1cm} [1]

where $\Delta r_i$ and $\Delta r_j$ are the fluctuation vectors of the atoms $i$ and $j$, respectively. The angle brackets represent an average over the sampled time period. The value of $CC\textsubscript{ij}$ ranges from -1 to 1. Positive $CC\textsubscript{ij}$ values describe a correlated motion between atoms $i$ and $j$, while negative $CC\textsubscript{ij}$ values describe anti-correlated motions.

The Generalized-Correlation ($GC\textsubscript{ij}$) analysis has also been employed.\textsuperscript{23} In comparison with the more traditional Pearson coefficients analysis, $GC\textsubscript{ij}$ has the advantages of being independent of the relative orientation of the atomic fluctuations, while also being able to capture non-linear correlations. Two variables can be considered independent if their joint probability distribution, $p(x_i, x_j)$, is equal to the product of their marginal distributions, $p(x_i) \cdot p(x_j)$. If $x_i$ assumes values restricting the range of values accessible to $x_j$, the joint probability is smaller than $p(x_i) \cdot p(x_j)$. This mutual information ($MI$) reveals the degree of correlation between $x_i$ and $x_j$ and is defined as:

$$MI [x_i, x_j] = \int \int p(x_i, x_j) \ln \frac{p(x_i, x_j)}{p(x_i)p(x_j)} dx_i dx_j$$ \hspace{1cm} [2]

Considering that the expectation value (or Hamiltonian, $H$) of a random variable $x$, having a probability distribution $p(x_i)$, following the concept of Shannon entropy, is:

$$H[x] = \int p(x) \ln p(x) dx$$ \hspace{1cm} [3]
MI of the variable $x_i$ and $x_j$ can be computed as:

$$MI \left[ x_i, x_j \right] = H \left[ x_i \right] + H \left[ x_j \right] - H \left[ x_i, x_j \right]$$  \[4\]

where $H \left[ x_i \right]$ and $H \left[ x_j \right]$ are the marginal entropies, and $H \left[ x_i, x_j \right]$ is the joint entropy. The g_correlation tool,\textsuperscript{23} which has been built within Gromacs 3.3\textsuperscript{24} and used in this work, the marginal entropies $H \left[ x_i \right]$ and $H \left[ x_j \right]$ and the joint entropy $H \left[ x_i, x_j \right]$ are estimated employing the k-nearest neighbor distances algorithm,\textsuperscript{25} applied to the atomic positions fluctuations from MD simulations. Since MI varies from 0 to $+\infty$, the $GC_{ij}$ coefficients are defined as in Eq. [5], ranging from 0 (independent variables) to 1 (correlated variables):

$$GC_{ij} \left[ x_i, x_j \right] = \left\{ 1 - e^{-2MI\left[ x_i, x_j \right]/d} \right\}^{-1/2}$$  \[5\]

Cross-correlation and generalized correlation analyses have been performed considering the last 1.2 $\mu$s of the trajectories. The trajectories have been further divided in windows of 400 ns, overlapping with each other, such allowing the average of the obtained matrices.

**Correlation Score ($Cs_i$).** Cross-Correlation Score ($Cs_i$) coefficients have been calculated for each residue as:

$$Cs_i = \sum_{i \neq j}^N CC_{ij}$$  \[6\]

$Cs_i$ coefficients are a measure of the number and intensity of the correlated ($0 < CC_{ij} < 1$) and anti-correlated ($-1 < CC_{ij} < 0$) motions displayed by each residue. To filter non-trivial correlations, per-residue $Cs_i$ have been computed considering the most positive ($CC_{ij} \geq 0.6$) and negative ($CC_{ij} \leq -0.6$) Pearson correlations. To shed light on intra and inter domain correlations, $Cs_i$ have been calculated for each residue $i$, with the residues $j$ belonging to the same protein domain of the residue $i$ ($Cs_i^{intra}$) and excluding the protein domain to which $i$ belongs ($Cs_i^{inter}$). By detecting the
protein residues that are highly correlated/anti-correlated, the $C_s$ function helps in identifying how specific protein regions mechanistically intervene in the overall motions. Finally, in order to detail the inter-dependent motions of the protein regions moving lockstep (i.e., as characterized by a $CC_{ij} > 0$) or showing opposite motions ($CC_{ij} < 0$), $C_{si}^{\text{inter}}$ have been accumulated over all residues $j$ of each specific Cas9 domain and plotted as a two-by-two matrix (Figure 3 of the main text). Complete analysis of $C_{si}^{\text{intra}}$ and $C_{si}^{\text{inter}}$ is reported as supplementary results below, as well as in the main text.

**Cluster analysis.** Cluster analysis has been performed using the method described by Daura et al., in which for each point (i.e., MD frame), the algorithm calculates the number of other frames for which the RMSD is lower than a given cutoff (neighbors). A cutoff of 2.5 Å has been used for the analyses of Cas9.
2. Supplementary Results

2.1. Analysis of RMSF and RMSD. Root Mean Square Fluctuations (RMSF) of the protein Cα atoms, calculated along the dynamics of the apo Cas9, Cas9:RNA, Cas9:RNA:DNA and Cas9:pre-cat systems is reported in Figure S1. Time evolution of the Root Mean Square Deviation (RMSD) for the protein and nucleic acids components are reported in Figure S2-3.

2.2. Correlation analyses. Cross-Correlation ($CC_{ij}$) and Generalized-Correlation ($GC_{ij}$) matrices have been computed over the last ~1.2 µs of MD simulations for the apo Cas9, Cas9:RNA, Cas9:RNA:DNA and Cas9:pre-cat systems (Figure S4-5). While $CC_{ij}$ analysis allows detecting the collinear correlations between atoms, such identifying if atoms move in lockstep ($0 < CC_{ij} < 1$) or in an opposite way ($–1 < CC_{ij} < 0$), non-linear correlations are captured via $GC_{ij}$ analysis. As a result, both $CC_{ij}$ and $GC_{ij}$ correlation matrices show highly correlated motions in the apo and RNA-bound states, while correlations become less relevant upon association with DNA. In the most complete system (Cas9:pre-cat), weak correlations are observed, indicating the stabilization of the system upon binding of both DNA strands.

Furthermore, we have computed the per-residue Correlation Scores ($Cs_{i}$), which are a measure of the number and intensity of the correlated and anti-correlated motions for each residue (full details in the Supplementary Methods). To shed light on intra and inter domain correlations, $Cs_{i}$ have been calculated for each residue $i$, with the residues $j$ belonging to the same protein domain of the residue $i$ ($Cs_{i}^{\text{intra}}$) and excluding the protein domain to which $i$ belongs ($Cs_{i}^{\text{inter}}$). Figure S6 reports the $Cs_{i}^{\text{intra}}$ (left column) and $Cs_{i}^{\text{inter}}$ (right column), as calculated from the $CC_{ij}$ matrix calculated over the last ~1.2 µs of MD simulations for each simulated system. Both $Cs_{i}^{\text{intra}}$ and $Cs_{i}^{\text{inter}}$ track intense correlations in the apo and RNA-bound states, while becoming less relevant in the DNA-bound Cas9. The plot of the $Cs_{i}^{\text{intra}}$ shows, for all simulated systems, mainly positive per-residue correlations within their protein domain. $Cs_{i}^{\text{inter}}$ shows that, in the apo and RNA-bound states, the α-helical region RECIII, as well as the PI and Cterm domains, which mediate the nucleic acids binding, are characterized by the most anti-correlated per-residue $Cs_{i}^{\text{inter}}$. In order to detail the inter-dependent motions of the protein regions moving lockstep (i.e., as characterized by a $CC_{ij} > 0$) or showing opposite motions ($CC_{ij} < 0$), $Cs_{i}^{\text{inter}}$ have
been accumulated over all residues $j$ of each specific Cas9 domain and plotted as a 2-by-2 matrix (Figure 3 of the main text).

Overall, correlation analyses characterize the motions of the protein domains (complete discussion is reported in the main text). Figure S7 reports the Ca-Ca vector map of the of Cas9:RNA (4ZT0.pdb) versus the apo Cas9 (4CMQ.pdb), allowing the comparison of our computational outcomes with the experimental data.

2.3. MD simulations of Cas9:pre-cat in the absence of the nt-DNA. Figure S8 reports the key outcomes from MD simulations of the Cas9:pre-cat system in the absence of the nt-DNA (w/o nt-DNA). Panel (a) reports the principal direction of movements (PC1) – i.e., “essential dynamics” – of the HNH domain plotted on the protein molecular surface. Panel (b) shows the Cross-Correlation ($CC_{ij}$, left) and Generalized-Correlation ($GC_{ij}$, right) matrices calculated over the last ~1.2 µs of MD.

2.4. H-bond network established by the nt-DNA in Cas9:pre-cat. Figure S9 shows the H-bonding network established by the nt-DNA with the L1 (residues 765–780) and L2 (residues 906–918) protein loops, during MD simulations of Cas9:pre-cat. It is noteworthy that, while an extended H-bonding network between the A-4–A-9 nucleotides of the nt-DNA with L1 is characteristic of the X-ray structure, as well as conserved during MD, H-bond interactions between L2 and the nt-DNA are not observed in the X-ray structure and in the first part of the dynamics, and the K119–C-3 H-bonding occurs at ~0.75 µs.

2.5. MD simulations of Cas9:RNA:DNA and Cas9:pre-cat in absence of the nucleic acids. MD simulations have also been performed deleting the nucleic acids from the DNA bound states (i.e., Cas9:RNA:DNA and Cas9:pre-cat), which differ for the orientation of the HNH domain, with the purpose of studying the conformational dynamics of HNH in the absence of the nucleic acids. As a result, particularly high RMSF values are detected for the HNH domain (Figure S11), which indicate the high conformational mobility of the HNH domain. It is interesting to note that, during the dynamics of Cas9:RNA:DNA, the HNH domain undergoes a conformational shift and resembles the configuration observed in the X-ray structure of Cas9 in complex with RNA (Figure S12). Contrariwise, in the Cas9:pre-cat system, the HNH domain undergoes an opposite conformational transition. A cluster analysis (details are
reported in the Supplementary Methods) reveals two distinctive conformational states of the HNH domain, that are predominant in the two systems.

2.6. Convergence tests. The analyses here reported (i.e., correlation analyses, PCA, cluster analysis) have been performed over the last ~1.2 µs of MD simulations for each simulated system, such allowing proper equilibration of the systems under investigation. In order to check the convergence of our MD runs, we have considered as a test-case the simulation of the Cas9:RNA:DNA system. We have utilized the generalized coefficients ($GC_{ij}$) matrix as a metrics for checking the convergence of our results. Indeed, spanning from 0 to 1, $GC_{ij}$ represent an easy tool for convergence tests. In detail, the $GC_{ij}$ matrix has been calculated in two different time windows of 0.6 µs (i.e., first ~0.6 µs and last ~0.6 µs). The resulting $GC_{ij}$ matrices are reported in Figure S13. Subsequently, for each of these two $GC_{ij}$ matrices, the RMSD with the $GC_{ij}$ matrix computed over the last ~1.2 µs has been computed, resulting in 0.055 (first ~0.6 µs) and 0.054 (last ~0.6 µs). This indicates that the MD results are robust and converged for a time window of ~0.6 µs.
3. Supplementary References

(1) Jinek, M.; Jiang, F.; Taylor, D. W.; Sternberg, S. H.; Kaya, E.; Ma, E.; Anders, C.; Hauer, M.; Zhou, K.; Lin, S.; Kaplan, M.; Iavarone, A. T.; Charpentier, E.; Nogales, E.; Doudna, J. A. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **2014**, *343*, 1247997.

(2) Jiang, F.; Zhou, K.; Ma, L.; Gressel, S.; Doudna, J. A. STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **2015**, *348*, 1477-1481.

(3) Anders, C.; Niewohner, O.; Duerst, A.; Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **2014**, *513*, 569-573.

(4) Jiang, F. G.; Taylor, D. W.; Chen, J. S.; Kornfeld, J. E.; Zhou, K. H.; Thompson, A. J.; Nogales, E.; Doudna, J. A. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **2016**, *351*, 867-871.

(5) Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Cassarino, T. G.; Bertoni, M.; Bordoli, L.; Schwede, T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **2014**, *42*, W252-W258.

(6) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., 3rd; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92*, 3817-3829.

(7) Banas, P.; Hollas, D.; Zgarbova, M.; Jurecka, P.; Orozco, M.; Cheatham, T. E.; Sponer, J.; Otyepka, M. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **2010**, *6*, 3836-3849.

(8) Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E.; Jurecka, P. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886-2902.

(9) Aqvist, J.: Ion-Water interaction Potentials Derived from Free Energy Perturbation Simulations. *J. Phys. Chem.* **1990**, *94*, 8021-8024.
(10) L. Casalino; G. Palermo; U. Rothlisberger; Magistrato, A.: Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns. J. Am. Chem. Soc. 2016, 138, 10374-10377.

(11) Palermo, G.; Cavalli, A.; Klein, M. L.; Alfonso-Prieto, M.; Dal Peraro, M.; De Vivo, M. Catalytic metal ions and enzymatic processing of DNA and RNA. Acc. Chem. Res. 2015, 48, 220-228.

(12) Palermo, G.; Stenta, M.; Cavalli, A.; Dal Peraro, M.; De Vivo, M. Molecular Simulations Highlight the Role of Metals in Catalysis and Inhibition of Type II Topoisomerase. J. Chem. Theory Comput. 2013, 9, 857-862.

(13) Steitz, T. A.; Steitz, J. A. A general two-metal-ion mechanism for catalytic RNA. Proc Natl Acad Sci U S A 1993, 90, 6498-6502.

(14) Allner, O.; Nilsson, L.; Villa, A. Magnesium Ion-Water Coordination and Exchange in Biomolecular Simulations. J. Chem. Theory Comput. 2012, 8, 1493-1502.

(15) Li, P. F.; Roberts, B. P.; Chakravorty, D. K.; Merz, K. M. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. J. Chem. Theory Comput. 2013, 9, 2733-2748.

(16) Oelschlaeger, P.; Klahn, M.; Beard, W. A.; Wilson, S. H.; Warshel, A. 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. 2007, 366, 687-701.

(17) Saxena, A.; Sept, D. Multisite Ion Models That Improve Coordination and Free Energy Calculations in Molecular Dynamics Simulations. J. Chem. Theory Comput. 2013, 9, 3538-3542.

(18) Casalino L.; Palermo, G.; Abdurakhmonova, N.; Rothlisberger, U.; Magistrato, A. Towards rational development of site-specific Mg2+-RNA force field parameters: A handbook from combined molecular dynamics and quantum mechanics simulations. 2016, Submitted.

(19) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys. 1983, 79, 926-935.

(20) Andersen, H. C. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. J. Comput. Phys. 1983, 52, 24-34.
(21) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. Gromacs: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701-1718.

(22) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33-38.

(23) Lange, O. F.; Grubmuller, H. Generalized correlation for biomolecular dynamics. *Prot: Struct., Funct., Bioinf.* **2006**, *62*, 1053-1061.

(24) Lindahl, E.; Hess, B.; Van Der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306-317.

(25) Floyd, R. W. Algorithm-97 - Shortest Path. *Commun. Acm.* **1962**, *5*, 345-345.

(26) Ricci, C. G.; Silveira, R. L.; Rivalta, I.; Batista, V. S.; Skaf, M. S. Allosteric Pathways in the PPAR gamma-RXR alpha nuclear receptor complex. *Sci. Rep-Uk* **2016**, *6*, 19940.

(27) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Edit.* **1999**, *38*, 236-240.
4. Supplementary Figures

**Figure S1.** Root Mean Square Fluctuations (RMSF) of the protein Cα atoms for the production MD runs (i.e., over the last ~1.2 µs of MD) of the apo Cas9 (1\textsuperscript{st} row), Cas9:RNA (2\textsuperscript{nd} row), Cas9:RNA:DNA (3\textsuperscript{rd} row) and Cas9:pre-cat (4\textsuperscript{th} row) systems. Modeled residues are shown with red lines. The protein sequence is shown on top of the graphs, highlighting individual protein domains with different colors.
Figure S2. Time evolution of the Root Mean Square Deviation (RMSD) of the protein backbone atoms with respect to the X-ray coordinates of the apo Cas9 (blue), Cas9:RNA (cyan), Cas9:RNA:DNA (yellow) and Cas9:pre-cat (red) systems. The box on the right reports a comparison between the backbone RMSD for the modeled and crystallized residues of the apo Cas9 (top) and Cas9:RNA:DNA (bottom) systems. In the case of the apo Cas9 a higher RMSD, with respect to the remaining systems, is observed. Moreover, an increased RMSD is also noted for the modeled residues (full details in the SI text).
**Figure S3.** Time evolution of the Root Mean Square Deviation (RMSD) of the RNA (orange), target DNA (t-DNA, blue) and non-target DNA (nt-DNA, black) during MD simulations of the Cas9:RNA (top panel, left), Cas9:RNA:DNA (top panel, right) and Cas9:pre-cat (bottom panel, left) systems. Higher fluctuations of the RMSD of the RNA are observed in Cas9:pre-cat, given that this system presents a longer and solvent exposed RNA strand (shown on the right).
Figure S4. Cross-Correlation ($CC_{ij}$) matrices, as calculated from long time scale MD of the apo Cas9, Cas9:RNA, Cas9:DNA and Cas9:pre-cat systems. The strength of the computed correlations is color-coded blue–to–red. The protein sequence is shown on top of the maps, highlighting individual protein domains with different colors.
Figure S5. Generalized-Correlation ($GC_{ij}$) matrices, as calculated from long time scale MD of the apo Cas9, Cas9:RNA, Cas9:DNA and Cas9:pre-cat systems. The strength of the computed correlations is color-coded green–to–violet. The protein sequence is shown on top of the maps, highlighting individual protein domains with different colors.
Figure S6. Intra ($C_{S_i}^{\text{intra}}$, left column) and Inter ($C_{S_i}^{\text{inter}}$, right column) per-residue Correlation Score ($C_S$), as calculated from the $CC_{ij}$ matrix averaged over ~1.2 μs of classical MD simulations for the apo Cas9 (1st row), Cas9:RNA (2nd row), Cas9:RNA:DNA (3rd row) and Cas9:pre-cat (4th row) systems. The plot of the $C_{S_i}^{\text{inter}}$ indicates the extent of the correlated/anti-correlated motions of each domain with respect to all the others.\textsuperscript{26}
Figure S7. Ca–Ca vector map of Cas9:RNA (4ZT0.pdb)² versus the apo Cas9 (4CMQ.pdb),¹ plotted over the X-ray structure of Cas9:RNA, suggesting the relative directions of the protein domains during the association with the RNA. The arrows indicate the directions of domain motions of the first structure relative to the second.
**Figure S8.** (a) Principal direction of movements (PC1) – i.e., "essential dynamics" – of the HNH domain plotted on the protein molecular surface, as calculated over the dynamics of the Cas9:pre-cat system without the non target DNA (w/o nt-DNA). (b) Cross-Correlation ($CC_{ij}$, left panel, the strength of the computed correlations is color-coded blue-to-red) and Generalized-Correlation ($GC_{ij}$, right panel, the strength of the computed correlations is color-coded green–to–violet) matrices, calculated over the last ~1.2 µs of MD of the Cas9:pre-cat system w/o nt-DNA.
**Figure S9.** (a) Time evolution of the number of hydrogen bonds established by the non target DNA (nt-DNA) with the L1 (residues 765–780, top graph) and L2 (residues 906–918, bottom graph) loops, along MD simulations of the Cas9:pre-cat system. The figure on top shows the L1 (magenta) and L2 (blue) loops interacting with the bases of the nt-DNA (black). (b) Time evolution of the key H-bonds (i.e., donor–acceptor distance) formed by the residue K913 of the L2 loop and the C-3 base of the nt-DNA strand. H-bonds are formed with the backbone (top graph) and side chain (bottom graph) at ~0.75 µs, as indicated by a dashed bar. The figure on top shows the K913–C-3 H-bonding.
Figure S10. Steric clash between the non-target DNA (nt-DNA) and the HNH, –RuvC interconnecting region, as observed upon superimposition of the 5F9R.pdb (Cas9:pre-cat, including a complete nt-DNA),\(^4\) with the 4ZT0 (Cas9:RNA)\(^2\) (a) and 4UN3 (Cas9:RNA:DNA) (b) X-ray structures. The RNA (orange), target DNA (t-DNA, blue) and nt-DNA (black) are shown as crystallized in the 5F9R.pdb, while the protein framework is shown for the superposed 4ZT0 (Cas9:RNA) and 4UN3 (Cas9:RNA:DNA) structures. The protein is in molecular surface, highlighting the HNH (green) and RuvC (blue) domains. The tightening between the HNH and RuvC domains in the 4ZT0 and 4UN3 X-ray structures hampers the accommodation of the nt-DNA, as crystallized in 5F9R, and results in a series of steric clashes. The close views on the right show the steric clashes between the nt-DNA with the HNH domain, as well as with the HNH–RuvC interconnecting regions, which are constituted by the L1 (magenta) and L2 (dark-blue) loops.
Figure S11. Root Mean Square Fluctuations (RMSF) of the protein Cα atoms for the equilibrium MD runs of the Cas9:pre-cat (1\textsuperscript{st} row) and Cas9:RNA:DNA (2\textsuperscript{nd} row) systems, performed in the presence (black lines) and absence (blue lines) of the nucleic acids. The protein sequence is shown on top of the graphs, highlighting individual protein domains with different colors.
Figure S12. Conformational change of the HNH domain, as observed during MD simulations of the Cas9:RNA:DNA (a) and Cas9:pre-cat (b) systems, performed in the absence of the nucleic acids. The protein is shown as cartoon, highlighting the HNH domain in yellow (Cas9:RNA:DNA) and green (Cas9:pre-cat). For each system, two representative configurations of the HNH domain are shown, indicating the conformational change observed during MD using an arrow. During simulation of Cas9:RNA:DNA, the HNH domain undergoes a conformational shift and resembles the configuration observed in the X-ray structure of the RNA-bound state, while an opposite conformational transition is observed in the complete Cas9:pre-cat. An RMSD-based clustering analysis, performed over the last ~1.2 µs of MD is shown on the right for each of the simulated systems, showing that the two distinctive conformational states of the HNH domain belong to the most populated clusters of each simulation. Details on cluster analysis are reported in the Supplementary Methods.
Figure S13. Generalized Correlation matrices ($GC_{ij}$) of the Cas9:RNA:DNA system, calculated over two different time windows of 0.6 µs: first ~0.6 µs (a) and last ~0.6 µs (b), as well as over the complete production run (i.e., last ~1.2 µs) (c). The difference matrix ($\Delta GC_{ij}$) between $GC_{ij}$ matrix in (a) with the $GC_{ij}$ matrix of the complete production (c) is reported in (d), while the $\Delta GC_{ij}$ matrix between $GC_{ij}$ matrix in (b) with the $GC_{ij}$ matrix of the complete production (c) is reported in (e). The calculated RMSD is also reported.
5. Supplementary Movies

Movie S1. “Essential dynamics” (i.e., Principal Component – PC) of the apo Cas9 (4CMQ.pdb). ¹ The protein is shown as tubes, color-coded by mobility (i.e., blue: less amplitude motions; red: highest amplitude motions).

Movie S2. “Essential dynamics” of the Cas9:RNA system (4ZT0.pdb). ² The protein is shown as tubes, color-coded by mobility (i.e., blue: less amplitude motions; red: highest amplitude motions).

Movie S3. “Essential dynamics” of the Cas9:RNA:DNA system (4UN3.pdb). ³ The protein is shown as tubes, color-coded by mobility (i.e., blue: less amplitude motions; red: highest amplitude motions).

Movie S4. “Essential dynamics” of the Cas9:pre-cat system (5F9R.pdb). ⁴ The protein is shown as tubes, color-coded by mobility (i.e., blue: less amplitude motions; red: highest amplitude motions).

Movie S5. Molecular dynamics simulations of Cas9:pre-cat (5F9R.pdb), ⁴ performed by deleting the non-target DNA (nt-DNA) strand from the system (i.e., Cas9:pre-cat w/o nt-DNA). The movie shows that, during the simulations, the HNH domain moves far apart from the catalytic site on the target DNA, reaching the catalytic H480 a distance of ~25/27 Å (initially ~18 Å) from the scissile phosphate on the target DNA (t-DNA) strand. The Cas9 protein in shown in molecular surface, with the HNH domain (green) highlighted as cartoon. The RNA (orange) and t-DNA (blue) strands are shown in ribbons. H480 (magenta) is shown as sticks.

Movie S6. Molecular dynamics simulations of Cas9:pre-cat (5F9R.pdb), ⁴ including both DNA strands, as captured by X-ray crystallography. The movie shows that, during the simulations, the HNH domain stably locates the catalytic H480 at a ~15 Å distance from the cleavage site on the target DNA (t-DNA) strand, as favored by a stable H-bonding interactions between C-3 in the non-target DNA (nt-DNA) and the K913 of the L2 loop, which occurs over the dynamics at ~0.75 µs. The Cas9 protein in shown in molecular surface, with the HNH domain (green) highlighted as cartoon. The RNA (orange), t-DNA (blue) and nt-DNA (black) strands are shown in ribbons. H480 and K913 (magenta) are shown as sticks.