Unveiling Cas8 Dynamics and Regulation within a transposon-encoded Cascade-TniQ Complex

Amun Patel[1,2], Souvik Sinha[2], Pablo Arantes[2], Giulia Palermo[2,3]*
[1]Biophysics Program, [2]Department of Bioengineering, [3]Department of Chemistry, The University of California, Riverside

Abstract: Cascade is a class 1, type 1 CRISPR-Cas system with a variety of roles in prokaryote defense, specifically against DNA-based viruses. The *Vibrio Cholerae* transposon, Tn6677, encodes a variant of the type 1F Cascade known as type 1F-3. This Cascade variant complexes with a homodimer of the transposition protein TniQ and leverages the sequence specificity of Cascade to direct the integration activity of the heteromeric transposase tnsA/B, resulting in site-specific transposition of Tn6677. We desire to uncover the molecular details behind R Loop formation of 'Cascade-TniQ.' Due to the lack of a complete model of Cascade-TniQ available at atom-level resolution, we first build a complete model using AlphaFold V2.1. We then simulate this model via classical molecular dynamics and umbrella sampling to study an important regulatory component within Cascade-TniQ, known as the Cas8 'bundle.' Particularly, we show that this alpha helical bundle experiences a free energy barrier to its large-scale translatory motions and relative free energies of its states primarily dependent on a loop within a Cas7 subunit in Cascade-TniQ. Further, we comment on additional structural and dynamical regulatory points of Cascade-TniQ during R Loop formation, such as Cascade-TniQ backbone rigidity, and the potential role TniQ plays in regulating bundle dynamics. In summary, our outcomes provide the first all-atom dynamic representation of one of the largest CRISPR systems, with information that can contribute to understanding the mechanism of nucleic acid binding and, eventually, to transposase recruitment itself. Such information may prove informative to advance genome engineering efforts.

Introduction

CRISPR/Cas systems, originating as key components to bacterial adaptive immunity, have been successfully re-engineered for use in the targeted genetic editing of plant and mammalian cells (1–3). Such systems, therefore, have massive potential as therapeutics for genetic disorders. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) / Cas (CRISPR-associated) systems use a guide, or CRISPR-RNA (crRNA) to detect a target DNA sequence, or 'target strand' (TS). Upon successful detection, the complex conformationally and dynamically adjusts to form a structure called the 'R Loop.' Proteins associated with the crRNA (the 'Cas' proteins) may then either directly cut the DNA or recruit an external nuclease to do so. By design, such a system is ideal for creating knockouts of genes, but insertion of genes remains a challenge. Intriguingly, though not surprisingly, nature is one step ahead of us. Bacteria have evolved variants of these standard CRISPR/Cas systems, which lack cleavage capability, for use in site-specific transposition events (4, 5). These variants originate from CRISPR-associated transposons (CASTs) and, upon identifying the TS, recruit key transposition proteins which guide a transposase to paste the transposon at a specified site. In this way, CAST systems leverage the sequence specificity of CRISPR/Cas and the 'cutting-pasting' enzymatic capability of transposases to establish an accurate gene insertion paradigm. Thus, bacteria have essentially evolved a robust tool for site-specific transposition events, one that we may study in the hopes of guiding current genome engineering efforts, especially gene insertions (knock-ins) with high fidelity. We here focus on a particular CAST, Tn6677, discovered in *Vibrio Cholerae* by the Sternberg group which encodes a variant of the multi-subunit type 1F CRISPR complex (also called 'Cascade'), known as 1F-3 (4, 6). 1F-3 associates with a homodimer of the transposition protein TniQ to form a 'Cascade-TniQ' complex. Cascade-TniQ then undergoes a search for the TS and, as with any CRISPR/Cas system, forms the R Loop upon successful detection. TniQ then recruits the protein tnsC which, upon activation, recruits the heteromeric transposase tnsA/B for the pasting of Tn6677 (Fig. 1a). It is important to note that tnsC recruitment details are still unknown, Hoffman, et al. have observed a heptamer of tnsC's bound to TniQ as the structure which potentially recruits the transposase (7). Regardless of downstream activity, the first regulatory checkpoint of the transposition of Tn6677 is the R Loop formation of Cascade-TniQ, and this is the focus of the present work.

The mechanistic details of R Loop formation remain unknown. Furthermore, there are very few structural differences of Cascade-TniQ before and after R Loop formation, according to deposited models (6, 8). In fact, the models are all nearly superimposable, with root mean square differences (RMSD) of their polypeptide backbones being ~1 Å (SI). We therefore believe there could be relevant structural changes upon or during R Loop formation of certain parts of the complex that could not be modeled from cryo-Electron Microscopy data alone. Indeed, a continuous segment of ~100 amino acids within the Cas8 subunit of Cascade, coined the "bundle," could not be resolved by cryo-EM. Further, experimental studies of the original type 1F system indicate

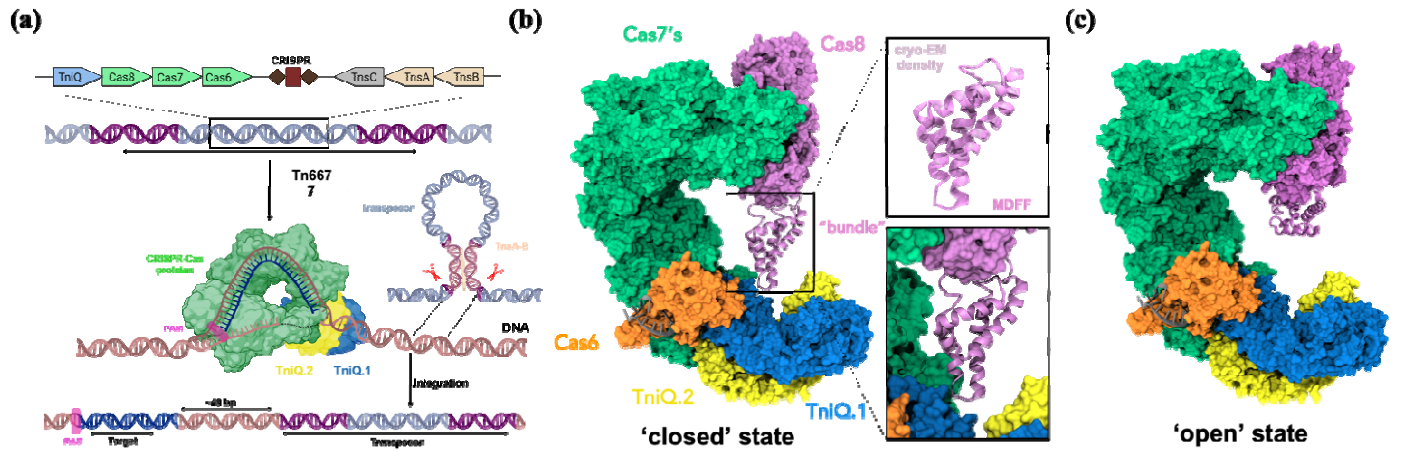

Figure 1: Details on the biology and organization of Cascade-TniQ. **(a)** Schematic of the transposition of Tn6677, top showing the components encoded within the transposon, middle showing a complete R Loop Cascade-TniQ complex, bottom showing location of integration of Tn6677 after pasting by tnsA/B. **(b)** 'Quicksurf' representation of the complete model in the closed state of Cascade-TniQ, top inset showing the overlay on the cryo-Electron Microscopy density of the bundle from EMD-20350[6]. Bottom inset showing a zoom of the bundle within the complex. **(c)** 'Quicksurf' representation of the complete model in the open state

the importance of its bundle in R Loop formation, particularly via canonical (vital to complete R Loop formation) non-target strand (NTS) interactions established after a large rotation of the bundle (9).

Under the assumption that NTS/bundle interactions promote complete R Loop formation in Cascade-TniQ, a complex exhibiting an incomplete R Loop ought to lack these interactions, implying Cascade-TniQ found difficulty in adjusting the bundle itself to an orientation which allows for these interactions. This implication would allow us to study the bundle without explicitly considering NTS interactions, and itself implies that the bundle is well regulated before and possibly during R Loop formation. We now make this regulation notion more specific. The electron density maps of Cascade-TniQ from Halpin-Healy, et al., EMD-20350 and EMD-20351, show weak electron density of the bundle in the same location for both the pre-R Loop (before TS binding) and incomplete R Loop states, respectively (6). We call this bundle position the 'closed' state of the bundle. The electron density maps of Jia, et al., EMD-21126, EMD-21146, on the contrary, do not exhibit any bundle density (indistinguishable from dust) and the DNA-bound model displays complete R Loop formation (8). Thus, there is likely a significant energetic minimum of the 'closed' bundle configuration for systems which cannot achieve a full R Loop, and which is attenuated for systems which do achieve a full R Loop. Now, if we knew the state the bundle ought to go to from the closed state, we could hypothesize that the transition barrier is higher in systems leading to partial R Loops than in the ones leading to complete R Loops. Naturally, we have no such information, so we look to the type 1F systems for some guidance. Guo, et al., obtained models for the pre-R Loop and incomplete R Loop for type 1F and observed that the bundle maintained its position, just like what Cascade-TniQ did in the Halpin-Healy, et al. results (10). The notable difference with Cascade-TniQ is that, in type 1F, the R Loop is just missing the last ~2 base pairs of TS/crRNA binding. In Cascade-TniQ, the incomplete state is missing the last ~10 base pairs. This discrepancy is most likely due to the differences in bundle position. In the type 1F models, the bundle's pre-R Loop state is close to the rest of Cas8, being able to interact with the NTS and therefore sequester it away from the TS, furthering R Loop formation via noncanonical interactions (as the R Loop remains incomplete). In Cascade-TniQ, the bundle's start, closed state is much further from the rest of Cas8, precluding these interactions and leading to a less developed partial R Loop. Additionally, during our Cascade-TniQ modeling, AlphaFold V2.1 gave 2 distinct configurations of the bundle, one of which was nearly the closed state (good overlap with electron density), while the other was similar, in spatial positioning, to Guo, et al.'s 1F bundle. We call this state the 'open' state of the bundle, details in 'Model Building' and in SI. If we now conjecture that this open state is at least attainable from the closed state, we can hypothesize that there exists at least one barrier between the closed and open states, the height of which determines whether the R Loop experiences further formation than as seen in the incomplete Cascade-TniQ (6). Explicitly, we are claiming that the closed-to-open transition will have a higher barrier for pre-R Loop systems which lead to partial R Loops, relative to the barrier for pre-R Loop systems which lead to complete R Loops. Additionally, due to the nature of the sampling required to test this hypothesis, the findings would serve the more general purpose of shedding light on the Cas8 bundle dynamics and regulation, which so far remain entirely unknown.

Methods + Experimental Design

To test our hypothesis, we create 2 pre-R Loop Cascade-TniQ models both in their open and closed states (as discussed in 'Model Building'). One open/closed pair is built from Halpin-Healy, et al.'s pre-R Loop structure (PDB: 6PIG) while the other is from Jia, et al.'s pre-R Loop structure (PDB: 6V9Q),

the closed states of which are named RNA-HH and RNA-J, respectively (6, 8). It is assumed that RNA-HH, upon DNA binding, will lead to an incomplete R Loop while RNA-J leads to a complete one for the following reasons. Key structural differences between RNA-HH and RNA-J are in the secondary structural ordering of the TniQ dimer and the presence of electron density of the bundle. In both metrics, the systems are much closer in form to their DNA-bound counterparts, in which the Halpin-Healy, et al.'s R Loop structure was incomplete (PDB: 6PIJ) while the Jia, et al.'s was complete (PDB: 6VBW) (6, 8). Thus, we posit that the closed-to-open transition of RNA-HH ought to exhibit a heightened free energy barrier relative to RNA-J's transition. As an additional point of inquiry, we also conjecture that these R Loop counterparts of the systems above, labeled DNA-HH and DNA-J respectively, will also exhibit the same pattern in their free energies. Of course, due to the lack of NTS in either system, this conjecture is very much weakened, but the results would be informative nonetheless by giving an idea of the direct impact of TS binding to crRNA. Thus, we prepare a total of 4 open/closed pairs labeled RNA-HH, RNA-J, DNA-HH, and DNA-J. These names refer to the system in potentially any state, whether we are discussing the closed, open, or any other state of these systems will be made clear with context or explicated.

To complete the structures for each of the systems, we build complete models as discussed in 'Model Building' below. To test our hypothesis itself, we used both classical molecular dynamics (cMD) simulations and enhanced sampling methods, which have been instrumental in characterizing conformational dynamics (11, 12), catalysis (13, 14), and specificity (15–18) of CRISPR-Cas9, and other genome editing systems (19, 20). The models were run under a cMD regime to test for their stability. We use Amber 2022's PMEMD engine and run in the NVT ensemble (21). We ran the simulations in 3 replicates for 2 microseconds each at 300 K, after an equilibration of each system (discussed below). The results of the cMD runs are available in the SI, and indeed show the systems to be dynamically stable across at least timescales up to the order of microseconds.

We ran an enhanced sampling method known as Umbrella Sampling (US) to deliberately 'pull' the bundle from its closed to its open state. This method effectively uses a series of harmonic potentials across a user-defined Reaction Coordinate (RC) to sample the system across said RC and uses a reweighting scheme to construct a Potential of Mean Force (PMF). We here ran 1D US simulations, using RMSD of the backbone atoms of Cas8 (Cα, C, N, O) with the open state of Cas8 as reference, as our single RC (details in SI). The simulations were run at 300 K using Amber 2022's PMEMD engine. In the first set, we ran all four systems, RNA-HH, RNA-J, DNA-HH, DNA-J. In the second set we ran only the RNA systems with the displaced Cas7.1 gate (RNA-Jd and RNA-HHd). Motivation for this second set is given in "An Important Find: The Cas7.1 Gate." The simulations were conducted across 30 windows and run for 60 ns each (which was the time it took to reach convergence of the PMF). The windows were centered at 0.5 Å increments, starting at 15.0 Å RMSD and going down to 0.5 Å, each supplying a harmonic force constant of 25 kcal/mol/Å^2. An initial pulling step of 0.4 ns was done to guide the Cas8 to an RMSD close to the centers of each window to avoid excessive forces during the simulations. Each US simulation was then reweighted using the variational Free Energy Perturbation approach (vFEP) (22). The plots demonstrating convergence, details regarding error analysis, and extent of overlap between adjacent windows are in the SI.

For both cMD and US, the systems were equilibrated in the same manner. We start with a short NVT heating phase for 0.1 ns linearly increasing the temperature from 10 K to 100 K, restraints of 10 kcal/mol/Å^2 are applied. We then continue heating linearly to 200 K in an NPT ensemble across 0.25 ns with restraints of 5 kcal/mol/Å^2, and then to 300 K across 0.25 ns with weaker restraints of 1.5 kcal/mol/Å^2. We finally run at 300 K without any restraints for 10 ns, still in the NPT ensemble.

Model Building

To build the complete structure, we divided the apo Cascade-TniQ complex (protein only) into 3 multimers labeled MM1 - MM3, the primary sequences of which were given as input to AlphaFold V2.1 via ColabFold (23, 24). The best ranked multimers for MM1 and MM3 were then reassembled into a model of Cascade-TniQ (figure S). To reassemble MM2, which contained Cas8 and so its bundle, we chose the best ranked result which agreed most closely with the electron density of the bundle as seen in electron maps of Cascade-TniQ (EMD-20350, EMD-20351) (figure S). The complete model was then fragmented and threaded ('cannibalized') into the incomplete cryo-EM based Cascade-TniQ model using PyRosetta (25, 26). In summary, we had essentially used AlphaFold to fill in the missing regions of the cryo-EM deposited structure. Because we input Cas8 via a partial multimer into AlphaFold, the bundle was not assembled considering the potential interactions with other members of the Cascade-TniQ complex, so we desired to refine the position of the bundle. A model produced by Li, et al. treated the bundle as a poly-alanine chain and yielded much better resolution of its polypeptide backbone (27). We ran a Molecular Dynamics based fitting program, called 'Molecular Dynamics Flexible Fitting' (MDFF) to fit the AlphaFold bundle to their poly-alanine bundle density (28) (Fig. SI). Indeed, the resulting bundle positioning much more closely agreed with the electron density seen of the bundle from the other cryo-EM maps (EMD-20350, EMD-20351) (Fig. 1b, inset). Finally, since the model is basically a chimera of two structures (one from AlphaFold, the other from cryo-EM), we 'relaxed' the model using MDFF to fit to the electron density

map associated with the cryo-EM based structure. The final structure resulted in the 'closed' state of Cascade-TniQ (Figure 1b). By plugging in the relevant cryo-EM model of interest into the methodology

above to thread AlphaFold results into, we were able to produce all the 'closed' state models used in this work.

When AlphaFold was given the primary sequence of MM2, its overall best ranked multimer contained a Cas8 with the bundle oriented in the open position (figure S for AF plots). We created the 'open' state of Cascade-TniQ in the same way as the closed state, but with the Cas8 in its open configuration (Fig. 1c). Since there is only one open state of the bundle, as it was obtained as the best ranked result of AlphaFold, every model we built had the same open Cas8 configuration. It should be noted that AlphaFold being able to output this open state is testament to its versatility and 'insight,' though of course caution should be exercised when both using and interpreting its outputs! Further details on the multimers used, the AlphaFold results, and the methodology is available in the SI.

Due to the complexity of the Cascade-TniQ complex, we will take here the opportunity to explain its structure. Cascade-TniQ is comprised of 10 proteins with the following stoichiometry, $Cas8_1$:$Cas7_6$:$Cas6_1$:$TniQ_2$. The overall shape is that of a croissant, with the bundle 'gluing' the two ends of the croissant together in its closed state. The Cas8 "body" (without the bundle) is on one end of the croissant, while Cas6, TniQ.1, and TniQ.2 are on the other end. Finally, the Cas7s are adjacent to one another and form the 'backbone' of the full complex. The crRNA 'threads' through the whole structure, primarily passing through the Cas7s and Cas6. We will refer to the end with the TniQ homodimer, Cas6, and Cas7.1 as simply the "TniQ end," and the opposite end with the Cas8 body as the "Cas8 end."

The Importance of the Cas8 Bundle

The results of the US simulations do indicate a larger barrier in the RNA-HH case relative to RNA-J, proving the hypothesis (Fig. 2a). Specifically, RNA-HH exhibits a ~15 kcal/mol increased barrier to the open state compared with RNA-J. Furthermore, the curves show RNA-HH's bundle constantly biased against traversal from the closed state. In contrast, RNA-J's bundle reaches a diffusive region from which it can easily progress towards the open state. The free energy trends persist in the DNA-HH and DNA-J comparisons, in which DNA-HH has a ~10 kcal/mol barrier to the open state while DNA-J has a ~5 kcal/mol barrier (Fig. 2b). Additionally, it does hold that in both sets of systems, the RNA-to-DNA transition decreases the free energy barrier to the open state and decreases the open state free energies themselves. These results suggest that the TS hybridizing to the crRNA assists in the process of re-orienting the bundle, considering bundle/NTS interactions are absent in these simulations. If we now factor in bundle/NTS interactions, which would also begin to occur during R Loop formation, we make the following claim. The formation of the R Loop is a gradual process, as more TS binds to crRNA and more NTS binds to the bundle, the relative free energy of the open:closed state decreases, as does the barrier to the open state, allowing continued R Loop formation to be spontaneous. The early interactions between both crRNA/TS and bundle/NTS are sufficient to drive this process in the case of successful R Loop formation, where the barrier is lower (RNA-J), but insufficient for the case where the barrier height is too great (RNA-HH). It is important to reiterate that the open state was similar in configuration to Guo, et al.'s Cas8 bundle, which didn't move from its pre-R Loop position, thus, our open state may very well be close to an intermediate state. We therefore posit that R Loop formation is regulated by the bundle's ability to reach its intermediate configuration(s), giving us a hint of a checkpoint during the R Loop formation process.

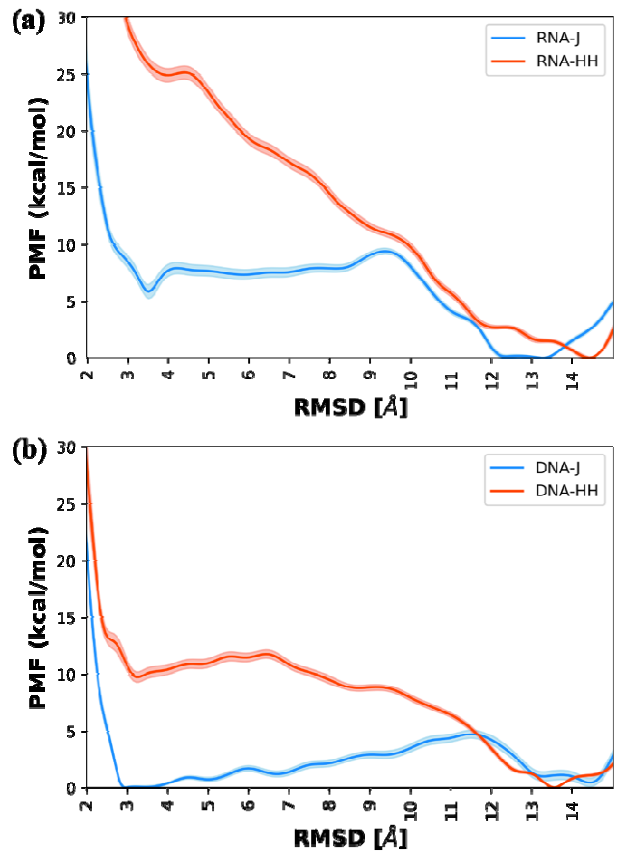

Figure 2: Free Energy Plots of the US Simulations. **(a)** PMF plots of RNA-HH (orange-red) and RNA-J (dodger blue), the thickened shading represents error as +/- standard deviation, obtained from bootstrapping the original sample. **(b)** PMF plots of DNA-HH (orange-red) and DNA-J (blue), where the thickened region again represents the error. The plots are cutoff at 2 Å since we are pulling to a non-equilibrated Cas8 structure (one that was only energy minimized), hence the energy hikes above. Additionally, the start, closed configuration for each system varied between ~12.5 to ~13.5 Å, consistent with the closed state minima above.

An Important Find: The Cas7.1 'Gate'

The obvious question that remains is why there ought to be a difference at all in the bundle transition barriers of RNA-HH

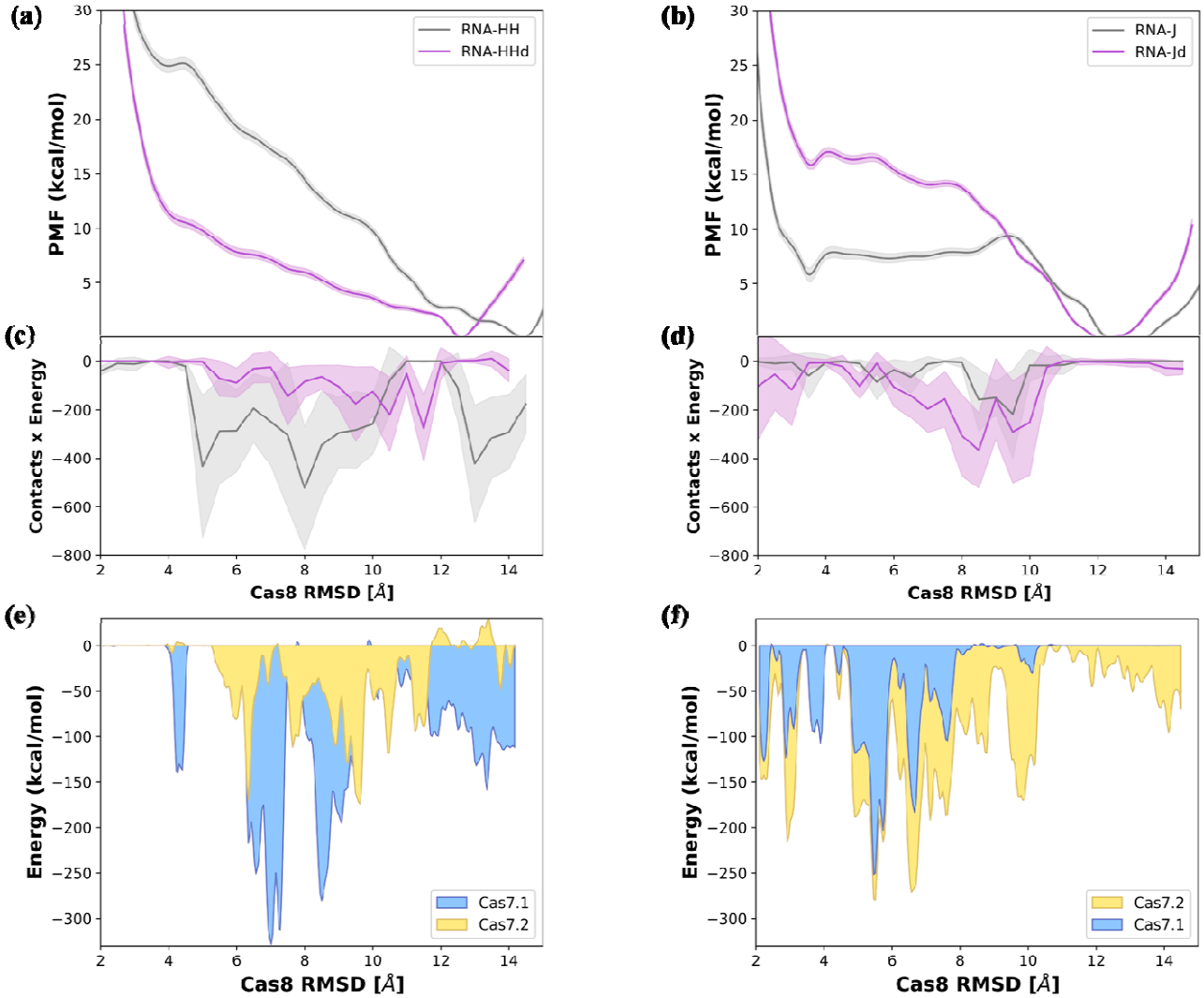

Figure 3: Comparisons of the RNA-bound systems and their Cas7.1 gate displaced counterparts. **Figures 3a and 3b** show the former PMF's of RNA-HH and RNA-J (gray) with their gate-displaced counterparts RNA-HHd and RNA-Jd (purple), respectively. **Figures 3c and 3d** show the values of the product of Linear Interaction Energy (LIE) and number of contacts between the hairpin loop of Cas7.2 (resid 48 - 63) and the Cas8 bundle. The color scheme follows figures 3a and 3b for 3c and 3d, respectively. Finally, **figures 3e and 3f** are the LIE results for RNA-HHd and RNA-Jd, respectively. For comparison, figures S3a, S3b contain the LIE values for RNA-HH and RNA-J.

and RNA-J. As mentioned before, the secondary structural ordering of the TniQ dimer was noticeably different. RNA-J had a greater number of alpha helical loops around the Zinc finger domains of both TniQ monomers, whereas RNA-HH was disordered in these regions, therefore we conjecture that this difference is the driving force behind the bundle transition barrier discrepancy. Specifically, we claim that the TniQ dimer 'holds' on to the bundle more tightly in RNA-HH, increasing the transition barrier, relative to the RNA-J case.

To test this, we examined the electrostatic interaction energies between the bundle and the TniQ dimer in both RNA-HH and RNA-J. Surprisingly, the interaction energy plots were ambiguous at best, it seems that the TniQ dimer, if anything, may have an allosteric role in driving the barrier discrepancy, but unlikely has a direct role. We then examined the electrostatic interaction energies between the bundle and all subunits of Cascade-TniQ which interact extensively with it, this excluded Cas6, Cas7.4, Cas7.5, and Cas7.6. Of the remaining subunits, the interaction energies with Cas7.1 and Cas7.2 exhibited the largest differences between RNA-HH and RNA-J (Fig. S3A, S3B). We then conducted a distance analysis (as electrostatic energies are inversely proportional to the square of the distance) between the bundle and amino acids within Cas7.1 and Cas7.2. A particular loop of 10 amino acids in the Cas7.1 subunit, which we are calling the Cas7.1 'gate,' demonstrated the biggest difference in positioning between RNA-HH and RNA-J (Fig. S3C). This was also the major driving factor behind the large interaction energies between the bundle and Cas7.1 in RNA-J, which was essentially absent in RNA-HH (figure S3A, S3B).

It is likely that this gate's orientation is crucial to the ease of bundle opening in Cascade-TniQ. Specifically, a gate

in its 'locked' state, which is the orientation exhibited in RNA-J, is the preferred configuration, while the 'unlocked' gate, present in RNA-HH, prevents the bundle from easily opening due to its allowance of bundle/Cas7.2 interactions. To test this, we ran a series of Molecular Dynamics Flexible Fitting (MDFF) runs to have the Cas7.1 gates swap places. The RNA-J system, in its closed state, was driven to push the gate towards the unlocked configuration, while RNA-HH in its closed state was driven to push its gate into the locked configuration. Since the rest of the systems remain practically still (they were still run under NAMD's molecular dynamics protocol, but the timescale was so short that no meaningful conformational change occurred), the only relevant change was to the Cas7.1 gate. We ran the same Umbrella Sampling protocol as before on these initial configurations with displaced gates (RNA-HHd and RNA-Jd) and confirmed convergence of the free energy profiles, the results do in fact show the importance of the gate (Fig. 3a, 3b). RNA-HHd indeed experienced a drop in the free energy profile compared to RNA-HH. Meanwhile RNA-Jd demonstrated a heightened free energy barrier and increased free energy of its open state compared to RNA-J. Details on the MDFF protocol and on the Cas7.1 Gate are in the SI (Fig. S4). Curiously, the form of the free energy profiles themselves do not change much. RNA-Jd still well-maintained a diffusive region, while RNA-HHd maintained a constant bias against opening. In the below we give a potential explanation of this behavior.

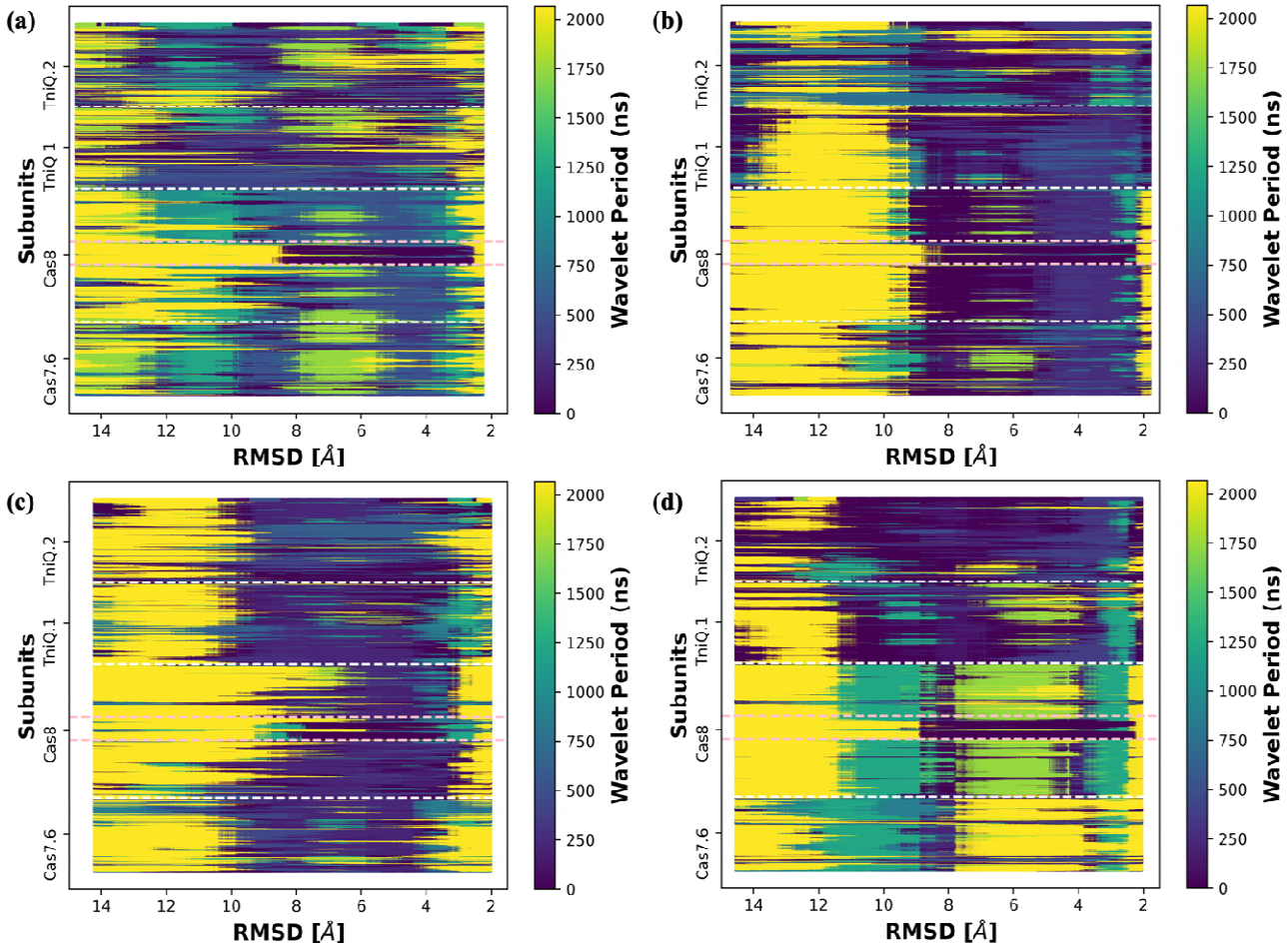

Figure 4: Wavelet Decomposition of the data trains for the C atoms of the subunits shown. **Figure 4a** shows the result for RNA-HH, **figure 4b** for RNA-J, **figure 4c** for RNA-HHd and **figure 4d** for RNA-Jd. The color gradient range was fixed for all systems and indicates the range of wavelet scales, a parameter proportional to the period of the wavelet. The white dashed lines indicate the cutoffs for each of the subunits shown, and the pink dashed lines indicate the 'bundle' within Cas8.

On The Shape of the Free Energy Curves

We posit that the Cas7.1 gate's role is predominantly to prevent the bundle from falling into a minimum. This minimum is one in which the bundle would primarily interact with the hairpin loops of Cas7.2. This is generally corroborated by the interaction energy analysis of the RNA-HH and RNA-Jd systems, in which the interactions with Cas7.2 were substantially greater than their gate-locked counterparts (Fig. 3e,f). To see whether the interaction between the bundle and these hairpin loops of Cas7.2 indeed drive the discrepancy in the free energy curves, a proxy for enthalpy was computed. This measure is the product of the number of contacts and interaction energies between two regions of interest, scaled by the average number of contacts

across the full simulation. Thus, values close to zero demonstrate little to no interactions, while large absolute values correspond to strong and/or extensive interactions. To compute this metric, the simulation was filtered based on the RMSD of the Cas8 backbone. Windows were generated from 2 Å RMSD to 14.5 Å RMSD, the centers of which were separated by 0.5 Å, each window spanning 0.25 Å on either side. This filtered the trajectory into a total of 26 sets of simulation frames, the mean and standard deviations of this 'contact*energy' product for each of these sets were then plotted. In Fig. 3c,d, this measure is applied to the bundle/Cas7.2 hairpin loop (residues 48-63 of Cas7.2). The RNA-HH curve, figure 3c, demonstrates a strong declining trend from 11 Å to 8 Å RMSD, corresponding to a constantly increasing number of stabilizing interactions. On the contrary, RNA-HHd shows a slightly upward trend for the same region, suggesting an appreciably lesser number of stabilizing interactions (Fig. 3c). These trends are flipped in the Jia, et. al systems. The RNA-Jd curve, figure 3d, shows a steeper decline from ~10 Å to ~8 Å RMSD relative to the RNA-J curve, exactly where the corresponding free energy profiles diverge (Fig. 3b). Lastly, below the 6 Å mark, the curves in figure 3d converge, as do the forms of the free energy curves in figure 3b. Thus, this analysis shows the flipped nature of the bundle/Cas7.2 loop interactions between the native systems and their gate-displaced counterparts and establishes those interactions as a key component of the free energy profiles.

There is one discrepancy that is not completely clarified in the above, both RNA-J and RNA-Jd have similar forms in Figure 3b, as do RNA-HH and RNA-HHd in figure 3a. This implies that there is something persistent in the 2 sets of systems, independent of how we displace the Cas7.1 gate. The discrepancy manifests as a 'pivot' at ~9.5 Å RMSD into a diffusive region in RNA-J, and in RNA-Jd this behavior persists at the adjusted point of ~8 Å RMSD, such pivots are not present in neither RNA-HH nor RNA-HHd. To address this, we investigated the systems from a purely dynamic perspective, trying to see what significant motions or correlations are established as the bundle transitioned to the open state. To do so, we decomposed the trajectories of each atom of interest (Cα of Cas8, Cas7.6, and TniQ) into wavelets and examined those scales which deviated appreciably from expectation, as outlined in Heidari, et. al (29). To describe this approach we will use an analogy, imagine you're a judge in an orchestra competition. You are judging two orchestras on their performance of Beethoven's No. 5. To decide the winner, you wish to see who was the closest to the actual piece. We can do this mathematically by breaking down the original symphony into a weighted sum of wavelets, or 'little waves' which can be visualized as a 2D cross-section of a ripple in a pond (there are numerous wavelets one can use, so this visual is just to give a soft introduction to wavelets). This 'decomposition' will capture the high frequency sounds using wavelets with low scale values and lower frequency sounds using high scale wavelets, as the scale is proportional to the period. It will also capture localized positions in time since wavelets, as the ripples in the pond, vanish outside a finite interval. Thus, all the components of the symphony can be coded into a sum of wavelets with appropriate coefficients which are chosen exactly as in Fourier Analysis, via the inner product of the symphony signal with each wavelet.

We can then do the same decomposition with the two performances, the distribution of weights that most closely matches the original would be deemed the winner! Suppose now that you wish to help the losing group for their future performances. You could use the same wavelet decomposition to see what parts of the signal deviated the most from the original piece at each point in time. Maybe 45 seconds in a violinist misplayed, suppose this was the only mistake at that time point, this would present as the maximal deviation from expectation in the coefficients corresponding to the wavelets that add up to the violin's melody at the time points around 45 seconds. This is precisely analogous to what we are doing here, in place of a symphony of sounds we have a collection of trajectories, and we wish to see at what wavelet scale each atom's trajectory deviates the most from the expectation of a normally distributed collection of coefficients, for each point in time. These scales are presented as colors on a heat map, where the 2 axes are the RMSD of the Cas8 backbone in each simulation frame (to compare to the PMFs directly) and the other axis the atoms in question. Of course, the technical details to measure said deviations and how the wavelet coefficients are computed can be found in Heidari, et. al (29). To sum, the plots in Figure 4 demonstrate the timescales for the motions which are most 'out of the ordinary.'

There are two major themes at play in the wavelet data, one, the coupling of subunits and two, the switch in the timescales as the Cas8 achieves its open configuration. In RNA-J, Fig. 4b, we can see strong coupling between Cas8, Cas7.6, and TniQ.1. We also see a distinctive shift from slow modes between 14.5 Å and 9.5 Å to fast modes for nearly the rest of the simulation that happens simultaneously across nearly all atoms. That 9.5 Å point is also where we see the pivot in figure 3b. Contrasting these findings with RNA-HH, Fig. 4a, we do not have clear coupling of virtually any extent across the subunits, and we have a cacophonous mix of modes as opposed to a clear switch between modes across the subunits. The RNA-HHd system in Fig. 4c nearly recovers the trends seen in RNA-J of extensive coupling between the relevant subunits, and a distinct shift from a slow to a fast mode across said subunits. The caveat here is that Cas8 lags in the switch, gradually switching to the fast mode. Finally, RNA-Jd, depicted in Figure 4d, acquires a mix of modes as seen in RNA-HH, yet does maintain clearer shifts of modes, happening at ~9 Å and ~7.8 Å. Importantly, these shifts of modes remain simultaneous in nature, reminiscent of the RNA-J behavior at 9.5 Å. At both RMSD values, RNA-Jd in Fig. 3b displays an inflection that tapers the free energy

profile, the 9 Å point's tapering is much weaker due to the still present bundle/Cas7.2 interactions as seen in Fig. 3d. Taken together, these results seem to suggest that the pivot points we see in the systems of Jia, et al. are the result of a 'dynamic driver,' switching the motions of the Cas8 body and other coupled subunits, namely Cas7.6 and TniQ.1, simultaneously to promote bundle opening. If this switch is attenuated by weakening the coupling between Cas8 and the other key subunits, we will lose the pivot point. Since the only major structural deviation between the systems are their Cas7.1 gate orientations, these findings unveil an alternate role of the Cas7.1 gate, being a potential coupling device, synchronizing the motions of Cas8 with Cas7.6 and TniQ.1 when locked. Figure 4d still shows decent coupling between Cas8 and Cas7.6, so while Cas7.1 gate locking might not be sufficient for extensive subunit coupling, it is likely necessary.

All in all, our findings suggest that the Cas7.1 gate plays both an 'enthalpic' role of blocking bundle access to a minimum involving a hairpin loop of Cas7.2, and a dynamic role of coupling the Cas8 subunit to Cas7.6 and TniQ, for the purpose of introducing a simultaneous shift in modes of an extensive collection of atoms within these subunits. The question of why this coupling even impacts bundle opening remains outstanding. We speculate that the simultaneity of the switch in modes across so many atoms is indicative of an energy transfer event, which may promote the overcoming of energy barriers to further bundle opening. Once these barriers are overcome, the push to the open state would become favorable, as the Cas8 open state is likely the favored state, as suggested by the fact that AlphaFold gave the open state as the expected conformation of Cas8.

Conclusions

In this work, we desired to elucidate the relevant components of R Loop formation for the Cascade-TniQ system. We began with building the complete model of the system using AlphaFold V2.1 and a threading approach to capture the structural deviations between the different experimentally determined structures. To probe the process of R Loop formation, we treated an AlphaFold result, the 'open' configuration of Cas8, as an endpoint for Umbrella Sampling simulations. Those simulations were quite information rich, and showed the clear discrepancies between a system that could not form a complete R Loop, and one that could. We learned that the formation process is likely tightly regulated and primarily driven by the ease of motions of the Cas8 bundle, itself being dependent upon a hairpin loop of the Cas7.1 protein we are calling the Cas7.1 gate. The locked configuration promotes opening of the bundle from an enthalpic route, in which it blocks access to a minimum at Cas7.2, and from a dynamic route, coupling Cas7.6 and the TniQ dimer to Cas8 which may lead to an energy transfer event pivoting the PMF into a diffusive regime. It is important, now, to point out the elephant in the room. If RNA-J and RNA-HH were identical, how could one have such a different Cas7.1 gate configuration? This comes down to a point mentioned earlier regarding their differences, namely that RNA-J had Zinc coordination within its TniQ dimer's ZnF sites, while RNA-HH did not. This led to structural deviations of their secondary elements, namely alpha helices and beta turns. This likely was the reason for a dynamical difference between the two, indeed numerous analyses show that the TniQ dimer is in general much more flexible in RNA-HH. These differences could impact regions of TniQ.1 which interact directly with the Cas7.1 gate, causing a difference in the Cas7.1 gate orientation between RNA-HH and RNA-J, ultimately leading to the bulk of the discrepancy in the free energy profiles. Naturally, a great deal is still left to explore, such as the allosteric role of the Cas7.1 gate, which would be crucial to drive coupling, or the details behind the impact of the active ZnF sites of TniQ, but hopefully the ideas presented in this work usher us a step forward towards a better understanding of this rather colossal system.

Acknowledgments

This material is based upon work supported by the National Institutes of Health (Grant No. R01GM141329, to G.P.) and the National Science Foundation (Grant No. CHE- 2144823, to G.P.). This work used Expanse at the San Diego Super-computing Center through allocation MCB160059 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Com- puter time was also provided by the National Energy Research Scientific Computing Center (NERSC) under Grant No M3807

References

1. S. Das, S. Bano, P. Kapse, G. C. Kundu, CRISPR based therapeutics: a new paradigm in cancer precision medicine. *Mol Cancer* **21**, 85 (2022).
2. K. El-Mounadi, M. L. Morales-Floriano, H. Garcia-Ruiz, Principles, Applications, and Biosafety of Plant Genome Editing Using CRISPR-Cas9. *Front Plant Sci* **11** (2020).
3. M. Jinek, *et al.*, A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (1979)* **337**, 816–821 (2012).
4. S. E. Klompe, P. L. H. Vo, T. S. Halpin-Healy, S. H. Sternberg, Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
5. J.-U. Park, *et al.*, Structures of the holo CRISPR RNA-guided transposon integration complex. *Nature* **613**, 775–782 (2023).
6. T. S. Halpin-Healy, S. E. Klompe, S. H. Sternberg, I. S. Fernández, Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system. *Nature* **577**, 271–274 (2020).

7. F. T. Hoffmann, *et al.*, Selective TnsC recruitment enhances the fidelity of RNA-guided transposition. *Nature* **609**, 384–393 (2022).
8. N. Jia, W. Xie, M. J. de la Cruz, E. T. Eng, D. J. Patel, Structure–function insights into the initial step of DNA integration by a CRISPR–Cas–Transposon complex. *Cell Res* **30**, 182–184 (2020).
9. M. F. Rollins, *et al.*, Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. *Mol Cell* **74**, 132-142.e5 (2019).
10. T. W. Guo, *et al.*, Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. *Cell* **171**, 414-426.e12 (2017).
11. G. Palermo, Y. Miao, R. C. Walker, M. Jinek, J. A. McCammon, CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. *Proceedings of the National Academy of Sciences* **114**, 7260–7265 (2017).
12. A. Saha, P. R. Arantes, G. Palermo, Dynamics and mechanisms of CRISPR-Cas9 through the lens of computational methods. *Curr Opin Struct Biol* **75**, 102400 (2022).
13. L. Casalino, Ł. Nierzwicki, M. Jinek, G. Palermo, Catalytic Mechanism of Non-Target DNA Cleavage in CRISPR-Cas9 Revealed by *Ab Initio* Molecular Dynamics. *ACS Catal* **10**, 13596–13605 (2020).
14. Ł. Nierzwicki, *et al.*, Principles of target DNA cleavage and the role of Mg2+ in the catalysis of CRISPR–Cas9. *Nat Catal* **5**, 912–922 (2022).
15. B. P. Mitchell, *et al.*, Spontaneous Embedding of DNA Mismatches Within the RNA:DNA Hybrid of CRISPR-Cas9. *Front Mol Biosci* **7** (2020).
16. L. Nierzwicki, *et al.*, Enhanced specificity mutations perturb allosteric signaling in CRISPR-Cas9. *Elife* **10** (2021).
17. C. G. Ricci, *et al.*, Deciphering Off-Target Effects in CRISPR-Cas9 through Accelerated Molecular Dynamics. *ACS Cent Sci* **5**, 651–662 (2019).
18. E. Skeens, *et al.*, High-fidelity, hyper-accurate, and evolved mutants rewire atomic-level communication in CRISPR-Cas9. *Sci Adv* **10** (2024).
19. S. Sinha, *et al.*, Unveiling the RNA-mediated allosteric activation discloses functional hotspots in CRISPR-Cas13a. *Nucleic Acids Res* **52**, 906–920 (2024).
20. A. Saha, *et al.*, An alpha-helical lid guides the target DNA toward catalysis in CRISPR-Cas12a. *Nat Commun* **15**, 1473 (2024).
21. Case, D. A. et al. AMBER 2022 (University of California, San Francisco, 2022).
22. T.-S. Lee, B. K. Radak, A. Pabis, D. M. York, A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations. *J Chem Theory Comput* **9**, 153–164 (2013).
23. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
24. M. Mirdita, *et al.*, ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
25. S. Chaudhury, S. Lyskov, J. J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
26. M. ’Ferla, Filling missing loops by cannibalising alphafold2. *https://blog.matteoferla.com/2021/10/filling-missing-loops-by-cannibalising.html* (2021).
27. Z. Li, H. Zhang, R. Xiao, L. Chang, Cryo-EM structure of a type I-F CRISPR RNA guided surveillance complex bound to transposition protein TniQ. *Cell Res* **30**, 179–181 (2020).
28. L. G. Trabuco, E. Villa, E. Schreiner, C. B. Harrison, K. Schulten, Molecular dynamics flexible fitting: A practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49**, 174–180 (2009).
29. Z. Heidari, D. R. Roe, R. Galindo-Murillo, J. B. Ghasemi, T. E. Cheatham, Using Wavelet Analysis To Assist in Identification of Significant Events in Molecular Dynamics Simulations. *J Chem Inf Model* **56**, 1282–1291 (2016).