Mechanism of expanded DNA recognition in xCas9

Kazi A. Hossain,[1,5] Lukasz Nierzwicki,[1] Modesto Orozco,[3,4] Jacek Czub,[5,6] and Giulia Palermo*[1,2]

[1] Department of Bioengineering and [2] Department of Chemistry, University of California Riverside, 900 University Avenue, Riverside, CA 52512, United States

[3] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.

[4] Departament de Bioquímica i Biomedicina, Facultat de Biologia, Universitat de Barcelona, 08028 Barcelona, Spain.

[5] Department of Physical Chemistry, Gdańsk University of Technology, Gdańsk 80-233, Poland

[6] BioTechMed Center, Gdańsk University of Technology, Gdańsk 80-233, Poland

CORRESPONDENCE:

Giulia Palermo (giulia.palermo@ucr.edu)

Keywords: CRISPR-Cas9, RNA, genome editing, protein dynamics, molecular dynamics

ABSTRACT

xCas9 is an evolved variant of the CRISPR-Cas9 genome editing system, engineered to improve specificity and reduce undesired off-target effects. How xCas9 expands the DNA targeting capability of Cas9 by recognizing a series of alternative Protospacer Adjacent Motif (PAM) sequences while ignoring others is unknown. Here, we establish the molecular determinants of xCas9's expanded PAM recognition. We show that while Cas9 enforces strict guanine selection through the rigidity of its interacting arginine dyad, xCas9 modulates the flexibility of R1335 to recognize specific PAM sequences selectively. This modulation confers a pronounced entropic preference for the canonical TGG PAM over SpCas9. Moreover, xCas9 expands PAM recognition by enhancing DNA binding in the early evolution cycles and improving binding to the canonical PAM in the final evolution cycle. This dual capability explains how xCas9 expands PAM recognition while also enhancing recognition of the canonical TGG PAM. These findings will facilitate the engineering of Cas9 variants more effective and specific across a broader spectrum of genetic sequences.

INTRODUCTION

Expanding the targeting scope of genome editing systems toward a broader spectrum of genetic sequences is a major priority to advance the CRISPR–Cas technology(Pacesa et al., 2024; Wang and Doudna, 2023). xCas9 is an evolved variant of the CRISPR-Cas9 system(Hu et al., 2018), engineered to improve specificity and reduce undesired off-target effects compared to the *S. pyogenes* Cas9 (SpCas9)(Anzalone et al., 2020; Pacesa et al., 2024). In this system, the endonuclease Cas9 associates with a guide RNA to recognize and cleave any desired DNA sequence enabling facile genome editing(Jinek et al., 2012). Site-specific recognition of the target DNA occurs through a short Protospacer Adjacent Motif (PAM) sequence next to the target DNA, enabling precise selection of the desired DNA sequence across the genome.

However, PAM recognition is also a significant bottleneck in fully exploiting the genome editing potential of CRISPR-Cas9(Anzalone et al., 2020; Pacesa et al., 2024). In SpCas9, recognition is strictly limited to 5'-NGG-3' PAM sequences, mediated by the binding to two arginine residues (R1333, R1335) within the PAM-interacting domain. This constraint significantly limits the DNA targeting capability of SpCas9, as the occurrence of NGG sites within a given genome is restricted. To address this inherent limitation, extensive protein engineering and directed evolution led to novel variants of the enzyme. The xCas9 3.7 variant (hereafter referred to as xCas9, Figure 1) has emerged as a notable advancement(Hu et al., 2018), expanding the PAM targeting capability toward a variety of sequences, including guanine- and adenine-containing PAMs. xCas9 not only demonstrates a significantly expanded PAM compatibility but also improves recognition of the canonical TGG PAM and reduces off-target effects compared to SpCas9(Hu et al., 2018). Despite these advancements, the mechanism by which xCas9 expands DNA targeting capability by recognizing a series of PAM sequences while ignoring others remains unknown.

xCas9 was developed through directed evolution introducing seven amino acid substitutions within SpCas9 (Figure 1). Notably, the only substitution within the PAM-interacting domain is E1219V and does not directly interact with the PAM sequence. Additionally, structures of xCas9 bound to AAG (PDB

6AEB) and GAT (PDB 6AEG) PAM sequences show no substantial differences in the PAM-interacting domain compared to SpCas9(Chen et al., 2019; Guo et al., 2019). Hence, this substitution does not explain xCas9's ability to recognize non-canonical PAM sequences, including those with adenine that typically do not favourably interact with arginine(Hossain et al., 2023; Luscombe, 2001). It is unclear how the E1219V mutation, rather than the R1333/R1335 substitution, facilitates binding to alternative PAMs. Additionally, the impact of other xCas9 mutations scattered across the protein on nucleic acid binding has not been addressed.

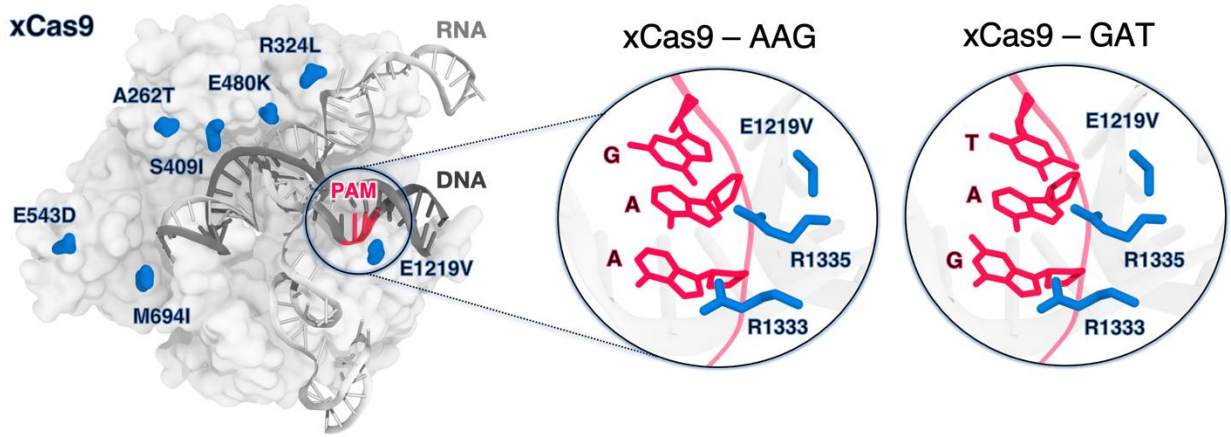

Figure 1. xCas9 variant of the *S. Pyogenes* Cas9 (SpCas9) protein bound to a guide RNA (gray) and a target DNA (charcoal) including the 5'-AGG-3' PAM recognition sequence (red) (PDB 6AEB)(Guo et al., 2019). xCas9 includes seven amino acid substitutions (blue) with respect to SpCas9. Close-up views of the PAM recognition region for xCas9 bound to AGG (PDB 6AEB, left)(Guo et al., 2019) and GAT (PDB 6AEG, right)(Guo et al., 2019). The PAM nucleobases (red) and the PAM interacting residues (R1333 and R1335, blue) are shown as sticks. The E1219V mutation is also shown.

Here, we establish the molecular determinants of the expanded DNA targeting capability of xCas9 through extensive molecular simulations, well-tempered metadynamics, and alchemical free energy calculations. We demonstrate that while SpCas9 enforces a strict guanine selection through the rigidity of its arginine dyad, xCas9 modulates the flexibility of R1335 to selectively recognize specific PAM sequences, conferring also a pronounced entropic preference for TGG over SpCas9. We also show that

directed evolution improves DNA binding, expanding the DNA targeting capability in the early evolution cycles, while achieving tight DNA binding with the canonical TGG PAM through the last evolution cycle. These findings will facilitate the development of improved Cas9 variants with expanded DNA recognition capabilities.

RESULTS

PAM recognition requires specific interactions

To elucidate the selection mechanism, we performed multi-μs molecular dynamics (MD) simulations of xCas9 bound to PAM sequences that are recognized well (TGG, GAT and AAG) and that are ignored (ATC, TTA and CCT), and compared to SpCas9 bound to its canonical PAM (TGG). We broadly explored the systems' dynamics and defined the statistical relevance of critical interactions through ~6 μs of sampling for each system (in four replicates, totalling ~42 μs runs).

Since in SpCas9, R1333 and R1335 provide anchoring for the PAM nucleotides (Figure 2a)(Anders et al., 2014), we performed an in-depth statistical analysis of their interactions with the DNA, using both distance and energetic criteria (details in Materials and Methods). We analysed the probability for R1333 and R1335 to interact with the PAM nucleobases (PAM NB), the PAM backbone (PAM BB), and non-PAM nucleotides (non-PAM) (Figure 2b). In SpCas9, both arginine residues steadily interact with the PAM NB(Anders et al., 2014; Bhattacharya and Satpati, 2024; Palermo et al., 2017), with R1335 exhibiting a higher probability compared to R1333. Analysis of the arginine' flexibility (Figure 2c) denotes that R1335 is remarkably constrained in SpCas9, likely a result of its interaction with the PAM NB and its vicinity to E1219, while R1333 is more flexible. Interestingly, in xCas9 bound to TGG, both R1333 and R1335 switch their interactions between the PAM NB and BB (Figure 2b). Similarly, for xCas9 bound to other recognized PAMs (AAG and GAT), the arginine dyad interacts with both PAM NB and BB, with R1335 being more specific than R1333, which also contacts non-PAM nucleotides.

Contrarily, for xCas9 bound to ignored PAM sequences (CCT, TTA, and ATC), no significant interactions with the PAM NB or BB were observed, while both arginine primarily interacted with non-PAM nucleotide. This is in line with SpCas9's specificity for NGG PAMs(Hu et al., 2018), and with the arginine's preference for guanine(Hossain et al., 2023; Luscombe, 2001). To further detail the interactions established by the arginine dyad and the nucleotides, we analysed the frequency of hydrogen bonds among them (Figures 2d). In SpCas9, R1335 secures its interaction with the G3 nucleobase (G3 NB), while R1333 interacts with G2 NB and contacts non-PAM nucleotides. In xCas9 bound to TGG, the arginine dyad maintains its interactions with PAM. Here, R1335 is more flexible (Figure 2c) and weakens its binding to the G3 in favour of the adjacent backbone (Figure 2d). When xCas9 binds AAG, R1333 forms hydrogen bonds with the A1 backbone, and R1335 predominantly interacts with the G3 nucleobase. Consistent with the analysis in Figure 2b, for ignored PAM sequences, both arginine fail to interact with the PAM nucleotides, while displaying a substantial increase in interactions with non-PAM residues.

We then computed a specificity index by measuring the frequency of hydrogen bond formation between a given arginine and the PAM nucleotides relative to the frequency of forming hydrogen bonds with non-PAM residues (Figure 2e). Remarkably, R1335 exhibits a distinct specificity pattern for PAM sequences recognized by xCas9 compared to those that are ignored. This suggests that R1335 may serve as a discriminator for recognizing specific PAM sequences in xCas9. In contrast, R1333, which possesses greater flexibility (Figure 2c), engages significantly with non-PAM residues (Figure 2b,d) rendering it non-specific (Figure 2e).

This demonstrates that in SpCas9, the arginine dyad is more rigid than in xCas9 (with R1335 more constrained than R1333). This rigidity enforces a strict selection of their preferred binding partners, which are guanine(Hossain et al., 2023; Luscombe, 2001). Contrarily, the E1219V substitution in xCas9 leads to increased flexibility of R1333/R1335, adjusting their interactions and expanding the number of PAM sequences that are recognized. However, to enable PAM recognition, effective interactions with the PAM nucleotides are required. Accordingly, the vast majority of recognized PAM sequences contain at least

one guanine, prone to interact with arginine(Hossain et al., 2023; Luscombe, 2001). Ignored PAM sequences do not form any interaction between the PAM nucleotides and the arginine dyad, which shifts away from PAM (Figure 2f).

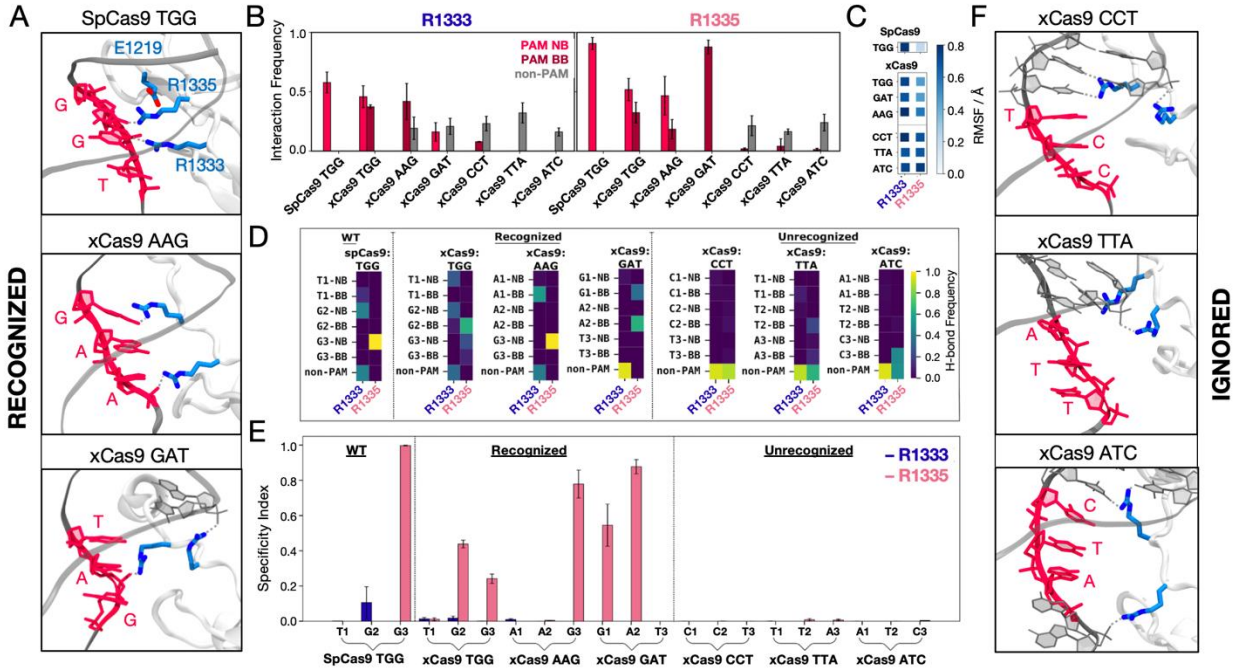

Figure 2. Binding of SpCas9 and xCas9 to PAM sequences that are recognized and ignored. A. Binding of SpCas9 and xCas9 to PAM sequences that are recognized (TGG, top; AAG, centre; GAT, bottom). **B.** Interaction pattern established by R1333 (left) and R1335 (right) with PAM nucleobases (NB), PAM backbone (BB) and non-PAM nucleotides in SpCas9 bound to TGG (i.e., the wilt-type system) and xCas9 bound to PAM sequences that are recognized (TGG, AAG, GAT) and ignored (CCT, TTA, ATC). Interaction frequencies are averaged over ~6 μs of collective ensemble for each system. Errors are computed as standard deviation of the mean over four simulations replicates. **C.** Root mean square fluctuations (RMSF) of the R1333 and R1335 side chains in SpCas9 bound to its TGG PAM, compared to xCas9 bound to recognized and ignored PAMs. **D.** Frequencies of hydrogen bond formation between the arginine side chains and the PAM NB, BB and non-PAM nucleotides (details in the SI). **E.** Specificity index, representing the frequency of hydrogen bond formation between a given arginine and the PAM nucleotides relative to the frequency of forming hydrogen bonds with non-PAM residues. Data are reported with the standard deviation of the mean over four simulations replicates. **F.** PAM recognition region in xCas9 bound to PAM sequences that are ignored (CCT, top; TTA, centre; ATC, bottom). Figure 2–figure supplement 1.

TGG binding is entropically favoured in xCas9

To delve deeper into the role of the R1335, we performed free energy simulations. Well-tempered metadynamics simulations(Barducci et al., 2008) were carried out to elucidate the preference of R1335 to bind either the G3 nucleobase or the neighbouring E1219 in SpCas9 (Figure 3a, details in Materials and Methods). Toward this aim, the free energy landscape was described along two collective variables (CVs) that define the distance of the R1335 guanidine from the E1219 carboxylic group (CV1) or the G3 nucleobase (CV2, details in Materials and Methods), and sampled through μs-long converged well-tempered metadynamics simulations (Figure 3–figure supplement 1). A well-defined free energy minimum at ~0.4 nm indicates that R1335 stably binds and is effectively "sandwiched" between E1219 and G3 (Figure 3a), which explains its rigidity and limited ability to recognize only NGG PAMs in SpCas9. We then sought to understand why xCas9 exhibits improved recognition of the TGG PAM sequence compared to SpCas9(Hu et al., 2018). To investigate this, we conducted well-tempered metadynamics simulations focusing on the binding of R1335 to the G3 nucleobase and the DNA backbone in both SpCas9 and xCas9 (Figure 3b-c). The free energy was described along the distances between the R1335 guanidine and either the backbone phosphate group (CV1) or the G3 functional group atoms (CV2, details in Materials and Methods). The obtained free energy landscapes reveal that in SpCas9, R1335 predominantly interacts with the G3 nucleobase (Figures 3b, Figure 3–figure supplement 2). This finding aligns with hydrogen bond analysis (Figure 2d) and is consistent with R1335 rigidity (Figure 2c). Conversely, in xCas9, R1335 exhibits dynamic interactions with both the G3 nucleobase and backbone (Figures 3c, Figure 3–figure supplement 3), lowering the entropic penalty for DNA binding, compared to SpCas9. This reduction in entropic penalty allows for a more adaptable interaction with the DNA, facilitating a smoother accommodation of the DNA's conformational freedom, and offering a rationale for the xCas9's enhanced recognition of TGG relative to SpCas9(Hu et al., 2018).

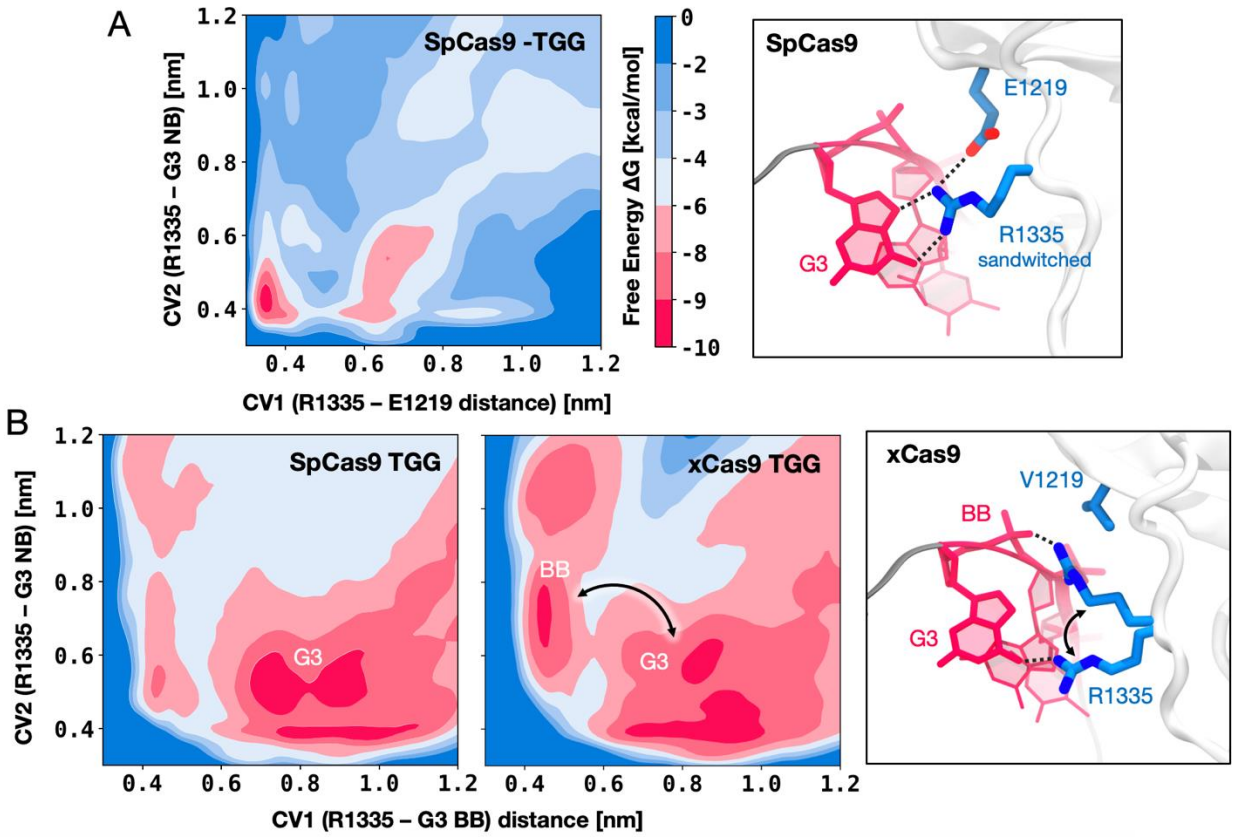

Figure 3. DNA binding preference of R1335 in SpCas9 vs. xCas9. A. Free energy surface (FES) describing the preference of R1335 for binding either the G3 nucleobase or E1219 in SpCas9. The FES is plotted along the distances between the R1335 guanidine and either the E1219 carboxylic group (CV1) or the G3 nucleobase (CV2, details in Materials and Methods). A well-defined minimum indicates that R1335 is "sandwiched" between G3 and E1219 (right). **B.** FES describing the binding of R1335 to G3 and the DNA backbone in SpCas9 (left) and xCas9 (centre). The FES is plotted along the distances between the R1335 guanidine and either the backbone phosphate (CV1) or G3 nucleobase (CV2). In SpCas9, R1335 mainly binds the G3 nucleobase, while in xCas9, it alternates interactions between the nucleobase and backbone (right). The free energy, ΔG, is expressed in kcal/mol. Figure 3–figure supplements 1-3.

Directed evolution improves DNA binding

xCas9 was developed through three cycles of directed evolution. The E480K, E543D, and E1219V amino acid substitutions were introduced in the first evolution cycle (leading to $xCas9_1$); A262T, S409I,

and M694I were introduced in the second evolution cycle (yielding $xCas9_2$); and R324L was included in the third evolution cycle (yielding $xCas9_3$)(Hu et al., 2018).

To assess how these substitutions contribute to expanding PAM recognition, we systematically introduced them into SpCas9 in a stepwise manner, following the evolution cycles, and evaluated their impact on DNA binding. We considered the TGG PAM sequence, effectively recognized by both SpCas9 and xCas9, and AAG, specifically recognized by xCas9. The contribution of each mutation cycle was computed as the difference in the DNA binding free energy ($\Delta\Delta G$) between SpCas9 and its mutant counterparts: $xCas9_1$, followed by $xCas9_2$, and finally $xCas9_3$ (Figure 4). We performed alchemical free energy calculations using a thermodynamic cycle (details in Materials and Methods), obtaining $\Delta\Delta G$ values by transforming the respective amino acid residues in the presence or absence of bound DNA (Figure 4–figure supplements 1-2).

The transition from SpCas9 to $xCas9_1$ for TGG PAM does not significantly improve binding ($\Delta\Delta G = -1.5 \pm 0.63$ kcal/mol), while the same transition for AAG PAM is notably favoured ($\Delta\Delta G = -4 \pm 0.47$ kcal/mol), consistent with xCas9's ability to recognize AAG. Upon introducing the second cycle of mutations (i.e., transforming $xCas9_1$ in $xCas9_2$), we observe a significant improvement in DNA binding in the presence of AAG ($\Delta\Delta G = -6 \pm 0.16$ kcal/mol) while no substantial change is noted for TGG PAM (Figure 3). Hence, the recognition of AAG improves in the second evolution cycle, as we further transition toward xCas9. It also shows that E1219V in the PAM-interacting domain is not the sole contributor to expanded PAM compatibility; and subsequent mutations play a crucial role. Finally, upon transitioning from $xCas9_2$ to $xCas9_3$, by introducing R324L, we observe a significant improvement in the DNA binding free energy in the presence of TGG (by -9 ± 0.82 kcal/mol), while the affinity for AAG only increased by -4 ± 0.18 kcal/mol. This suggests that the third cycle of mutation is the key player behind the enhanced recognizability of the canonical TGG PAM by xCas9. This observation suggests that, while xCas9 expands DNA recognition toward AAG by improving DNA binding in the early evolution

cycles, the enhanced recognizability of the canonical TGG PAM by xCas9 compared to SpCas9 observed experimentally primarily arises from the third cycle of mutations.

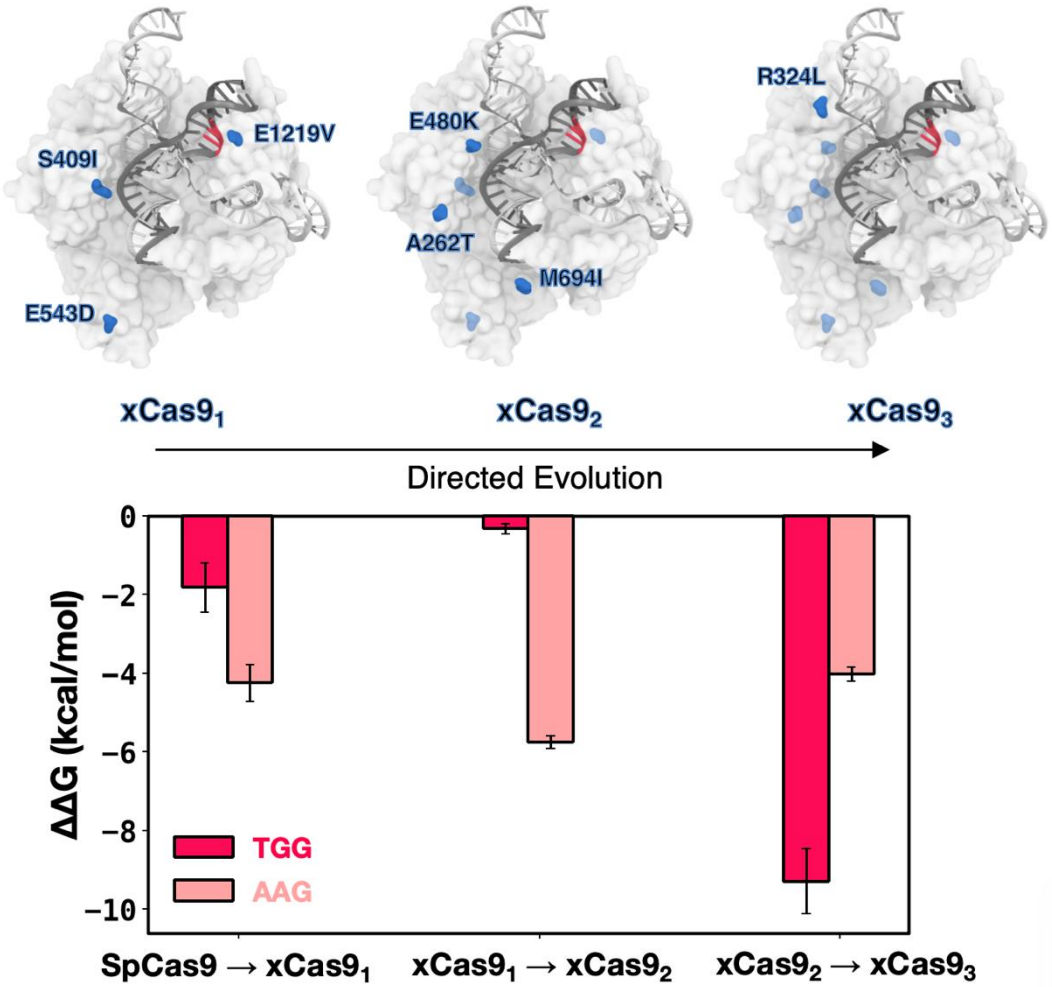

Figure 4. DNA binding free energy difference (ΔΔG) between SpCas9 and its xCas9$_{1\text{-}3}$ mutants. Relative changes in the DNA binding free energy (ΔΔG) upon transitioning from SpCas9 to xCas9$_1$, from xCas9$_1$ to xCas9$_2$, and from xCas9$_2$ to xCas9$_3$ in the presence of TGG (red) and AAG (pink) PAM sequences. Binding free energies from alchemical free energy calculations denoted with the associated error computed through the Multistate Bennett Acceptance Ratio (MBAR) method(Shirts and Chodera, 2008) (details in Materials and Methods). Figure 4–figure supplements 1-2.

AAG-induced conformational change

To better understand the PAM binding mechanism, we analysed the individual per-residue contributions to the DNA binding free energy. These changes denote the difference in enthalpic contribution (ΔE) to the DNA binding free energy (ΔΔG) upon transitioning from SpCas9 to $xCas9_1$, from $xCas9_1$ to $xCas9_2$, and from $xCas9_2$ to $xCas9_3$ (details in Materials and Methods).

For the SpCas9 → $xCas9_1$ transition, we computed these contributions for the critical residue R1335 (adjacent to the E1219V) and the E480K and E543D substitutions (Figure 5a). We observe that R1335 interacts more favourably with the AAG-bound $xCas9_1$ compared to TGG-bound $xCas9_1$, while the contribution of E480K is similar in both the TGG- and AAG-bound systems. Interestingly, the contribution of E543D is unfavourable in the presence of TGG, owing to a conformational change in the REC3 domain upon AAG binding. This finding aligns with the structural characterizations of xCas9. Superimposing xCas9 bound to TGG (PDB 6K4P)(Chen et al., 2019) onto the TGG-bound SpCas9 (PDB 4UN3)(Anders et al., 2014) reveals no significant differences, with a backbone RMSD of ~0.80 Å (Figure 5–figure supplement 1). In contrast, comparing TGG-bound SpCas9 with xCas9 bound to AAG (PDB 6AEB)(Guo et al., 2019) reveals a notable shift in the REC3 domain[7] (backbone RMSD of ~3.21 Å), consistent with our computational results. We also note that by computing the enthalpic contribution for the remaining protein residues, i.e., the ΔE as the overall interaction energy between the rest of the protein (without the selected residues) and the DNA, we observed no substantial difference for the TGG- and AAG-bound systems (Figure 5a). This indicates that the contribution from the E543D substitution, being the only substitution of the first cycle located in REC3, is mainly influenced by the conformational change of REC3.

To further investigate whether the observed conformational change in the REC3 domain is due to the mutations incorporated in xCas9 or to a PAM-induced allostery, we introduced the xCas9 mutations in the X-ray structure of SpCas9 (PDB 4UN3) and performed MD simulations in the presence of different

PAM sequences, i.e., TGG and AAG. These simulations were conducted in replicates, each reaching ~6 μs of sampling, similar to our equilibrium MD simulations of xCas9 (details in Materials and Methods).

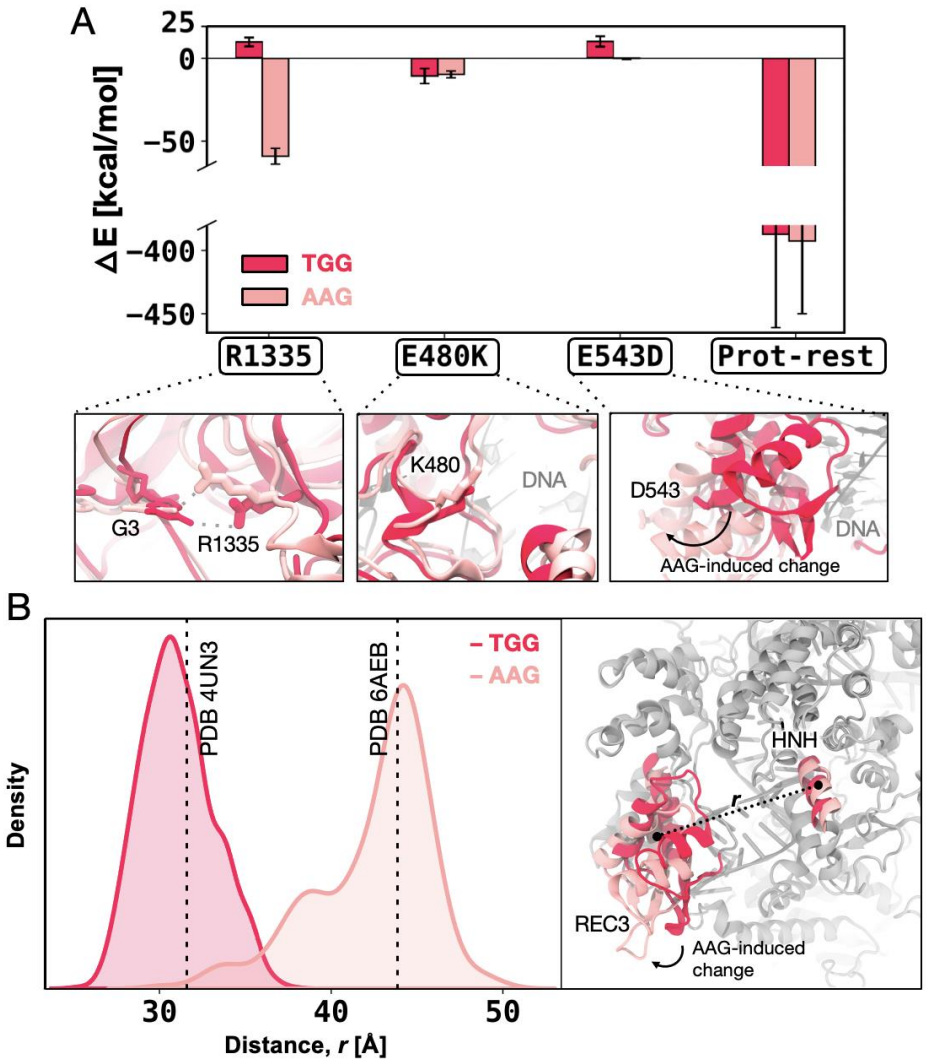

Figure 5. Enthalpic contribution to the DNA binding free energy for the SpCas9 → xCas9$_1$ transformation. A. Enthalpic contribution to the ΔΔG of DNA binding while transitioning from SpCas9 to xCas9$_1$ in the presence of the TGG (red) and AAG (pink) PAM sequences, computed as the average changes in the interaction energy (ΔE) between selected amino acid residues and the DNA. The ΔE is also computed as the overall interaction energy between the rest of the protein (Prot-rest, without the selected residues) and the DNA. Error bars were computed by averaging the results from different segments of the given trajectory (details in Materials and Methods). **B.** Probability density of the distance (r) between the centers of mass (COM) of the REC3 and HNH domains from molecular dynamics simulations of SpCas9 (PDB 4UN3[1]) incorporating the xCas9 mutations in the presence of TGG (red) and

AAG (pink) PAM sequences. Data from ~6 μs of aggregate sampling for each system. The values of the distance r in the X-ray structures of SpCas9 (PDB 4UN3[1], 31.6 Å) and xCas9 bound to AAG (PDB 6AEB[3], 43.8 Å) are indicated using a vertical dashed bar. The statistical significance between the two distributions was evaluated using Z-score statistics with a two-tailed hypothesis (P-value was less than 0.0001). The distance (r) between the REC3 and HNH COMs is indicated on the three-dimensional structure of Cas9 (right), highlighting the AAG-induced conformational change using an arrow. Figure 5–figure supplements 1-3.

To monitor the conformational change of REC3, we computed the centres of mass (COM) distance between REC3 and HNH domains from the obtained MD trajectories (Figure 5b). The simulations reveal that in the presence of TGG, the REC3 domain maintains the configuration observed in the 4UN3 X-ray structure. On the other hand, in the presence of AAG, REC3 exhibits a notable opening of ~12/13 Å with respect to the 4UN3 X-ray structure, reaching a conformation that is similar to the X-ray structure of the AAG-bound xCas9 (PDB 6AEB). Together, our computations show that the REC3 conformational change is attributed to alternative PAM binding rather than the xCas9 amino acid substitutions, and occurs in the early cycle of directed evolution, suggesting that PAM binding acts as a positive allosteric effector of the REC3 function(Nierzwicki et al., 2020; Palermo et al., 2017; Sternberg et al., 2015; Zuo and Liu, 2020).

Analysis of the enthalpic contribution to the ΔΔG upon the second evolution cycle (i.e., from $xCas9_1$ to $xCas9_2$) reveals that despite being distal from the DNA and PAM, the $xCas9_2$ mutations promote favourable interactions between R1335 and AAG (Figure 5–figure supplement 2). Upon the third evolution cycle (i.e., from $xCas9_2$ to $xCas9_3$), the enthalpic contribution to DNA binding does not report substantial differences for R1335 and R324L between TGG- and AAG-bound xCas9 (Figure 5–figure supplement 3). On the other hand, the overall interaction energy between the protein and the DNA is highly favourable for the TGG-bound $xCas9_3$, compared to the AAG-bound system. This is notable considering that the overall ΔE for protein DNA binding between the TGG- and AAG-bound systems was negligible in the first cycle (Figure 5a), while more favourable for AAG in the second cycle (Figure 5–

figure supplement 2). This indicates that the third evolution cycle highly adapts to binding TGG, resulting in an energetically favourable DNA binding compared to AAG. This observation contributes to clarifying how xCas9 expands PAM recognition while also improving recognition of the canonical TGG PAM compared to SpCas9.

DISCUSSION

Here, we establish the molecular determinants underlying the enhanced DNA targeting capabilities of xCas9 by integrating extensive molecular dynamics simulations, alchemical free energy calculations, and well-tempered metadynamics simulations. Our motivation stems from the observation that while structural studies demonstrate xCas9's binding to alternative PAM sequences such as AAG and GAT (Guo et al., 2019), they do not clarify how the enzyme expands the DNA targeting capability, while also improving the recognition of the canonical PAM.

We demonstrate that in SpCas9, the arginine dyad R1333 and R1335 within the PAM-interacting domain – particularly the rigidity of R1335 – enforces strict guanine selection, thereby restricting PAM compatibility to sequences closely aligned with the canonical NGG motif. In contrast, xCas9 introduces remarkable flexibility in R1335 through the E1219V mutation, enabling R1335 to effectively interact with both PAM nucleobases and backbone (Figure 2a), thereby enhancing the recognition of a broader range of PAM sequences. However, effective interactions with the PAM nucleotides are required for successful PAM recognition. In this respect, it is notable that most recognized PAM sequences contain at least one guanine, which readily interacts with arginine(Hossain et al., 2023; Luscombe, 2001). Conversely, when faced with ignored PAM sequences, the arginine dyad shifts away from the PAM nucleotides, indicating selective targeting (Figure 2f). The flexibility of R1335 not only facilitates binding to non-canonical PAMs such as AAG and GAT but also confers a pronounced entropic preference for the canonical TGG PAM over SpCas9, as demonstrated through free energy simulations (Figure 3). These well-tempered metadynamics simulations further reveal that the adaptability of R1335 interactions in xCas9 reduces the

entropic penalty during DNA binding, providing a mechanistic basis for its enhanced recognition of TGG observed experimentally.

xCas9 was developed through three cycles of directed evolution, progressively introducing a total of seven amino acid substitutions with respect to SpCas9. We evaluated the contribution of these substitutions to expanded PAM recognition by incorporating them into SpCas9 in a stepwise fashion and assessed their effects on DNA binding. We found that the directed evolution process tunes the specificity of xCas9 for non-canonical PAM sequences by progressively enhancing the DNA binding affinity. Specifically, the substitutions incorporated during the first evolution cycle primarily improved DNA binding with non-canonical PAMs, as evidenced by the substantial increase in binding free energy for AAG (Figure 4). Subsequent mutations in the second evolution cycle further stabilized these interactions, particularly with R1335, while the final mutation cycle significantly enhanced the binding affinity for the canonical TGG PAM, achieving an energetically favourable DNA binding configuration. These findings indicate that the expanded DNA targeting capability is established early in the directed evolution process, while the enhanced recognition of the canonical TGG PAM by xCas9, as compared to SpCas9, predominantly arises from mutations introduced during the third cycle. This is particularly significant in the context of the extensive research on directed evolution by Romero & Arnold(Romero and Arnold, 2009), Tawfik(Tokuriki et al., 2008), and Shoichet(Wang et al., 2002), who demonstrated that early substitutions in the evolution process commonly reshape functionality, resulting in improved or novel functions. This pattern is evident in the enhanced DNA binding affinity of xCas9 for the AAG PAM, which was achieved in the initial evolution cycle. Conversely, structural mutations that stabilize the complex tend to occur in later stages, as exemplified by the substantial improvement in DNA binding affinity for the canonical TGG PAM-bound xCas9 observed during the third cycle of evolution. This stepwise improvement underscores the cumulative effect of directed evolution in expanding PAM recognition while also fine-tuning the recognition of the canonical PAM sequence.

The recognition of alternative PAMs also triggers a conformational change in the REC3 domain (Figure 5), driven by AAG binding rather than the amino acid substitutions in xCas9. This conformational shift is observed upon AAG binding in the initial cycle of directed evolution and is maintained throughout the evolution of xCas9. The REC3 domain eventually adopts an open conformation similar to that seen in the X-ray structure of xCas9 bound to AAG (PDB 6AEB)(Guo et al., 2019). This observation suggests that the binding of alternative PAM sequences modifies the conformational dynamics of the distally located REC3 domain, implying an allosteric regulatory mechanism. This finding aligns with the concept of Cas9 functioning as an allosteric engine, where the binding of the canonical PAM sequence primes the protein for double-stranded DNA cleavage (Nierzwicki et al., 2020; Palermo et al., 2017; Sternberg et al., 2015; Zuo and Liu, 2020). The impact of alternative PAM sequences on this mechanism, particularly the changes in REC3 domain dynamics observed through the simulations (Figure 5b), underscores the need for novel experimental studies to accurately elucidate these molecular changes(East et al., 2020; Nierzwicki et al., 2021; Skeens et al., 2024).

When bound to TGG, xCas9 retains a closed conformation of the REC3 domain, akin to that observed in the X-ray structure of the TGG-bound SpCas9 (PDB 4UN3)(Anders et al., 2014). We recall that the third round of evolution significantly enhances DNA binding in the presence of TGG (Figure 4), indicating that a compact and energetically favourable conformation is essential for the efficient recognition of the canonical TGG PAM. xCas9 also reduces the entropic penalty associated with TGG binding at the PAM site compared to SpCas9 (Figure 3), suggesting that multiple factors synergistically contribute to xCas9's improved TGG recognition over SpCas9(Hu et al., 2018).

CONCLUSIONS

Here, we used extensive molecular dynamics and free energy simulations to establish the molecular determinants of xCas9's expanded PAM recognition, offering critical insights to broaden the

DNA targeting capability of the CRISPR-Cas9 genome editing tool. We demonstrate that while the *S. pyogenes* Cas9 (SpCas9) enforces a strict guanine selection through the rigidity of its arginine dyad, xCas9 expands PAM recognition by modulating the flexibility of R1335. This modulation allows xCas9 to shift away from ignored PAM sequences while establishing favorable interactions with recognized PAM nucleotides. By adjusting the flexibility of the arginine dyad, xCas9 also acquires an entropic preference for the canonical TGG PAM. Furthermore, xCas9 expands the DNA targeting capability in the early cycles of directed evolution, while the enhanced recognition of the canonical TGG PAM by xCas9, as compared to SpCas9, predominantly arises from mutations introduced during the third cycle. These outcomes elucidate how xCas9 expands PAM recognition while also improving the recognition of TGG compared to SpCas9(Hu et al., 2018). These findings have significant implications for the future development of CRISPR-Cas9 systems. Building on the insights from this study, future engineering efforts can harness the dynamic interactions between xCas9 and a broader range of genomic sequences to achieve enhanced specificity. Understanding the molecular determinants of PAM recognition will enable the rational design of Cas9 variants with tailored specificities, thereby expanding the range of targetable sequences in the genome, a key priority for enhancing genome editing technologies(Anzalone et al., 2020; Pacesa et al., 2024; Wang and Doudna, 2023). This could greatly enhance the versatility and precision of genome editing applications, from basic research to therapeutic interventions.

MATERIALS AND METHODS

Structural Models. Molecular simulations have been based on the *S. pyogenes* Cas9 (SpCas9) and its variant xCas9 3.7 (hereafter referred to as xCas9) bound to several PAM sequences. The wild-type SpCas9 bound to the 5'-TGG-3' canonical PAM has been based on the X-ray crystallographic structure (PDB 4UN3)(Anders et al., 2014), solved at 2.59 Å resolution. Three systems of xCas9 bound to PAM sequences that are recognized were based on the X-ray structures of xCas9 including 5'-TGG-3' (PDB 6K4P(Chen et al., 2019), solved at 2.90 Å resolution), 5'-GAT-3' (PDB 6AEG(Guo et al., 2019), solved

at 2.70 Å resolution) and 5'-AAG-3' (PDB 6AEB solved at 3.00 Å resolution(Guo et al., 2019)). xCas9 was also simulated bound to PAM sequences that are not recognized (5'-ATC-3', 5'-TTA-3', and 5'-CCT-3'). These systems were based on the PDB 6AEB(Guo et al., 2019). Two additional simulation systems were considered to examine the effect of alternative PAM binding on the conformational dynamics of Cas9. In detail, starting from the X-ray structure of SpCas9 (PDB 4UN3(Anders et al., 2014)), xCas9 mutations were introduced in the presence of both TGG and AAG PAM sequences. All simulation systems have been embedded in explicit waters, adding Na^+ and Cl^- counterions to provide physiological ionic strength (0.15 M), and reaching ~340,000 atoms each (periodic box ~148.5 x 185.0 x 125.1 $Å^3$).

Molecular dynamics simulations. Molecular dynamics (MD) simulations were performed through a simulation protocol tailored for protein/nucleic acid complexes(Sinha et al., 2023), which we also employed in studies of genome editing systems(Pacesa et al., 2022; Saha et al., 2024; Sinha et al., 2024). We used the Amber ff19SB force field(Tian et al., 2020), incorporating the OL15(Galindo-Murillo et al., 2016) corrections for DNA and the OL3(Banas et al., 2010; Zgarbova et al., 2011) corrections for RNA. The TIP3P model was employed for explicit water molecules(Jorgensen et al., 1983). The Li & Merz 12-6-4 model was used for Mg^{2+} ions(Li and Merz, 2014). All simulations were performed using the Gromacs 2018.5 code(Abraham et al., 2015). The simulations were conducted in the NPT ensemble using the leap-frog algorithm for integrating the equations of motion, with a time step of 2 fs. The temperature was maintained at 310 K using the v-rescale thermostat(Bussi et al., 2007) with a time constant of 0.1 ps. The pressure was controlled at 1 bar with the isotropic Parrinello–Rahman barostat(Parrinello and Rahman, 1981). Periodic boundary conditions were applied in three dimensions. Long-range electrostatic interactions were computed using the particle mesh Ewald (PME) approach(York et al., 1993) with a real-space cut-off of 1.2 nm and a Fourier grid spacing of 0.12 nm. Van der Waals interactions were modelled using the Lennard-Jones potential with a cut-off of 1.2 nm and a switching distance of 1 nm. The P-LINCS(Hess, 2008) algorithm was used to constrain bond lengths of the protein, DNA and RNA, and the

SETTLE(Miyamoto and Kollman, 1992) algorithm was used to preserve the water molecules' geometry. Production runs were carried out collecting ~1.5 μs for each system and in four replicates. This resulted in a collective ensemble of ~6 μs for each system, totalling ~54 μs of simulated runs (i.e., ~6 μs for nine simulation systems).

Well-tempered metadynamics. To explore the molecular energetics underlying the interactions of R1335 and exhaustively sample the conformational space, we conducted two-dimensional well-tempered metadynamics simulations. Metadynamics is a non-equilibrium simulation method enabling the exploration of higher-dimensional free energy surfaces by reconstructing the probability distribution as a function of a few predefined collective variables (CVs)(Bussi and Laio, 2020; Laio and Parrinello, 2002). In metadynamics, the system's evolution is biased by a history-dependent potential, constructed through the cumulative addition of Gaussian functions deposited along the trajectory in the CVs space. As this bias potential compensates for the underlying free energy surface, the latter can be computed as a function of the CVs. Here, we performed well-tempered metadynamics(Barducci et al., 2008), an improvement that enhances the CV-space exploration by introducing a tuneable parameter ΔT that regulates the bias potential. We further improved the sampling through multiple walkers(Minoukadeh et al., 2010), to ensure efficient scan across the specified reaction CVs. In well-tempered metadynamics, the bias deposition rate decreases over the course of the simulation, which is achieved by using a modified expression for the bias potential, $V(s,t)$:

$$V(s,t) = \sum_{t'=0,t_G,2t_G,\ldots}^{t'<t} \omega t_G e^{-V(s(q(t'),t'))/\Delta T} e^{-\sum_{i=1}^{2}\left[\left(s_i(q)-s_i\left(q(t')\right)\right)^2/2\sigma_i^2\right]} \quad (1)$$

where ω is the deposition rate and t_G is the and deposition stride of the Gaussian hills. Here, each simulation applied a biasing potential with the deposition rate (ω) and deposition stride (t_G) of the Gaussian hills of 0.239 kcal/mol/ps (1.0 kJ/mol/ps) and 10 ps, respectively. The bias factor $(T + \Delta T)/T$ was set to 4 at 300 K and the free energy, $F(s,t)$, was then computed as:

$$F(s,t) = -\frac{T+\Delta T}{\Delta T}(V(s,t) - C(T)) \quad (2)$$

where ΔT is the difference between the temperature of the CV and the simulation temperature T. The bias potential is grown as the sum of the Gaussian hills deposited along the CV space, with the sampling of the CV space being controlled by the tunable parameter ΔT. All well-tempered metadynamics simulations were started with a well-equilibrated structure generated from unbiased MD simulation.

Well-tempered metadynamics was carried out using two CVs for each simulation. To examine the interactions between R1335 and either E1219 or the G3 nucleobase in SpCas9 (Figure 3a), the free energy landscape was studied along the distances between the canters of mass (COM) of the R1335 guanidinium group and either the carboxylic group of E1219 (CV1) or the G3 functional group atoms exposed in the major groove (O6 and N7; CV2). To compare the interactions of R1335 in SpCas9 and xCas9 (Fig. 3b-c, main text) and evaluate its preference for binding either the G3 nucleobase or the DNA backbone, the CVs the CVs were defined as the COM distances between the R1335 guanidinium group and either the backbone phosphate group atoms (OP1, OP2, P; CV1) or the COM of the G3 functional group atoms (O6 and N7; CV2). To restrict the sampled range of coordinates, one-sided harmonic potentials with a force constant of 3500 kJ/mol·nm2 were employed, limiting the range of the collective variables to 0.3–1.2 nm. Each well-tempered metadynamics simulation was carried out for ~1 μs, reaching converged free energy surfaces (Figure 3–figure supplement 1-3). Well-tempered metadynamics simulations were performed using the Gromacs 2018.5 code(Abraham et al., 2015) and the open-source, community-developed PLUMED library(Tribello et al., 2014).

Alchemical free energy calculations. Alchemical free energy calculations were performed to assess the impact of mutations introduced into SpCas9 during directed evolution, and resulting in xCas9, on the DNA binding affinity. This involved determining the binding free energy difference (ΔΔG) between SpCas9 and its xCas9 mutants. We considered three xCas9 mutants: $xCas9_1$, $xCas9_2$, and $xCas9_3$, which emerged through three successive cycles of directed evolution(Hu et al., 2018). In detail, starting from the X-ray structure of SpCas9 (PDB 4UN3)[1], the E480K, E543D, and E1219 mutations were introduced

during the first evolution cycle (yielding $xCas9_1$); A262T, S409I, and M694I arose in the second cycle ($xCas9_2$); and R324L was introduced in the third cycle ($xCas9_3$).

Alchemical free energy calculations were carried out to compute the relative ΔΔG of binding in SpCas9 while introducing the xCas9 mutations in three steps, moving from SpCas9 to $xCas9_1$, from $xCas9_1$ to $xCas9_2$, and from $xCas9_2$ to $xCas9_3$. This approach involved defining two end states, commonly referred to as "state A" ($\lambda = 0$; e.g., SpCas9) and "state B" ($\lambda = 1$; e.g., $xCas9_1$), using different molecular topologies to represent the initial and final states of a chemical process. We utilized a thermodynamic cycle (Figure 4–figure supplement 1) enabling us to compute the free energies associated with the "alchemical" transformation of specific amino acid residues in the absence (ΔG_{m1}) or presence (ΔG_{m2}) of the DNA substrate bound to the respective Cas9. This transformation was achieved by simulating the system independently for various values of the scaling parameter λ (referred to as λ-windows) ranging from 0 to 1, to linearly interpolate between the potential energy functions of the physical end states. Due to the need for multiple amino acid mutations in each transition (e.g., from SpCas9 to xCas9), a large number of λ-windows were required, significantly increasing the computational cost. In detail, the distribution of the λ windows was optimized using a gradient descent algorithm to maximize the probabilities of exchange between adjacent states (https://gitlab.com/KomBioMol/converge_lambdas)(Wieczor and Czub, 2022). The neighbouring λ-windows were allowed to exchange their configurations every 0.5 ps according to the Metropolis criterion, and the values of λ were optimized to achieve the acceptance rate of at least 10%. Since each intermediate λ state is technically a hybrid between the A and B end point states, we generated their dual coordinates and topologies using the PMX server (http://pmx.mpibpc.mpg.de/)(Gapsys and de Groot, 2017). The relative free energy changes (i.e., ΔΔG) for the transformation were computed using the Multistate Bennett Acceptance Ratio (MBAR) method(Matsunaga et al., 2022; Shirts and Chodera, 2008) to integrate the free energies over the different λ values(Klimovich et al., 2015). Each system underwent simulation for a minimum of ~80 ns in each λ-window, collecting ~18 μs of simulated runs, until

reasonable convergence of ΔΔG was attained (Figure 4–figure supplement 2). Hamiltonian-replica exchange(Hritz and Oostenbrink, 2008) was used in each simulation, with exchanges attempted every 1000 steps (or 2 ps), to enhance the sampling efficiency and ensure adequate overlap between neighbouring windows. The enthalpic contributions to the DNA binding free energy (ΔΔG) were computed as the average changes in the interaction energy (ΔE) between selected amino acid residues and the DNA. In detail, the ΔE values were calculated from the FEP trajectories, considering only the physical states (i.e., $\lambda = 0$ and $\lambda = 1$) and discarding the first ~10% frames as the equilibration phase. The values presented in Figure 5a and Figure 5–figure supplements 2-3 represent the average over the trajectory. The error estimation of the average was determined based on block averages over five blocks using the Gromacs 2018.5 energy module(Abraham et al., 2015).

Analysis of structural data. The probability for R1333 and R1335 to interact with the PAM nucleobases (PAM NB), the PAM backbone (PAM BB), and non-PAM nucleotides (non-PAM) (Figure 2b) was computed as follows. The contacts between two arginine residues (R1333 and R1335) and the DNA duplex were defined based the two criteria: (1) the distance between the centre of mass (COM) of the arginine guanidinium group and either of the COM of the DNA bases' heavy atoms or the COM of the backbone phosphate groups and (2) the interaction energies between the arginine residues and the DNA bases or the backbone phosphate groups. As possible contacts, each base–guanidine and phosphate–guanidine pairs were selected for which the COM distance was below 0.6 nm and 0.5 nm, respectively, in at least 5% of the cumulative trajectories. For these pairs, interaction energies between arginine residues and either the DNA bases or the backbone phosphates were computed. As a criterion to define effective contacts, the interaction strength of at least 100 kJ/mol and 350 kJ/mol was used for the arginine–base and arginine–phosphate pairs, respectively. These criteria allowed distinguishing the pairs that form efficient interactions from the pairs that are close to each other, but do not properly interact (such as contacts between the arginine guanidinium groups and bases adjacent to the properly interacting nucleotide). The interaction frequencies (Figure 2b) were computed through the following process. Initially, we classified

whether a given interaction occurred using predefined energy and distance criteria (as described above). This classification yielded binary data, which we treated as a Bernoulli distribution to compute the variance of the interaction frequencies. Next, to estimate the number of independent data points, we calculated the autocorrelation of the interaction energy data. For each interaction, we utilized the largest autocorrelation time derived from four independent simulation replicates (of ~1.5 μs each, totalling ~6 μs per system) to determine the number of effective independent samples. Finally, we used the computed variance and number of independent samples to compute the error of each interaction mean. This methodology ensured a robust estimation of the mean interaction frequency and its associated error.

Hydrogen bonds (Hb) between the arginine side chains and the PAM NB, BB, and non-PAM nucleotides were analysed using the Gromacs 2018.5 hbond analysis tool (Abraham et al., 2015). The standard geometrical criteria applied were a cut-off value of 3.5 Å for the acceptor-donor distance and 30° for the hydrogen-donor-acceptor angle. Hb frequencies (Figure 2d) were calculated through a binary classification of the presence or absence of hydrogen bonds. The mean frequency was determined as the ratio of the number of Hbs in a given trajectory to the total number of frames. Data normalization was achieved using a standard method, which involved dividing by the sum of all elements in the given dataset (e.g., Hbs for R1333 with PAM-NB, PAM-BB, and non-PAM nucleotides). This normalization ensured comparability across different interaction types or studied systems. These calculations were performed considering the overall ensemble of ~6 μs for each system (Figure 2d), as well as for the independent simulation replicates of ~1.5 μs each (Figure 2–figure supplement 1).

FIGURE SUPPLEMENTS CAPTIONS

Figure 2–figure supplement 1. Frequencies of hydrogen bond formation across four simulation replicates. A. Frequencies of hydrogen bond formation between the R1333 and R1335 arginine side chains and the PAM nucleobases (NB), PAM backbone (BB) and non-PAM nucleotides, computed for SpCas9 bound to TGG (**A**), and xCas9 bound to PAM sequences that are recognized (TTG (**B**), AAG (**C**), GAT (**D**)) and that are ignored (CCT (**E**), TTA (**F**), ATC (**G**)). Hydrogen bonds are computed using an acceptor-donor distance and angles of 3.5 Å and 30°, respectively (details in Materials and Methods). For each simulated system, data are reported for each simulation replicate of ~1.5 μs each. The analysis on the overall ensemble of ~6 μs for each system is reported in Figure 2d.

Figure 3–figure supplement 1. Convergence of well-tempered metadynamics simulations characterizing the preference of R1335 for binding either the G3 nucleobase or E1219 in SpCas9. The two-dimensional free energy surfaces (FES) were computed along two collective variables (CVs): the distances between the centres of mass (COM) of the R1335 guanidine and the E1219 carboxylic group (CV1) or the G3 functional group atoms exposed in the major groove (O6 and N7, CV2). The FES were computed over non-overlapping windows of the trajectory, specifically covering the intervals 0–100 ns, 100–300 ns, 300–400 ns, 400–500 ns, 500–600 ns, 600–700 ns, 700–800 ns, 800–900 ns, and 900–1000 ns of well-tempered metadynamics.

Figure 3–figure supplement 2. Convergence of well-tempered metadynamics simulations characterizing the binding of R1335 to G3 and the DNA backbone in SpCas9. The two-dimensional free energy surfaces (FES) were computed along two collective variables (CVs): the distances between the R1335 guanidine and either the backbone phosphate group atoms (CV1) or the COM of the G3 functional group atoms exposed in the major groove (O6 and N7, CV2). The FES were computed over non-overlapping windows of the trajectory, specifically covering the intervals 0–100 ns, 100–300 ns, 300–400 ns, 400–500 ns, 500–600 ns, 600–700 ns, 700–800 ns, 800–900 ns, and 900–1000 ns of well-tempered metadynamics.

Figure 3–figure supplement 3. Convergence of well-tempered metadynamics simulations characterizing the binding of R1335 to G3 and the DNA backbone in xCas9. The two-dimensional free energy surfaces (FES) were computed along two collective variables (CVs): the distances between the R1335 guanidine and either the backbone phosphate group atoms (CV1) or the COM of the G3 functional group atoms exposed in the major groove (O6 and N7, CV2). The FES were computed over non-overlapping windows of the trajectory, specifically covering the intervals 0–100 ns, 100–300 ns, 300–400 ns, 400–500 ns, 500–600 ns, 600–700 ns, 700–800 ns, 800–900 ns, and 900–1000 ns of well-tempered metadynamics.

Figure 4–figure supplement 1. Thermodynamic cycle. Thermodynamic cycle used to calculate the difference in the DNA-binding free energy (ΔΔG) between Cas9 and its mutants, representing the contribution of each group of mutations (e.g., E480K, E543D and E1219V in first cycle, moving from SpCas9 to $xCas9_1$) to the DNA binding affinity (Cas9-RNA complex in light gray, DNA in charcoal, PAM highlighted in red). ΔΔG is obtained by subtracting the free energies of "alchemically" transforming amino acid residues in the absence and the presence of DNA (ΔG_{m1} and ΔG_{m2}, respectively).

Figure 4–figure supplement 2. Convergence of alchemical free energy calculations. Convergence of the difference in the DNA-binding free energy (ΔΔG) between SpCas9 and its xCas9 mutants (i.e., for the transformations of SpCas9 → $xCas9_1$ (**A**), $xCas9_1$ → $xCas9_2$ (**B**), $xCas9_2$ → $xCas9_3$ (**c**)) with the length of the alchemical simulations, in the presence of a TGG (red) or an AAG (pink) PAM sequence.

Figure 5–figure supplement 1. Comparison of existing structural data. Superimposition of the X-ray crystal structure of the TGG-bound SpCas9 (PDB 4UN3[1]) with the TGG-bound xCas9 (PDB 6K4P[2]) (**A**), and with the AAG-bound xCas9 (PDB 6AEB[3]) (**B**). There is no substantial conformational difference between SpCas9 and xCas9 bound to a TGG PAM (backbone RMSD = 0.80 Å). When xCas9 binds AAG, the REC3 domain displays an opening (outbound movement) with respect to SpCas9 (resulting in a protein backbone RMSD = 3.31 Å)[3]. This suggests that the binding of PAM induces a positive allosteric effect, in line with the notion that PAM acts as an allosteric effector of Cas9 dynamics[34–37].

Figure 5–figure supplement 2. Enthalpic contribution to the DNA binding free energy for the $xCas9_1 \rightarrow xCas9_2$ transformation. Enthalpic contribution to the DNA-binding free energy (ΔΔG) while transitioning from $xCas9_1$ to $xCas9_2$ in the presence of the TGG (red) and AAG (pink) PAM sequences, computed as the average changes in the interaction energy (ΔE) between selected amino acid residues and the DNA. The ΔE is also computed as the overall interaction energy between the rest of the protein (Prot-rest, without the selected residues) and the DNA. Error bars were computed by averaging the results from different segments of the given trajectory (details in Materials and Methods). When transitioning from $xCas9_1$ to $xCas9_2$, R1335 provides an increased enthalpic contribution in the presence of the AAG PAM sequence, which facilitates its recognition. The M694I mutation reduces the unfavourable contribution in the presence of the AAG. The overall interaction energy for the rest of the protein (prot-rest) is more favourable for the AAG-bound system.

Figure 5–figure supplement 3. Enthalpic contribution to the DNA binding free energy for the $xCas9_2 \rightarrow xCas9_3$ transformation. Enthalpic contribution to the DNA-binding free energy (ΔΔG) while transitioning from $xCas9_2$ to $xCas9_3$ in the presence of the TGG (red) and AAG (pink) PAM sequences, computed as the average changes in the interaction energy (ΔE) between selected amino acid residues and the DNA. The ΔE is also computed as the overall interaction energy between the rest of the protein (Prot-rest, without the selected residues) and the DNA. Error bars were computed by averaging the results from different segments of the given trajectory (details in Materials and Methods). In this transition, no significant difference is observed in the ΔE for R1335 and R324L in the presence of AAG or TGG PAM sequence. The overall interaction energy between the rest of the protein and DNA (prot-rest) is highly favourable in the presence of the TGG PAM sequence. This is notable considering that the overall ΔE

between the TGG- and AAG-bound systems was negligible in the first cycle (Figure 5a), while more favourable for AAG in the second cycle (Figure 5–figure supplement 2). This indicates that the third evolution cycle highly adapts to binding TGG (further discussion in the main text).

ACKNOWLEDGMENTS

We thank Dr. Pablo R. Arantes for useful discussions. This material is based upon work supported by the National Institutes of Health (Grant No. R01GM141329 to GP) and the National Science Foundation (Grant No. CHE-2144823 to GP). KAH and MO acknowledge support by EU-HPC, BioExcel 101093290, Horizon-EU MDDB 101094561 and Spanish MCIIN PDI2021-122478NB-I00. This work used Expanse at the San Diego Supercomputing Center through allocation MCB160059 and Bridges2 at the Pittsburgh Supercomputer Center through allocation BIO230007 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation supports grants #2138259, #2138286, #2138307, #2137603, and #2138296.

COMPETING INTERESTS

The authors declare no competing financial interest.

DATA AVAILABILITY

The files from the molecular dynamics simulations, including trajectory and visualization data, can be accessed on Figshare using the identifier 10.6084/m9.figshare.26767465.

REFERENCES

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**:19–25.

Anders C, Niewoehner O, Duerst A, Jinek M. 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**:569–573.

Anzalone A V., Koblan LW, Liu DR. 2020. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat Biotechnol*. **38**:824–844.

Banas P, Hollas D, Zgarbova M, Jurecka P, Orozco M, Cheatham TE 3rd, Sponer J, Otyepka M. 2010. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J Chem Theor Comput* **6**:3836–3849.

Barducci A, Bussi G, Parrinello M. 2008. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys Rev Lett* **100**:020603.

Bhattacharya S, Satpati P. 2024. Why Does the E1219V Mutation Expand T-Rich PAM Recognition in Cas9 from *Streptococcus pyogenes*? *J Chem Inf Model* **64**:3237–3247.

Bussi G, Donadio D, Parrinello M. 2007. Canonical sampling through velocity rescaling. *Journal of Chemical Physics* **126**:014101.

Bussi G, Laio A. 2020. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics* **2**:200–212.

Chen W, Zhang H, Zhang Y, Wang Y, Gan J, Ji Q. 2019. Molecular basis for the PAM expansion and fidelity enhancement of an evolved Cas9 nuclease. *PLoS Biol* **17**:e3000496.

East KW, Newton JC, Morzan UN, Narkhede YB, Acharya A, Skeens E, Jogl G, Batista VS, Palermo G, Lisi GP. 2020. Allosteric Motions of the CRISPR-Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics. *J Am Chem Soc* **142**:1348–1358.

Galindo-Murillo R, Robertson JC, Zgarbová M, Šponer J, Otyepka M, Jurečka P, Cheatham TE. 2016. Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* **12**:4114–4127.

Gapsys V, de Groot BL. 2017. pmx Webserver: A User Friendly Interface for Alchemistry. *J Chem Inf Model* **57**:109–114.

Guo M, Ren K, Zhu Y, Tang Z, Wang Y, Zhang B, Huang Z. 2019. Structural insights into a high fidelity variant of SpCas9. *Cell Res* **29**:183–192.

Hess B. 2008. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J Chem Theory Comput* **4**:116–122.

Hossain KA, Kogut M, Słabońska J, Sappati S, Wieczór M, Czub J. 2023. How acidic amino acid residues facilitate DNA target site selection. *Proceedings of the National Academy of Sciences* **120**.

Hritz J, Oostenbrink C. 2008. Hamiltonian replica exchange molecular dynamics using soft-core interactions. *J Chem Phys* **128**:144121.

Hu JH, Miller SM, Geurts MH, Tang W, Chen L, Sun N, Zeina CM, Gao X, Rees HA, Lin Z, Liu DR. 2018. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature*. **5**:57–63.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**:816–821.

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**:926–935.

Klimovich P V., Shirts MR, Mobley DL. 2015. Guidelines for the analysis of free energy calculations. *J Comput Aided Mol Des* **29**:397–411.

Laio A, Parrinello M. 2002. Escaping free-energy minima. *Proc Natl Acad Sci U S A* **99**:12562–12566.

Li P, Merz KM. 2014. Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions. *J Chem Theory Comput* **10**:289–297.

Luscombe NM. 2001. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res* **29**:2860–2874.

Matsunaga Y, Kamiya M, Oshima H, Jung J, Ito S, Sugita Y. 2022. Use of multistate Bennett acceptance ratio method for free-energy calculations from enhanced sampling and free-energy perturbation. *Biophys Rev* **14**:1503–1512.

Minoukadeh K, Chipot C, Lelièvre T. 2010. Potential of Mean Force Calculations: A Multiple-Walker Adaptive Biasing Force Approach. *J Chem Theory Comput* **6**:1008–1017.

Miyamoto S, Kollman PA. 1992. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem* **13**:952–962.

Nierzwicki Ł, Arantes PR, Saha A, Palermo G. 2020. Establishing the Allosteric Mecanism in CRISPR-Cas9. *WIREs Comput Mol Sci* **11**:e1503.

Nierzwicki L, East KW, Morzan UN, Arantes PR, Batista VS, Lisi GP, Palermo G. 2021. Enhanced specificity mutations perturb allosteric signaling in CRISPR-Cas9. *Elife* **10**.

Pacesa M, Lin C-H, Cléry A, Saha A, Arantes P, Bargsten K, Irby M, Allain F, Palermo G, Cameron P, Donohoue P, Jinek M. 2022. Structural basis for Cas9 off-target activity. *Cell* **185**: 4067-4081.

Pacesa M, Pelea O, Jinek M. 2024. Past, present, and future of CRISPR genome editing technologies. *Cell* **187**:1076–1100.

Palermo G, Ricci CG, Fernando A, Basak R, Jinek M, Rivalta I, Batista VS, McCammon JA. 2017. Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9. *J Am Chem Soc* **139**:16028–16031.

Parrinello M, Rahman A. 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* **52**:7182–7190.

Romero PA, Arnold FH. 2009. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10**:866–876.

Saha A, Ahsan M, Arantes PR, Schmitz M, Chanez C, Jinek M, Palermo G. 2024. An alpha-helical lid guides the target DNA toward catalysis in CRISPR-Cas12a. *Nat Commun* **15**:1473.

Shirts MR, Chodera JD. 2008. Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys* **129**:124105.

Sinha S, Molina Vargas AM, Arantes PR, Patel A, O'Connell MR, Palermo G. 2024. Unveiling the RNA-mediated allosteric activation discloses functional hotspots in CRISPR-Cas13a. *Nucleic Acids Res* **52**:906–920.

Sinha S, Pindi C, Ahsan M, Arantes PR, Palermo G. 2023. Machines on Genes through the Computational Microscope. *J Chem Theory Comput* **19**:1945–1964.

Skeens E, Sinha S, Mohd A, D'Ordine AM, Jogl G, Palermo G, Lisi GP. 2024. High-Fidelity, Hyper-Accurate, and Evolved Mutants Rewire Atomic Level Communication in CRISPR-Cas9. *Sci Adv* **10**:eadl1045.

Sternberg SH, LaFrance B, Kaplan M, Doudna JA. 2015. Conformational control of DNA target cleavage by CRISPR–Cas9. *Nature* **527**:110–113.

Tian C, Kasavajhala K, Belfon KAA, Raguette L, Huang H, Migues AN, Bickel J, Wang Y, Pincay J, Wu Q, Simmerling C. 2020. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J Chem Theory Comput* **16**:528–552.

Tokuriki N, Stricher F, Serrano L, Tawfik DS. 2008. How Protein Stability and New Functions Trade Off. *PLoS Comput Biol* **4**:e1000002.

Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G. 2014. PLUMED 2: New feathers for an old bird. *Comput Phys Commun* **185**:604–613.

Wang JY, Doudna JA. 2023. CRISPR technology: A decade of genome editing is only the beginning. *Science (1979)* **379**.

Wang X, Minasov G, Shoichet BK. 2002. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. *J Mol Biol* **320**:85–95.

Wieczor M, Czub J. 2022. Gromologist: a Gromacs-Oriented Utility Library for Structure and Topology Manipulation. 2022, *ChemRxiv* doi:1026434/chemrxiv-2022-dhswc.

York DM, Darden TA, Pedersen LG. 1993. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald and truncated list methods. *J Chem Phys* **99**:8345–8348.

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

Zuo Z, Liu J. 2020. Allosteric regulation of CRISPR-Cas9 for DNA-targeting and cleavage. *Curr Opin Struct Biol* **62**:166–174.