Allosteric Motions of the CRISPR-Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics

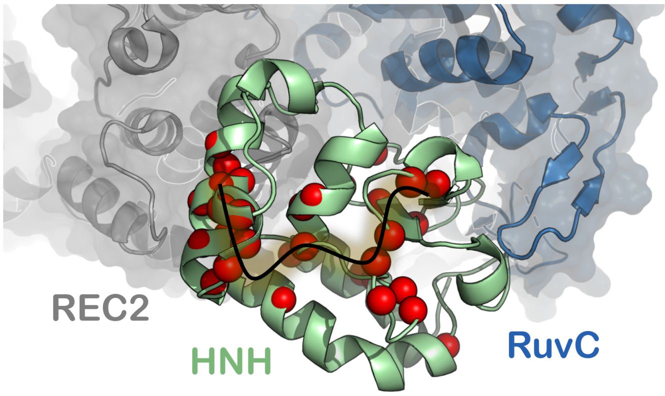
Kyle W. East,[1] Jocelyn C. Newton,[1] Uriel N. Morzan,[2] Atanu Acharya,[2,a] Erin Skeens,[1] Gerwald Jogl,[1] Victor S. Batista,[2] Giulia Palermo*,[3] and George P. Lisi*,[1]

1. Department of Molecular Biology, Cell Biology & Biochemistry, Brown University, Providence, RI 02903, United States

2. Department of Chemistry, Yale University, New Haven, CT 06520 , United States

3. Department of Bioengineering, University of California Riverside, 900 University Avenue, Riverside, CA 52512, United States

KEYWORDS: CRISPR-Cas9, protein dynamics, allostery, NMR, MD simulations



ABSTRACT: CRISPR-Cas9 is a widely employed genome-editing tool with functionality reliant on the ability of the Cas9 endonuclease to introduce site-specific breaks in double-stranded DNA. In this system, an intriguing allosteric communication has been suggested to control its DNA cleavage activity through flexibility of the catalytic HNH domain. Here, solution NMR experiments and a novel Gaussian accelerated Molecular Dynamics (GaMD) simulations method – flanked by mixed machine learning and structure-based prediction of NMR chemical shifts – are used to capture the structural and dynamic determinants of allosteric signaling within the HNH domain. We reveal the existence of a millisecond timescale dynamic pathway that spans HNH from the region interfacing the adjacent RuvC nuclease and propagates up to the DNA recognition lobe in the full-length CRISPR-Cas9. These findings reveal a potential route of signal transduction within the CRISPR-Cas9 HNH nuclease, advancing our understanding of the allosteric pathway of activation. Further, considering the role of allosteric signaling in the specificity of CRISPR-Cas9, this work poses the mechanistic basis for novel engineering efforts aimed at improving its genome editing capability.

The CRISPR-Cas9 enzyme machine has exciting applications in genome editing and numerous investigations have sought to harness its mechanism for therapeutic bioengineering.[1-2] Cas9 is an RNA-guided DNA endonuclease, which generates double-stranded breaks in DNA by first recognizing its protospacer-adjacent motif (PAM) sequence and then cleaving the two DNA strands via the HNH and RuvC nuclease domains.[3] Structural studies of Cas9 have employed crystallographic[4-6] and cryo-EM[7-8] techniques, revealing several well-defined structural subdomains, including the catalytic domains, a recognition (REC) lobe and a PAM interacting (PI) region (Figure 1A). In parallel, Förster Resonance Energy Transfer (FRET) techniques provided insight into the large-scale conformational changes that occur during nucleic acid processing.[9-11] These and other biophysical studies have been invaluable to our current understanding of Cas9 function.[12-13] Building on this experimental information, computational investigations have been carried out to describe the conformational and dynamic requirements underlying Cas9 mechanistic action. All-atom Molecular Dynamics (MD) simulations have described the conformational activation of the Cas9 protein toward the binding and enzymatic processing of nucleic acids.[14-16] These investigations also revealed the ability of the Cas9 protein to propagate the DNA binding signal across the HNH and RuvC nuclease domains for concerted cleavage of the two DNA strands.[17] Notably, biochemical experiments and MD simulations have jointly indicated a dynamically driven allosteric signal throughout Cas9, where the intrinsic flexibility of the catalytic HNH domain regulates the conformational activation of both nucleases, therefore controlling the DNA cleavage activity [9, 17] Detailed knowledge of this allosteric mechanism and of the conformational control exerted by HNH is essential for understanding Cas9 function and for engineering efforts aimed at improving the specificity of this system through modulation of its allosteric signaling.[18] In this respect, an in-depth investigation necessitates the use of experimental techniques such as solution nuclear magnetic resonance (NMR) to quantify the motional timescales critical to this allosteric crosstalk. NMR can readily detect subtle conformational fluctuations at the molecular level, with precise information about the local dynamics on picosecond (ps) to nanosecond (ns) timescales (i.e. the so-called fast dynamics), as well as those occurring over microseconds (μs) to milliseconds (ms) (i.e. slow dynamics). These slow dynamics are of particular interest because several biological processes – including allosteric regulation – usually occur in this regime.[19] The power of solution NMR is magnified when coupled to MD simulations,[20-21] that capture protein fluctuations and conformations on the same timescales of NMR experiments, offering an interpretation at the atomic scale while also describing the subtle changes that characterize protein allostery.[22-24]

Here, we probe the structural and dynamic determinants of allosteric signaling in the Cas9 HNH nuclease by means of solution NMR and all-atom MD simulations. A novel construct of the HNH nuclease domain from S. pyogenes Cas9 has been determined through NMR and X-ray crystallography, which maintains the fold of the wild-type (WT, i.e. full-length) Cas9 protein and allows the characterization of its multi-timescale conformational dynamics by solution NMR spectroscopy and MD simulations. To comprehensively access the long timescale dynamics of the system at the atomic scale, accelerated MD simulations have been performed, employing a Gaussian accelerated MD (GaMD) method.[25] Accelerated MD is an enhanced sampling methodology that enables us to probe Cas9 dynamics over μs and ms timescales, in remarkable agreement with NMR experiments.[26-28] Thus, GaMD is ideal for the study of Cas9 motions that are relevant to its allosteric signaling.[19] As a result, we experimentally and theoretically identify a dynamic pathway that connects HNH and RuvC through contiguous ms timescale motions, while also highlighting its propagation to the REC lobe to enable the information transfer for concerted cleavage of the two DNA strands. Mixed machine learning and structure-based prediction tools of the NMR chemical shifts further reveal the agreement between experiments and computations, indicating that the structural/dynamic features derived from GaMD simulations represent the experimental results well at the molecular level. Overall, the integrated approach employed in this study enabled access to the intrinsic conformational fluctuations of the Cas9 HNH nuclease, which are essential for allosteric signaling in CRISPR-Cas9. Our combined NMR and theoretical approach paves the way for the complete mapping of its allosteric signaling and determination of its role in the enzymatic function and specificity.

Results

Structural features of the HNH nuclease

To determine the structural features of the isolated HNH domain, we employed solution NMR and X-ray crystallography. First, the structure of the HNH domain (Figure 1A) was derived from the $^1H^{15}N$ HSQC NMR spectrum (Figure 1B). Backbone assignments were uploaded to the CS23D server in order to predict the structure based on composite NMR chemical shift information.[29] Figure 1A shows (as a close-up view) the model of the HNH structure determined from NMR data using the CS23D server (green) overlaid with that of HNH from the full-length Cas9 (gray). The predicted structure reveals a remarkable overlap with the X-ray structure of the full-length Cas9 (PDB code: 4UN3)[5] displaying Cα root-mean-squared-deviation RMSD = 0.688 Å. The NMR model also highlights small helical turns in regions of poor electron density in the full-length Cas9 structure, as well as an extension of the C-terminal α-helix. The secondary structure of this construct determined from Cα and Cβ chemical shift indices is in good agreement with that of the HNH domain from the full-length Cas9 (Figure 1C), indicating that the engineered protein is a good representation of this fold in solution. Circular dichroism (CD) spectroscopy is consistent with a predominantly α-helical protein (Figure S1), in agreement with the X-ray structure of the full-length Cas9.[5]

The similarity of our construct to that of HNH from the full-length Cas9 supports the reliability of the predicted structure. A further confirmation is provided by the X-ray structure of the HNH construct solved at 1.9 Å resolution (Figure 2). This X-ray structure aligns well to that of the full-length Cas9 (PDB code 4UN3)[30] and the predicted NMR structure, with a Cα RMSD values of 0.549 Å and 0.479 Å, respectively, with the most significant difference due to a crystal contact in the experimental lattice pushing the N-terminal helix inward (Figure 2, inset top).

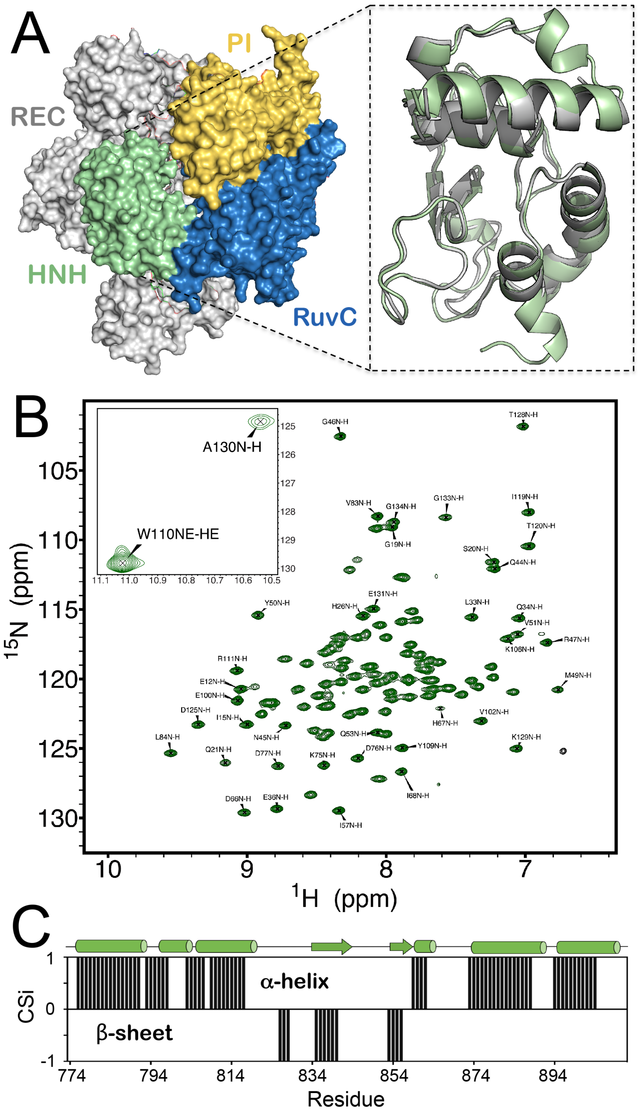


Figure 1. NMR Spectrum of HNH. (A) Architecture of the Cas9 protein (PDB code: 4UN3),[5] highlighting its protein domains as follows: HNH (green), RuvC (blue), PAM interacting region (PI, gold) and recognition lobe (REC, gray). In the close-up view, a model of the HNH structure determined from NMR data (green) is overlaid with that of HNH from the full-length Cas9 (gray). **(B)** $^1H^{15}N$ HSQC NMR spectrum of the HNH nuclease domain from *S. pyogenes* Cas9 (the inset reports two peaks out of range). **(C)** Consensus chemical shift index (CSi), indicating the predicted secondary structure for the HNH construct based on the NMR chemical shifts (black bars from 0 to 1 indicate α-helix, while bars from 0 to -1 indicate β-sheet, see the Methods section) compared to that of HNH from the full-length Cas9 (shown on top of the graph as sequence, with α-helical and β-sheet regions indicated as tubes and arrows).

The overall fold of HNH from full-length Cas9 is therefore well maintained in the isolated domain. The residues L791–E802 and T858–S872 form two flexible loop regions, as suggested by NMR. An α-helix is introduced in residues Q794–E798 and an additional solvent exposed loop comprised of residues T858–S872 forms a small α-helix at D861–R864 (Figure 2, inset bottom), also observed in the structural model from the NMR chemical shifts. Lastly, a small extension of the C-terminal α-helix is also confirmed in isolated HNH.

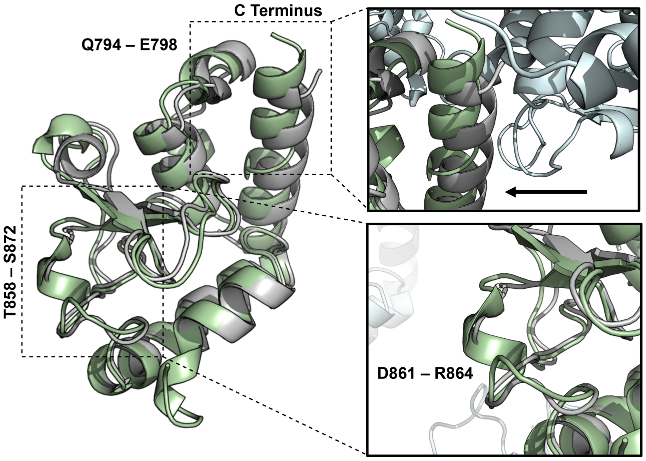


Figure 2. X-ray structure of HNH. The X-ray structure of the isolated HNH domain (PDB code: 6O56, green), solved at 1.90 Å resolution, is overlaid with the X-ray structure of the HNH domain from the full-length *S. pyogenes* Cas9 (PDB code: 4UN3, gray).[5]

Experimental dynamics of the HNH nuclease

Here, we analyzed the dynamics of HNH by means of the method of Bracken and coworkers.[31] The sites of *ps–ns* and *μs–ms* flexibility have been identified through the analysis of the R_1R_2 product. With respect to the individual longitudinal and transverse relaxation rates, the R_1R_2 product attenuates the contribution of motional anisotropy and more clearly illuminates sites of chemical exchange. As a result, the R_1R_2 values for each residue in HNH (Figure 3A and Table S1) highlight several locations of *ps–ns* and *μs–ms* flexibility. Twenty residues display R_1R_2 values above 1.5σ of the 10% trimmed mean, due to the significant influence of R_{ex} related to *μs–ms* motion. Measured R_{ex} parameters are consistent with this interpretation (Figures 3A, S3 and Table S1). A lower number of residues (i.e., 13) fall below 1.5σ of the mean, suggesting potential influence of *ps–ns* dynamics at these sites, with the mean R_1R_2 value corresponding to an average order parameter (S^2) value of 0.85, where $S_{av}^2 = \sqrt{\langle R_1R_2\rangle / R_1R_2^{max}}$. Steady-state ^{1}H-[^{15}N] NOE were also measured and the order parameter (S^2) was determined for assigned residues in HNH with RELAX.[32] Regions of *ps–ns* flexibility (*i.e.* high configurational entropy) are observed in residues 822–843 and 890–904. Consistent with these data, in the X-ray structure of full-length Cas9 residues 822–843 are exposed toward the solvent, while residues 890–904 comprise flexible loop regions.[5] Millisecond timescale dynamics of the HNH nuclease were quantified by Carr–Purcell–Meiboom–Gill (CPMG) relaxation dispersion experiments (Figure 3B and Table S2). Residues displaying slow timescale (*ms*) dynamics correspond to K782–E786, I788, K789, L791, Q794–E798, Y815, L816, N818, V824, E827–D829, I841, S851, D853, K855, E873 and L900 in the full-length Cas9. Rates of conformational exchange (k_{ex}) at these sites range from 800 – 2900 s^{-1} with an average k_{ex} = 1761 ± 414.

Allosteric signaling pathway

The slow dynamics identified via CPMG relaxation dispersion experiments are of particular interest for the identification of the signaling pathways, since the allosteric regulation is usually transmitted toward slow dynamical motions.[19] Residues displaying slow timescale (*ms*) dynamics (Table S2) form clusters in three regions of HNH (Figure 4A), two of which are the interface with the region REC2 of the recognition lobe and with RuvC (i.e., the HNH–REC2 and HNH–RuvC interfaces), while the third region is located in the core of HNH. This well-defined subset of flexible residues within HNH therefore bridges the RuvC and REC2 interfaces, forming a contiguous dynamic pathway within the isolated HNH domain. This pathway of flexible residues connecting HNH–RuvC and HNH–REC2 agrees well with the available experimental evidences that have indicated the existence of an allosteric communication within CRISPR-Cas9. Indeed, a tight dynamic inter-connection between HNH and RuvC has been originally reported by Sternberg and colleagues[9] and supported by MD simulations studies.[17] Moreover, the REC2 region has been recently suggested to be involved in the activation of HNH through an allosteric regulation that also implicates the REC3 region.[30] The authors have shown that, upon binding a complementary RNA:DNA structure prone to undergo DNA cleavage, the REC3 region modulates the motions of the neighboring REC2, which in turn contacts HNH and sterically regulates its access to the scissile phosphate. MD simulations of the fully activated CRISPR-Cas9 complex revealed that highly coupled motions between REC2, REC3 and HNH are critical for the activation of the catalytic domain toward cleavage, supporting the existence of an allosteric signal.[33] A recent experimental study has further suggested that REC2 is critical in regulating the rearrangements of the DNA for double strand cleavage via the HNH and RuvC nuclease domains.[34] Taken together, these findings strongly support the outcomes of the NMR experiments reported here, suggesting that the dynamic pathway spanning the isolated HNH domain is responsible for the information transfer between RuvC and REC2.

To gain insights into the allosteric signaling pathway within the full-length Cas9,[5] the latter has been object of extensive analysis by employing computational methods that are suited for the detection of allosteric effects.[35-38]

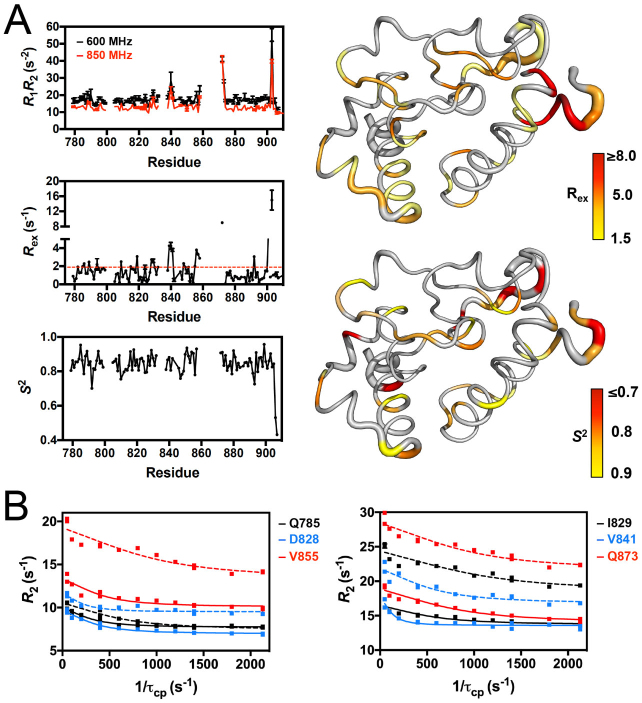


Figure 3. HNH Dynamics Measured by NMR. (A) Plots of R_1R_2, R_{ex} and the order parameter (S^2, determined from model-free analysis of R_1, R_2, ^{1}H-[^{15}N]-NOE measurements) for the HNH nuclease. The R_1R_2 parameters were measured at 600 (black) and 850 (red) MHz. On the plot of R_{ex}, the red dashed line denotes 1.5σ from the 10% trimmed mean of the data. On the right panel, the R_{ex} (top) and S^2 (bottom) values are mapped onto the HNH structure and colored according to the adjacent legends. **(B)** Selected CPMG relaxation dispersion curves collected at 600 (solid lines) and 850 (dashed lines) MHz.

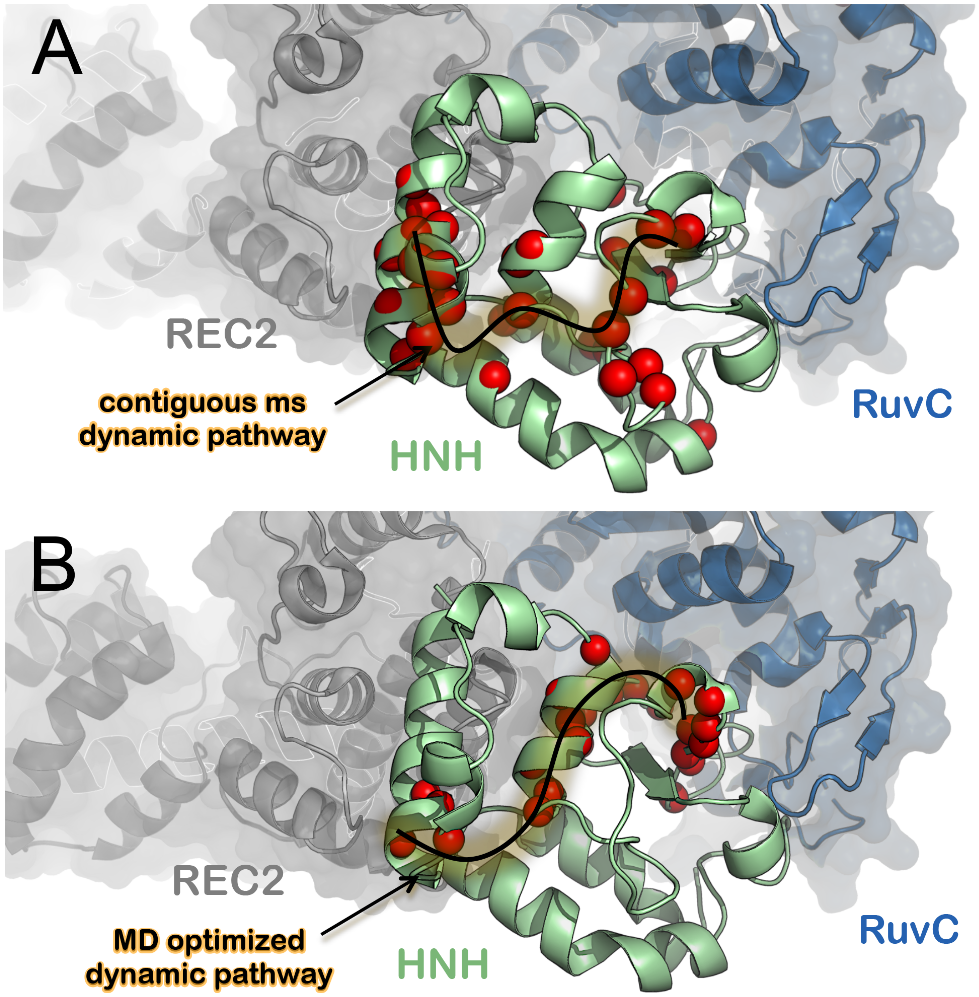


Figure 4. Allosteric signaling across HNH. (A) Flexible residues in the HNH construct measured by CPMG relaxation dispersion NMR. The majority of these sites highlight a contiguous *ms* dynamic pathway that spans the RuvC (blue) and REC (gray) domains when HNH (green) is placed into the full-length complex. **(B)** Allosteric pathway optimizing the overall correlation between HNH residues 789 and 841 (which are adjacent to RuvC and REC2 respectively), computed from correlation analyses and dynamical network models of the full-length Cas9. The theoretical pathway identifying the information transmission spanning HNH from the interface with the RuvC domain up to the REC2 region remarkably resembles the experimental pathway, derived in the isolated construct of HNH from CPMG experiments (panel A).

We combined correlation analyses and network models derived from graph theory to determine the most relevant pathways across HNH communicating RuvC with REC2. The computed pathways are composed by residue–to–residue steps that optimize the overall correlation (i.e., the momentum transport) between amino acids 789 and 841 (belonging to HNH but adjacent to RuvC and REC2, respectively). This yields an estimation of the principal channels of information transmission between RuvC and REC2. Interestingly, the pathway that maximizes the motion transmission between RuvC and REC2 through HNH (Figure 4B) agrees remarkably well with the pathway experimentally identified in the HNH construct via CPMG relaxation dispersion (Figure 4A). Residues belonging to the computational pathway are K789*, L791*, K810, L813, Y814, Y815*, K816*, Q817, N818*, G819*, D835, Y836, D837, V838, D839*, A840*, I841*, P843, D850*, S852, D853* and N854; where the asterisk indicates that they are also characterized by slow dynamics in the HNH construct (as experimentally identified via CPMG relaxation dispersion and R_1R_2 (+1.5σ), Tables S1-S2). This consensus between the dynamic pathways experimentally observed in the HNH construct and in the full-length Cas9 (as compared in Figures 4A and 4B, respectively) indicates that the REC2–HNH–RuvC communication channel is conserved in the full-length Cas9.

Conformational dynamics of HNH in the full-length Cas9

In order to compare the conformational dynamics of this novel HNH construct with those of the full-length Cas9, and to further interpret the outcomes of solution NMR experiments, we performed MD. All-atom MD simulations were conducted on the structure of the HNH domain predicted by NMR and of the X-ray structure of the full-length Cas9.[5] To access the long timescale dynamics of the systems, we performed accelerated MD simulations, using a Gaussian accelerated MD (GaMD) method,[25] which has shown to describe well the *μs* and *ms* dynamics of CRISPR-Cas9.[14, 39-40] Indeed, while classical MD can detect fast stochastic motions responsible for spin relaxation, more sophisticated methods that enhance the sampling of the configurational space are required to access the slower motion probed by solution NMR. Accelerated MD is a biased-potential method,[41] which adds a boost potential to the potential energy surface (PES), effectively decreasing the energy barriers separating low-energy states, thus accelerating the occurrence of slower dynamic events. As shown by several independent reports, the method accurately reproduces the slow dynamics captured by solution NMR in biomolecular systems,[26-28] therefore providing comparison with the experimental results reported here.

The simulated trajectories have been analyzed to compare the conformational dynamics of HNH in its isolated form and in the full-length Cas9. By performing Principal Component Analysis (PCA), the dynamics of HNH along the first principal mode of motion – usually referred as *"essential dynamics"*[42] – reveals remarkable similarities in the full-length Cas9 and in the isolated form (Figure S4). Interestingly, the residues of HNH that experimentally display *ms* dynamics (i.e., as captured from the CPMG relaxation dispersion and R_1R_2 (+1.5σ) measurements) are characterized by short amplitude motions in both the isolated form of HNH and when embedded in the full-length Cas9. Analysis of the root mean square fluctuations (RMSF) of individual Cα atoms further shows that the residues with slow timescale motions (experimentally identified via NMR) display low fluctuations in the simulations of both the isolated HNH and in the full-length Cas9 (Figure S6). This indicates that short amplitude motions and low fluctuations are conserved in the regions that form a continuous *ms* dynamic pathway connecting REC2–HNH–RuvC (Figure 4). In this respect, it is important to note that short amplitude motions, as well as low fluctuations, do not directly correspond to slow time scale dynamical motions. However, the consensus observed in both the HNH construct and within the full-length Cas9 indicates similar intrinsic dynamics along the pathway connecting REC2–HNH–RuvC, which has been experimentally derived via NMR (Figure 4A). Inspection of the conformational ensemble accessed during the simulations reveals that the isolated HNH domain resembles the ensemble of the full-length system overall, with a remarkable similarity in terms of short amplitude motions and low fluctuations for the residues within the REC2–HNH–RuvC pathway (Figure S7). Overall, the analysis of the conformational dynamics shows that the HNH construct maintains the fold observed in full-length Cas9, supporting the connection between conformational dynamics captured via solution NMR and those of HNH inside full-length Cas9.

Simulated ensemble and NMR experiments

To gain insight into how well the structural and dynamical features captured by GaMD simulations represent the NMR experiments at the molecular level, the simulated trajectories were used to compute NMR chemical shifts through ensemble machine learning and a mixed alignment/structure-based method with the SHIFTX2 code.[43] As a result, we detect qualitative agreement between predicted and experimental chemical shifts for the isolated HNH domain (Figure 5A, upper panel). Notably, the experimental distributions include side chain atoms, while the simulated spectra only consider backbone atoms. This results in a slightly broader distribution of the experimental ^{1}H shifts (while both experimental and predicted distributions are centered on the same value of ~8 ppm). Aside from these minor differences, the agreement between the simulated and the experimental chemical shift distribution plots for the isolated HNH domain is remarkable. This is a strong indication that the GaMD ensemble properly represents the NMR experiments at a molecular level. Another important aspect of these simulations is the similarity of HNH in full-length Cas9 and its isolated form. The lower panels of Figure 5A show that these forms of HNH display very similar ^{1}H and ^{15}N shift distributions, indicating that HNH presents similar spectral trends when it is isolated or in full-length Cas9. This is therefore a further indication that the structural dynamics of the HNH construct predicted by NMR are comparable to those of HNH in full-length Cas9, supporting the comparison performed here. Importantly, the observed agreement between the computed and experimental spectra is also observed in the simulation replicas (Figure S8).

Finally, to provide a comprehensive comparison of the molecular motions captured by NMR and sampled during the simulations, we analyzed the ^{15}N-^{1}H autorelaxation-derived order parameters S^2, which reflect the fluctuations of a backbone N-H bond vector due to its internal motion.

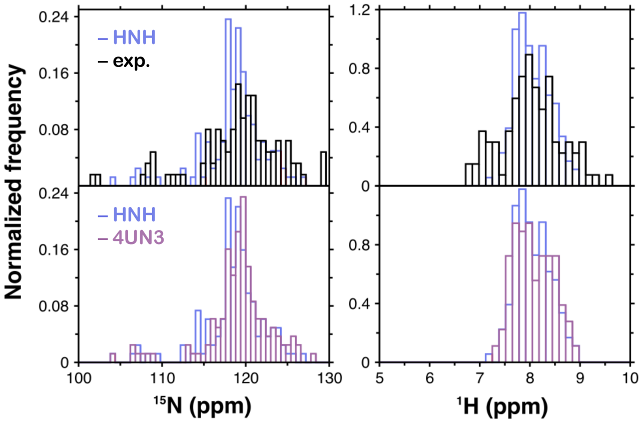


Figure 5. Experimental vs. simulated chemical shifts. Experimental and simulated ^{15}N and ^{1}H NMR chemical shifts of the HNH domain, plotted as normalized histograms. The upper panels compare the experimental (black line) and the simulated HNH (light blue) isolated domains. The lower panels compare the simulated HNH domain under two conditions: inside the Cas9 complex (purple) and in isolation (light blue). All simulated spectra were computed as described in Methods utilizing GaMD trajectories.

The orientational fluctuations of internuclear vectors sampled via MD can be compared to the experimentally measured S^2, through the Lipari-Szabo formalism. However, the distribution of the motions described by S^2 occurs on shorter time scales than the sampled GaMD trajectories. On the other hand, sampling of fast motions in different sub-states is challenging using classical MD, whose conformational range is likely to depend on the initial conformation. Hence, 40 equally distributed conformations have been extracted from the GaMD trajectories of the isolated HNH domain and of full-length Cas9 and subjected to independent classical MD simulation runs, enabling us to account for the increased statistical sampling of different sub-states explored by GaMD while gaining insight into the fast motions occurring within the diverse conformational states.[26-28] We find reasonable agreement between experimental and simulated order parameters (Figure S9), that highlights modest fast-timescale flexibility in HNH residues 791–796 and 846–852 as well as the N-/C-termini.

Discussion

The power of the CRISPR-Cas9 system is its ability to perform targeted genome editing *in vivo* with high efficiency and increasingly improved specificity.[30, 44-46] In this system, an intriguing allosteric communication has been suggested to propagate the DNA binding signal across the HNH and RuvC nuclease domains to facilitate their concerted cleavage of the two DNA strands.[9, 17] In this process, the intrinsic flexibility of the catalytic HNH domain would regulate the information transfer, exerting conformational control. Here, solution NMR experiments are used to capture the intrinsic motions responsible for the allosteric signaling across the HNH domain. We reveal the existence of a *ms* timescale dynamic pathway that spans the HNH domain from the region interfacing the RuvC domain and propagates up to the REC lobe at the level of the REC2 region (Figure 4A). In-depth analysis of the allosteric signaling within the full-length Cas9 has been performed, by employing theoretical approaches that are suited for the detection of allosteric effects.[35-38] As a result, the dynamic pathway experimentally observed in the HNH construct is conserved in the full-length Cas9 (Figure 4B), confirming the existence of a communication channel between REC2–HNH–RuvC. This continuous pathway confirms the direct communication between the two catalytic domains, originally identified by the experimental work of Sternberg[9] and supported by MD simulations,[17] and also discloses their connection to the REC2 region. In this respect, single molecule FRET experiments have indicated that REC2 is critical for the activation of HNH through an allosteric mechanism that also involves the REC3 region.[30] Accordingly, in the fully activated complex, the REC3 region would modulate the motions of the neighboring REC2, which in turn contacts HNH and regulates its access to the scissile phosphate. By doing so, the REC region would act as a *"sensor"* for the formation of a RNA:DNA structure prone to DNA cleavage, transferring the DNA binding information to the catalytic HNH domain in an allosteric manner. A tight dynamical interplay between REC2–REC3 and HNH has also been detected via MD simulations of the fully activated CRISPR-Cas9 complex, revealing that highly coupled motions of these regions are at the basis of the activation of HNH for DNA cleavage.[33] A recent important contribution further suggested that REC2 regulates the rearrangements of the DNA to attain double strand cleavages via the HNH and RuvC nucleases.[34] Altogether, these experimental outcomes strongly support the finding of a continuous dynamic pathway spanning HNH from RuvC to REC2, and suggest its functional role for the allosteric transmission. To further investigate the motions associated with allosteric signaling, the conformational dynamics of the HNH domain was investigated by means of accelerated MD simulations, which can probe long-timescale *μs* and *ms* motions in remarkable agreement with NMR experiments.[26-28] Analysis of these conformational dynamics indicates that the HNH construct maintains the overall fold observed in full-length Cas9, and indicates conserved short amplitude motions and low fluctuations in the regions that form a continuous *ms* dynamic pathway connecting REC2–HNH–RuvC. Taken together, these computational outcomes suggest that the intrinsic conformational dynamics experimentally identified in the HNH construct reasonably resemble the dynamics of HNH in the full complex, supporting the connection between the two systems. Finally, mixed machine learning and structure-based prediction of the NMR chemical shifts from the simulated trajectories have also revealed the agreement between experiments and computations, indicating that the structural/dynamic features derived via GaMD simulations represent the experimental results at the atomic and molecular level.

Overall, by combining solution NMR experiments and MD simulations, we identified the dynamic pathway for information transfer across the catalytic HNH domain of the CRISPR-Cas9 system. This pathway, which spans HNH from the RuvC nuclease interface up to the REC2 region in the full-length Cas9, is suggested to be critical for allosteric transmission, propagating the DNA binding signal across the recognition lobe and the nuclease domains (HNH and RuvC) for concerted cleavage of the two DNA strands. This study also represents the first step toward a complete mapping of the allosteric pathway in Cas9 through solution NMR experiments. In this respect, despite modern experimental practices such as perdeuteration,[47] transverse relaxation-optimized spectroscopy (TROSY),[48] sparse isotopic labeling,[49] and ^{15}N-detection,[50] the complete characterization of the slow dynamical motions responsible for the allosteric signaling has remained challenging, due to the size of the polypeptide chain of the Cas9 protein (~ 160 kDa). Future investigations – reliant upon ongoing experiments and computations in our research groups – will include the investigation of the information transfer between HNH and RuvC and the allosteric role of their flexible interconnecting loops.[9, 17] Further, our joint NMR/MD investigations are being employed to understand the role of the recognition region within the allosteric activation. This is of key importance, since mutations within the REC lobe – at distal sites with respect to HNH – can control the activation of HNH and the specificity of the enzyme toward on-target DNA sequences.[30, 44-46] As such, by providing fundamental understanding of the intrinsic allosteric signaling within the catalytic HNH domain, the present study poses the basis for the complete mapping of the allosteric pathway in Cas9 and its role in the on-target specificity, helping engineering efforts aimed at improving the genome editing capability of the Cas9 enzyme.

Materials and Methods

Protein Expression and Purification. The HNH domain of *S. pyogenes* Cas9 (residues 775-908) was engineered into a pET15b vector with an N-terminal His$_6$-tag and expressed in Rosetta(DE3) cells in M9 minimal medium containing MEM vitamins, $MgSO_4$ and $CaCl_2$. Cells were induced with 0.5 mM IPTG after reaching an OD_{600} of 0.8 – 1.0 and grown for 16 – 18 hours at 22 °C post induction. The cells were harvested by centrifugation, resuspended in a buffer containing 20 mM HEPES, 500 mM KCl, and 5 mM imidazole at pH 8.0, lysed by ultrasonication and purified on a Ni-NTA column. NMR samples were dialyzed into a buffer containing 20 mM HEPES, 80 mM KCl, 1 mM DTT and 7.5% (v/v) D_2O at pH 7.4.

X-ray Crystallography. Following TEV cleavage of the His$_6$-tag, HNH was subsequently purified by HiPrep 16/60 Sephacryl 100 S-100 HR gel filtration chromatography. Crystals were obtained with sitting drop vapor diffusion at room temperature with 48 mg/mL HNH 1:1 with the Molecular Dimensions Morpheus I Screen condition E4 (0.1 M mixture of [imidazole and MES] pH 6.5, 25% (v/v) mixture of [2-methyl-2,4-pentanediol, PEG1000, and PEG3350], and 0.3 M mixture of [diethylene glycol, triethylene glycol, tetraethylene glycol, and pentaethlyene glycol]). Diffraction data were collected on a Rigaku MicroMax-003i sealed tube X-ray generator with a Saturn 944 HG CCD detector and processed and scaled using XDS[51] and Aimless in the CCP4 program suite.[52] The HNH domain from full-length *S. pyogenes* Cas9 was used for molecular replacement (PDB: 4UN3)[5] with Phaser in the PHENIX software package.[53] Iterative rounds of manual building in Coot[54] and refinement in PHENIX yielded the final HNH domain structure.

NMR Spectroscopy. NMR spin relaxation experiments were carried out at 600 and 850 MHz on Bruker Avance NEO and Avance III HD spectrometers, respectively. All NMR spectra were processed with NMRPipe [55] and analyzed in SPARKY.[56] Backbone chemical shift data was uploaded to the CS23D server for secondary structure calculations. Carr-Purcell-Meiboom-Gill (CPMG) NMR experiments were adapted from the report of Palmer and coworkers,[57] and performed at 25 °C with a constant relaxation period of 40 ms, a 2.0 second recycle delay, and τ_{cp} points of 0.555, 0.625, 0.714, 0.833, 1.0, 1.25, 1.5, 1.667, 2.5, 5, 10, and 20 ms. Relaxation dispersion profiles were generated by plotting R_2 vs. $1/\tau_{cp}$ and exchange parameters were obtained from fits of these data carried out with in-house scripts and in RELAX under the R2eff, NoRex, Tollinger (TSMFK01), and Carver-Richards (CR72 and CR72-Full) models.[32, 58] Two-field relaxation dispersion data were fit simultaneously and uncertainty values were obtained from replicate spectra (see the Supporting Information, SI). Longitudinal and transverse relaxation rates were measured with relaxation times of 0(x2), 40, 80, 120, 160(x2), 200, 240, 280(x2), 320, 360, and 400 ms for R_1 and 4.18, 8.36(x2), 12.54, 16.72, 20.9(x2), 25.08(x2), 29.26, 33.44, 37.62, and 41.8 ms for R_2. Peak intensities were quantified in Sparky and the resulting decay profiles were analyzed in Mathematica with errors determined from the fitted parameters. Steady-state ^{1}H-[^{15}N] NOE were measured with a 6 second relaxation delay followed by a 3 second saturation (delay) for the saturated (unsaturated) experiments. All relaxation experiments were carried out in a temperature-compensated interleaved manner. Model-free analysis using the Lipari-Szabo formalism was carried out on dual-field NMR data in RELAX with fully automated protocols.[32]

Computational Structural Models. Two model systems were built for MD simulations, the first of which was based on the X-ray structure of the full-length wild-type Cas9 protein in complex with RNA and DNA, solved at 2.58 Å resolution (PDB code: 4UN3).[5] The second model system was based on the NMR structure of the HNH domain obtained in this work. The RMSD between the HNH domain in the X-ray structure of the WT Cas9 complex[5] and the HNH domain structure determined here is 0.688 Å. Both model systems were embedded in explicit water, adding Na+ counter-ions to neutralize the total charge, reaching a total of ~220,000 atoms and a box size of ~145 x 110 x 147 Å^3 for the CRISPR-Cas9 complex and ~25,000 atoms and a box size of ~72 x 62 x 60 Å^3 for the HNH domain.

Molecular Dynamics (MD) Simulations. The above-mentioned model systems were equilibrated through conventional MD. We employed the Amber ff12SB force field, which includes the ff99bsc0 corrections for DNA[59] and the ff99bsc0+χ*OL3* corrections for RNA.[60-61] Hydrogen atoms were added assuming standard bond lengths and constrained to their equilibrium position with the SHAKE algorithm. Temperature control (300 K) was performed via Langevin dynamics,[62] with a collision frequency γ = 1. Pressure control was accomplished by coupling the system to a Berendsen barostat,[63] at a reference pressure of 1 atm and a relaxation time of 2 ps. All simulations have been carried out through a well-established protocol described in the SI. MD simulations were carried out in the NVT ensemble, collecting ~100 ns for each system (for a total of ~400 ns of production runs). These well-equilibrated systems have been used as the starting point for Gaussian accelerated MD (GaMD, details below). Classical MD simulations have also been performed on 40 equally distributed conformations extracted from the GaMD trajectories of the isolated HNH domain and of the full-length Cas9. Specifically, ~10 ns of MD have been collected for each of the 40 configurations of the isolated HNH domain (for a total of ~400 ns), while ~30 ns of MD have been carried out for each of the 40 configurations of the full-length Cas9 (for a total of ~1,2 μs), in agreement with the decay time detected experimentally. All simulations were performed with the GPU version of AMBER 16.[64]

Gaussian Accelerated MD Simulations (GaMD). Accelerated MD (aMD) is an enhanced sampling method that adds a boost potential to the Potential Energy Surface (PES), effectively decreasing the energy barriers and accelerating transitions between low-energy states.[41] The method extends the capability of MD simulations over long timescales, capturing slow μs and ms motions in excellent comparability with solution NMR experiments.[26-28] Here, we applied a novel and robust aMD method, namely a Gaussian aMD (GaMD),[25] which uses harmonic functions to construct a boost potential that is adaptively added to the PES, enabling unconstrained enhanced sampling and simultaneous reweighting of the canonical ensemble.

Considering a system with N atoms at positions $\vec{r} = \{\vec{r_1}, \ldots \vec{r_N}\}$, when the system potential $V(\vec{r})$ is lower than a threshold energy E, the energy surface is modified by 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.* $V_1^*(\vec{r}) < V_2^*(\vec{r})$. Second, if $V_1(\vec{r}) < V_2(\vec{r})$, the potential difference observed on the smoothed 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 with Eqn [1] and [2]:

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

where V_{min} and V_{max} are the system minimum and maximum potential energies. To ensure that Eqn. [4] is valid, k must satisfy $k \leq 1/V_{max} - V_{min}$. By defining $k \equiv k_0\, 1/V_{max} - V_{min}$, then $0 < k \leq 1$. Lastly, the standard deviation of ΔV must be narrow enough 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 gies, $\sigma_{\Delta V}$ 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, $E = V_{min}$, according to Eqn. [4], k_0 can be calculated as:

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

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

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

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.[25] GaMD yields a canonical average of an ensemble by reweighting each point in the configuration space on the modified potential by the strength of the Boltzmann factor of the bias energy, $exp\ [\beta \Delta V(r_{t(i)})]$ at that particular point.

Based on extensive tests on the CRISPR-Cas9 system,[14, 39-40] the system threshold energy is $E = V_{max}$ for all GaMD simulations. The boost potential was 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 was used. The maximum, minimum, average, and standard deviation values of the system potential (V_{max}, V_{min}, V_{avg} and σ_V) were obtained from an initial ~12 ns NPT simulation with no boost potential. GaMD simulations were applied to the CRISPR-Cas9 complex and our HNH domain construct. Each GaMD simulation proceeded with a ~50 ns run, in which the boost potential was updated every 1.6 ns, thus reaching equilibrium. Finally, ~400 ns of GaMD simulations were carried in the NVT *ensemble* out for each system in two replicas, for a total of ~1.6 μs of GaMD. This simulation length (i.e., ~400 ns per replica) has shown to exhaustively explore the conformational space of the CRISPR-Cas9 system.[14, 39]

Determination of the Allosteric Pathways across the HNH domain. The allosteric pathway for information transfer has been investigated by employing correlation analysis and graph theory.[35-38] First, the generalized correlations (GC*ij*), which capture non-collinear correlations between pairs of residues *i* and *j*, are computed.[65] In this correlation analysis, two variables (x_i,x_j) can be considered correlated when their joint probability distribution, $p(x_i, x_j)$, is smaller than the product of their marginal distributions, $p(x_i) \cdot p(x_j)$. The mutual information (MI) is a measure of the degree of correlation between x_i and x_j defined as function of $p(x_i, x_j)$ and $p(x_i) \cdot p(x_j)$ according to:

$$MI\,[x_i, x_j] = \iint p(x_i, x_j)\, ln \frac{p(x_i, x_j)}{p(x_i) \cdot p(x_j)} dx_i dx_j \tag{6}$$

Notably, MI is related to the definition of the Shannon entropy, $H[x]$, i.e., the expectation value of a random variable x, having a probability distribution $p(x_i)$

$$H[x] = \int p(x)\, ln\, p(x) dx \tag{7}$$

and it can be thus computed as:

$$MI\,[x_i, x_j] = H\,[x_i] + H\,[x_j] - H\,[x_i, x_j] \tag{8}$$

where $H\,[x_i]$ and $H\,[x_j]$ are the marginal Shannon entropies, and $H\,[x_i, x_j]$ is the joint entropy. Since MI varies from 0 to + ∞, normalized generalized correlation coefficients (GC_{ij}), ranging from 0 (independent variables) to 1 (fully correlated variables), are defined as:

$$GC_{ij}\,[x_i, x_j] = \left\{1 - e^{-2MI[x_i, x_j]/d}\right\}^{-1/2} \tag{9}$$

where d=3 is the dimensionality of x_i and x_j. GC*ij* have been computed using have been computed using a code developed within our group, utilizing the MI defined by Lange.[65] In a second phase, the GC*ij* are used as a metric to build a dynamical network model of the protein.[37] In this model, the protein amino acids residues constitute the nodes of the dynamical network graph, connected by edges (residue pair connection). Edge lengths, i.e., the inter-node distances in the graph, are defined using the GC_{ij} coefficients according to:

$$w_{ij} = -\log GC_{ij} \tag{10}$$

In the present work, two nodes have been considered connected if any heavy atom of the two residues is within 5 A° of each other (i.e., *distance cutoff*) for at least the 70 % of the simulation time (i.e., *frame cutoff*). This leads to the definition of a set of elements w_{ij} of the graph. In the third phase of the protocol, the optimal pathways for the information transfer between two nodes (i.e., two amino acids) are defined using the Dijkstra algorithm, which finds the roads, composed by inter-node connections, that minimize the total distance (and therefore maximize the correlation) between amino acids. In the present study, this protocol was applied on the trajectories of the full-length Cas9 simulated for ~400 ns of GaMD simulations and averaged over two replicas. The Dijkstra algorithm was applied between the amino acids 789 and 841, which belong to HNH and are located at the interface with RuvC and REC2, respectively. As a result, the routes that maximize the correlation between amino acids 789 and 841 are identified, providing residue–to–residue pathways that optimize the correlations (i.e., the momentum transport). With the optimal motion transmission pathway the following 50 *sub-optimal* information channels where computed and accumulated and plotted on the 3D structure (Figure 4B), to account for the contribution of the most likely sub-optimal pathways.

through a well-established protocol described in the SI. MD simulations were carried out in the NVT ensemble, collecting ~100 ns for each system (for a total of ~400 ns of production runs). These well-equilibrated systems have been used as the starting point for Gaussian accelerated MD (GaMD, details below). Classical MD simulations have also been performed on 40 equally distributed conformations extracted from the GaMD trajectories of the isolated HNH domain and of the full-length Cas9. Specifically, ~10 ns of MD have been collected for each of the 40 configurations of the isolated HNH domain (for a total of ~400 ns), while ~30 ns of MD have been carried out for each of the 40 configurations of the full-length Cas9 (for a total of ~1,2 μs), in agreement with the decay time detected experimentally. All simulations were performed with the GPU version of AMBER 16.[64]

Gaussian Accelerated MD Simulations (GaMD). Accelerated MD (aMD) is an enhanced sampling method that adds a boost potential to the Potential Energy Surface (PES), effectively decreasing the energy barriers and accelerating transitions between low-energy states.[41] The method extends the capability of MD simulations over long timescales, capturing slow *μs* and *ms* motions in excellent comparability with solution NMR experiments.[26-28] Here, we applied a novel and robust aMD method, namely a Gaussian aMD (GaMD),[25] which uses harmonic functions to construct a boost potential that is adaptively added to the PES, enabling unconstrained enhanced sampling and simultaneous reweighting of the canonical ensemble.

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

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

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

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.* $V_1^*(\vec{r}) < V_2^*(\vec{r})$. Second, if $V_1(\vec{r}) < V_2(\vec{r})$, the potential difference observed on the smoothed 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 with Eqn [1] and [2]:

$$V_{max} \leq E \leq V_{min} + 1/k, \qquad [3]$$

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

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

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

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

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.[25] GaMD yields a canonical average of an ensemble by reweighting each point in the configuration space on the modified potential by the strength of the Boltzmann factor of the bias energy, $exp\ [\beta \Delta V(r_{t(i)})]$ at that particular point.

Based on extensive tests on the CRISPR-Cas9 system,[14, 39-40] the system threshold energy is $E = V_{max}$ for all GaMD simulations. The boost potential was 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 was used. The maximum, minimum, average, and standard deviation values of the system potential (V_{max}, V_{min}, V_{avg} and σ_V) were obtained from an initial ~12 ns NPT simulation with no boost potential. GaMD simulations were applied to the CRISPR-Cas9 complex and our HNH domain construct. Each GaMD simulation proceeded with a ~50 ns run, in which the boost potential was updated every 1.6 ns, thus reaching equilibrium. Finally, ~400 ns of GaMD simulations were carried in the NVT *ensemble* out for each system in two replicas, for a total of ~1.6 μs of GaMD. This simulation length (i.e., ~400 ns per replica) has shown to exhaustively explore the conformational space of the CRISPR-Cas9 system.[14, 39]

Determination of the Allosteric Pathways. The allosteric pathway for information transfer has been investigated by employing correlation analysis and graph theory.[35-38] First, the generalized correlations (GC*ij*), which capture non-collinear correlations between pairs of residues *i* and *j*, are computed.[65] In this correlation analysis, two variables (x_i,x_j) can be considered correlated when their joint probability distribution, $p(x_i, x_j)$, is smaller than the product of their marginal distributions, $p(x_i) \cdot p(x_j)$. The mutual information (MI) is a measure of the degree of correlation between x_i and x_j defined as function of $p(x_i, x_j)$ and $p(x_i) \cdot p(x_j)$ according to:

$$MI\ [x_i, x_j] = \iint p(x_i, x_j)\, ln \frac{p(x_i, x_j)}{p(x_i) \cdot p(x_j)} dx_i dx_j \qquad [6]$$

Notably, *MI* is related to the definition of the Shannon entropy, $H[x]$, i.e., the expectation value of a random variable x, having a probability distribution $p(x_i)$

$$H[x] = \int p(x)\, ln\, p(x) dx \qquad [7]$$

and it can be thus computed as:

$$MI\ [x_i, x_j] = H\ [x_i] + H\ [x_j] - H\ [x_i, x_j] \qquad [8]$$

where $H\ [x_i]$ and $H\ [x_j]$ are the marginal Shannon entropies, and $H\ [x_i, x_j]$ is the joint entropy. Since MI varies from 0 to + ∞, normalized generalized correlation coefficients (GC_{ij}), ranging from 0 (independent variables) to 1 (fully correlated variables), are defined as:

$$GC_{ij}\ [x_i, x_j] = \{1 - e^{-2MI[x_i, x_j]/d}\}^{-1/2} \qquad [9]$$

where d=3 is the dimensionality of x_i and x_j. GC*ij* have been computed using have been computed using a code developed within our group, utilizing the *MI* defined by Lange.[65] In a second phase, the GC*ij* are used as a metric to build a dynamical network model of the protein.[37] In this model, the protein amino acids residues constitute the nodes of the dynamical network graph, connected by edges (residue pair connection). Edge lengths, i.e., the inter-node distances in the graph, are defined using the *GCij* coefficients according to:

$$w_{ij} = -\log GC_{ij} \qquad [10]$$

In the present work, two nodes have been considered connected if any heavy atom of the two residues is within 5 Å of each other (i.e., *distance cutoff*) for at least the 70 % of the simulation time (i.e., *frame cutoff*). This leads to the definition of a set of elements w_{ij} of the graph. In the third phase of the protocol, the optimal pathways for the information transfer between two nodes (i.e., two amino acids) are defined using the Dijkstra algorithm, which finds the roads, composed by inter-node connections, that minimize the total distance (and therefore maximize the correlation) between amino acids. In the present study, this protocol was applied on the trajectories of the full-length Cas9 simulated for ~400 ns of GaMD simulations and averaged over two replicas. The Dijkstra algorithm was applied between the amino acids 789 and 841, which belong to HNH and are located at the interface with RuvC and REC2, respectively. As a result, the routes that maximize the correlation between amino acids 789 and 841 are identified, providing residue–to–residue pathways that optimize the correlations (i.e., the momentum transport). With the optimal motion transmission pathway, the following 50 *sub-optimal* information channels where computed and accumulated and plotted on the 3D structure (Figure 4B), to account for the contribution of the most likely sub-optimal pathways.

Corresponding Authors

* George P. Lisi (george_lisi@brown.edu)
* Giulia Palermo (giulia.palermo@ucr.edu)

Present Address

Atanu Acharya: School of Physics, Georgia Institute of Technology, Atlanta, GA 30332, United States.

Acknowledgments

We thank Prof. Martin Jinek for the WT Cas9 plasmid and Prof. Patrick Loria for helpful discussion of spin relaxation. We thank Prof. Luca Mollica for useful discussions.

Funding

This work was supported by start-up funds from Brown University and funds from the COBRE Center for Computational Biology of Human Disease (P20GM109035) to GPL, and by start-up funds from the University of California, Riverside to GP. Computer time for MD has been awarded by XSEDE via grant TG-MCB160059 (to GP).

References

1. Hsu, P. D.; Lander, E. S.; Zhang, F. *Cell* **2014**, *1576*, 1262-1278.
2. Doudna, J. A.; Charpentier, E. *Science* **2014**, *346*, 1258096.
3. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; Charpentier, E. *Science* **2012**, *337*, 816-821.
4. 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. *Science* **2014**, *343*, 12479971-11.
5. Anders, C.; Niewoehner, O.; Duerst, A.; Jinek, M. *Nature* **2014**, *513*, 569-573.
6. Nishimasu, H.; Ran, F. A.; Hsu, P. D.; Konemann, S.; Shehata, S. I.; Dohmae, N.; Ishitani, R.; Zhang, F.; Nureki, O. *Cell* **2014**, *156*, 935-949.
7. Jiang, F. G.; Taylor, D. W.; Chen, J. S.; Kornfeld, J. E.; Zhou, K. H.; Thompson, A. J.; Nogales, E.; Doudna, J. A. *Science* **2016**, *351*, 867-871.
8. Huai, G.; Li, G.; Yao, R.; Zhang, Y.; Cao, M.; Kong, L.; Jia, C.; Yuan, H.; Chen, H.; Lu, D.; Huang, Q. *Nat Commun* **2017**, *8*, 9.
9. Sternberg, S. H.; LaFrance, B.; Kaplan, M.; Doudna, J. A. *Nature* **2015**, *527*, 110-113.
10. Dagdas, Y. S.; Chen, J. S.; Sternberg, S. H.; Doudna, J. A. *Sci Adv* **2017**, *3*, eaao002.
11. Osuka, S.; Isomura, K.; Kajimoto, S.; Komori, T.; Nishimasu, H.; Shima, T.; Nureki, O.; Uemura, S. *EMBO J* **2018**, *37*, e96941.
12. Raper, A. T.; Stephenson, A. A.; Suo, Z. *J Am Chem Soc* **2018**, *140*, 2971-2984.
13. Shibata, M.; Nishimasu, H.; Kodera, N.; Hirano, S.; Ando, T.; Uchihashi, T.; Nureki, O. *Nat Commun* **2017**, *8*, 1430.
14. Palermo, G.; Miao, Y.; Walker, R. C.; Jinek, M.; McCammon, J. A. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 7260-7265.
15. Zuo, Z.; Liu, J. *Sci Rep* **2017**, *7*, 17271.
16. Palermo, G.; Miao, Y.; Walker, R. C.; Jinek, M.; McCammon, J. A. *ACS Cent Sci* **2016**, *2*, 756-763.
17. Palermo, G.; Ricci, C. G.; Fernando, A.; Basak, R.; Jinek, M.; Rivalta, I.; Batista, V. S.; McCammon, J. A. *J Am Chem Soc* **2017**, *139*, 16028-16031.
18. Chen, J. S.; Doudna, J. A. *Nat Rev Chem* **2017**, *1*, 78.
19. Kern, D.; Zuiderweg, E. R. *Curr Opin Struct Biol* **2003**, *13*, 748-757.
20. Lisi, G. P.; Loria, J. P. *Chem Rev* **2016**, *116*, 6323-6369.
21. Nussinov, R., Introduction to Protein Ensembles and Allostery. *Chem Rev* **2016**, *116*, 6263–6266.
22. Wodak, S. J.; Paci, E.; Dokholyan, N. V.; Berezovsky, I. N.; Horovitz, A.; Li, J.; Hilser, V. J.; Bahar, I.; Karanicolas, J.; Stock, G.; Hamm, P.; Stote, R. H.; Eberhardt, J.; Chebaro, J.; Dejaegere, A.; Cecchini, M.; Changeux, J. P.; Bolhuis, P. J.; Vreede, J.; Faccioli, P.; Orioli, S.; Ravasio, R.; Yan, L.; Brito, C.; Wyart, M.; Gkeka, P.; Rivalta, I.; Palermo, G.; McCammon, J. A.; Panecka-Hofman, J.; Wade, R. C.; Di Pizio, A.; Niv, M. Y.; Nussinov, R.; Tsai, C. J.; Jang, H.; Padhorny, D.; Kozakov, D.; McLeish, T., Allostery in its many disguises: from theory to applications. *Structure* **2019**, *27*, 566-578.
23. Guo, J.; Zhou, H. X. *Chem Rev* **2016**, *116*, 6503-6515.
24. Dokholyan, N. V. *Chem Rev* **2016**, *116*, 6463-6487.
25. Miao, Y.; Feher, V. A.; McCammon, J. A. *J Chem Theory Comput* **2015**, *11*, 3584-3595.
26. Markwick, P. R. L.; Bouvignies, G.; Blackledge, M. *J Am Chem Soc* **2007**, *129*, 4724-4730.
27. Mukrasch, M. D.; Markwick, P.; Biernat, J.; von Bergen, M.; Bernado, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J Am Chem Soc* **2007**, *129*, 5235-5243.
28. Salmon, L.; Pierce, L.; Grimm, A.; Ortega, R., J.-L.; Mollica, L.; Jensen, M. R.; van Nuland, N.; Markwick, P.; McCammon, J. A.; Blackledge, M. *Angew Chem Int Ed* **2012**, *51*, 6103-6106.
29. Wishart, D. S.; Arndt, D.; Berjanskii, M.; Tang, P.; Zhou, J.; Lin, G. *Nucleic Acids Res* **2008**, *36*, 496-502.
30. Chen, J. S.; Dagdas, Y. S.; Kleinstiver, B. P.; Welch, M. M.; Harrington, L. B.; Sternberg, S. H.; Joung, J. K.; Yildiz, A.; Doudna, J. A. *Nature* **2017**, *550*, 407–410.
31. Kneller, J. M.; Bracken, C. *J Am Chem Soc* **2002**, *124*, 1852-1853.
32. Bieri, M.; d'Auvergne, E. J.; Gooley, P. R. *J Biomol NMR* **2011**, *50*, 147-155.
33. Palermo, G.; Chen, J. S.; Ricci, C. G.; Rivalta, I.; Jinek, M.; Batista, V. S.; Doudna, J. A.; McCammon, J. A. *Q. Rev. Biophys.* **2018**, *51*, 1-11.
34. Sung, K.; Park, J.; Kim, J.; Lee, N. K.; Kim, S. K. *J Am Chem Soc* **2018**, *140*, 7778-7781.
35. Rivalta, I.; Sultan, M. M.; Lee, N. S.; Manley, G. A.; Loria, J. P.; Batista, V. S. *Proc Natl Acad Sci U S A* **2012**, *109*, E1428-36.
36. Negre, C. F. A.; Hendrickson, H.; Rhitankar Pal, R.; Rivalta, I.; Ho, J.; Batista, V. S. *Proc Natl Acad Sci U S A* **2018**, *115*, 12201-12208.
37. Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. *Proc Natl Acad Sci U S A* **2009**, *106*, 6620-6625.
38. Lisi, G. P.; East, K. W.; Batista, V. S.; Loria, J. P. *Proc Natl Acad Sci U S A* a **2017**, *114*, E3414-E3423.
39. Palermo, G., Structure and Dynamics of the CRISPR-Cas9 catalytic complex. *J. Chem. Inf. Model.* **2019**, *59*, 2394-2406.
40. Ricci, C. G.; Chen, J. S.; Miao, Y.; Jinek, M.; Doudna, J. A.; McCammon, J. A.; Palermo, G. *ACS Cent Sci* **2019**, *5*, 651-662.
41. Hamelberg, D.; Mongan, J.; McCammon, J. A. *J Chem Phys* **2004**, *120*, 11919-11929.
42. Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412-425.
43. Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. *J Biomol NMR* **2001**, *50*, 43–57.
44. Kleinstiver, B. P.; Pattanayak, V.; Prew, M. S.; Tsai, S. Q.; Nguyen, N. T.; Zheng, Z. L.; Joung, J. K. *Nature* **2016**, *529*, 490-495.
45. Casini, A.; Olivieri, M.; Petris, G.; Montagna, C.; Reginato, G.; Maule, G.; Lorenzin, F.; Prandi, D.; Romanel, A.; Demichelis, F.; Inga, A.; Cereseto, A. *Nat Biotechnol* **2018**, *36*, 265–271.
46. Slaymaker, I. M.; Gao, L.; Zetsche, B.; Scott, D. A.; Yan, W. X.; Zhang, F. *Science* **2016**, *351*, 84-88.
47. Venters, R. A.; Farmer, B. T. I.; Fierke, C. A.; Spicer, L. D. *J Mol Biol* **1996**, *264*, 1101-1106.
48. Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K. *Proc Natl Acad Sci U S A* **1997**, *94*, 12366-12371.

49. Tugarinov, V.; Kanelis, V.; Kay, L. E. *Nat Protoc* **2006**, *1*, 749-754.

50. Takeuchi, K.; Arthanari, H.; Imai, M.; Wagner, G.; Shimada, I. *J Biomol NMR* **2016**, *64*, 143-151.

51. Kabsch, W. *Acta Cryst Sect D* **2010**, *66*, 125-132.

52. Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. *Acta Cryst Sect D* **2011**, *67*, 235-242.

53. Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. *Acta Cryst Sect D* **2010**, 213-221.

54. Emsley, P.; Lohkamp, B.; Scott, W.G.; Cowtan, K. D. *Acta Cryst Sect D* **2010**, *66*, 486-501.

55. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277-293.

56. Goddard, T. D.; Kneller, D. G. **2008**, *University of California, San Francisco*.

57. Loria, J. P.; Rance, M.; Palmer, A. G., 3rd. *J Am Chem Soc* **1999**, *121* , 2331-2332.

58. Morin, S.; Linnet, T.; Lescanne, M.; Schanda, P.; Thompson, G. S.; Tollinger, M.; Teilum, K.; Gagne, S.; Marion, D.; Griesinger, C.; Blackledge, M.; d'Auvergne, E. J. *Bioinformatics* **2014**, *30*, 2219-2220.

59. Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E. r.; Laughton, C. A.; Orozco, M. *Biophys J* **2007**, *92*, 3817-3829.

60. Banas, P.; Hollas, D.; Zgarbova, M.; Jurecka, P.; Orozco, M.; Cheatham, T. E. r.; Sponer, J. *J Chem Theory Comput* **2010**, *6*, 3836-3849.

61. Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E.; Jurecka, P. *J Chem Theory Comput* **2011**, *7*, 2886-2902.

62. Turq, P.; Lantelme, F.; Friedman, H. L. *J Chem Phys* **1977**, *66*, 3039.

63. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J Chem Phys* **1984**, *81*, 3684.

64. Case, D. A.; Betz, R. M.; Botello-Smith, W.; Cerutti, D. S.; Cheatham, I., T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; M., Y. D.; A., K. P., AMBER 2016. *University of California, San Francisco* **2016**.

65. Lange, O. F.; Grubmuller, H. *Proteins: Struct Funct Bioinf* **2006**, *62*, 1053-1061.