

Machine-Learning-Inferred and Energy-Landscape-Guided Analyses Reveal Kinetic Determinants of CRISPR/Cas9 Gene Editing

Lei Jin  and Risi Liyanage 

Department of Physics and Astronomy, [University of Missouri](#), Columbia, Missouri 65211, USA

Dongsheng Duan 

Department of Molecular Microbiology and Immunology and Department of Neurology,
School of Medicine, [University of Missouri](#), Columbia, Missouri 65211, USA;

Department of Biomedical Sciences, College of Veterinary Medicine, [University of Missouri](#), Columbia, Missouri 65211, USA;
and Department of Bioengineering, [University of Missouri](#), Columbia, Missouri 65211, USA

Shi-Jie Chen*

Department of Physics and Astronomy and Department of Biochemistry, Institute of Data Sciences and Informatics,
[University of Missouri](#), Columbia, Missouri 65211, USA



(Received 26 March 2025; accepted 4 February 2026; published 4 March 2026)

The CRISPR-Cas9 nuclease system is currently the most powerful tool in genome engineering. Existing data-driven learning and thermal equilibrium-based modeling cannot capture the kinetic mechanism of the gene cleavage process. This study integrates large-scale genome editing data with a random forest approach—a supervised ensemble learning algorithm—and free energy landscape analysis to uncover the kinetic pathways and pivotal steps that govern cleavage efficiency. Major findings highlight a direct correlation between CRISPR cleavage efficiency and R-loop unfolding rate, establishing the R-loop unfolding rate as the critical factor in gene editing performance. The kinetic model significantly improves over other methods in predicting both on- and off-target gene editing efficiencies. Furthermore, energy landscape analysis reveals a novel three-state mechanism and identifies key nucleotides essential for gene editing. These insights lay the biophysical groundwork for new strategies to enhance CRISPR-Cas9 accuracy and efficiency, offering new avenues for single guide RNA design and target site selection.

DOI: [10.1103/rtxh-6fdl](https://doi.org/10.1103/rtxh-6fdl)

I. INTRODUCTION

The clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein 9 (Cas9) system, one of the most powerful tools in biological science [1], is widely employed in live-cell imaging [2,3], gene engineering [4,5], gene knockdown [6,7], and gene therapy [8]. In the CRISPR-Cas9 genome editing system, pairing a 100-nucleotide (nt) single guide RNA (sgRNA) and guided by a 20-nt target-recognition segment in the sgRNA [1], the *Streptococcus pyogenes* (Sp) Cas9 protein binds to the double-stranded (ds) DNA target by recognizing the protospacer adjacent motif (PAM), followed by base pairing between the 20-nt guide RNA segment and the DNA target. Subsequently, SpCas9 endonuclease cleaves the dsDNA with its RuvC and HNH domains [9–11]. The complementary sequence match between

the 20-nt target-recognition segment of the sgRNA and the target DNA strand constitutes an “on-target” cleavage site on the dsDNA [9–13]. Besides the on-target sites, partial complementarity between sgRNA and DNA sequences can also result in the binding of Cas9 and cleavage at such “off-target” sites [9–13]. Off-target cleavage can cause undesirable genetic alterations and should be avoided [14]. Maximizing on-target efficiency and minimizing off-target effects pose significant challenges in developing the CRISPR-Cas9 system [15–23].

Bioinformatics and machine-learning-based algorithms and thermodynamic models have been developed to evaluate on-target efficiency and identify potential off-target sites [24–45]. For instance, the *in vivo* library-on-library methodology sgRNA Scorer developed by Chari *et al.* [24] and sgRNA design rules Azimuth devised by Doench *et al.* [26] can be used to assess sgRNA on-target activity and help sgRNA selection; MIT (MIT_Zhang’s model) [28], CCTop (CRISPR-Cas9 target online predictor) [29], and cutting frequency determination (CFD) [26] can be used to evaluate off-target cleavage potential. However, these data-driven approaches rely solely on the available experimental datasets and do not address the physical mechanisms underlying the cleavage. The advanced gRNA design tool WU-CRISPR [33] integrates multiple thermodynamic features of the sgRNA with

*Contact author: chenshi@missouri.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](#) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

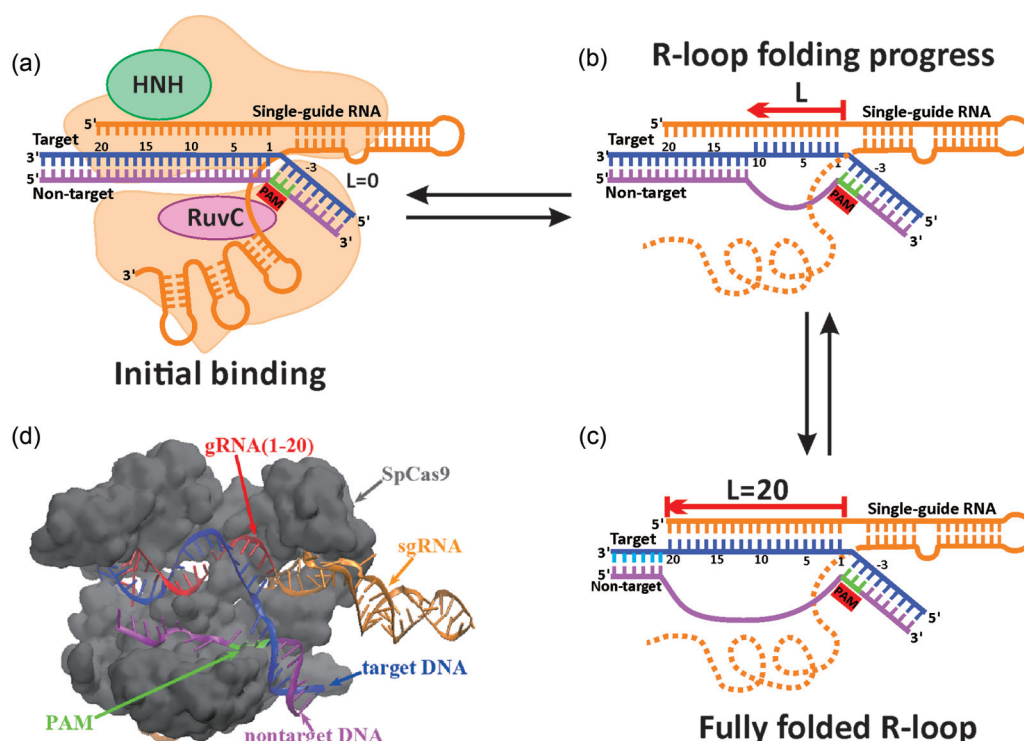


FIG. 1. [(a)–(c)] The R-loop folding process: Cas9-sgRNA complex recognizes the PAM site of the DNA followed by hybridization between the sgRNA and the target DNA strand and the displacement of the nontarget DNA strand. The progression of the R-loop formation can be described by L , the number of base pairs formed from the PAM end (equal to the length of the DNA-RNA duplex extended from the PAM end). We use R_L to denote the R-loop of a given L . (d) The structure of the Cas9-sgRNA-dsDNA ternary complex [10] (Protein Data Bank ID code 5F9R).

extended bioinformatic data in a statistical model, resulting in improved reliability for on-target evaluations. In contrast, we here explore the biophysical mechanisms for both on- and off-target cleavages. The free-energy-based physical approach uCRISPR [44] shows that on-target cleavage efficiency and off-target effects can be described in a unified framework by considering the folding stability of the R-loop, a complex consisting of the sgRNA-DNA duplex and the displaced single-stranded nontarget DNA in the structural context of the Cas9 protein. Although by accounting for the thermal equilibrium stability of the system, the physical model enables a better performance in evaluating the cleavage efficiency of both on- and off-target sites, the general Spearman rank correlation with the test set is below 0.50 and the predictions for off-target site are not always satisfactory.

The essential structures in the cleavage reaction of the CRISPR-Cas9 system [46,47] have been experimentally determined and a structure-based molecular mechanism of the cleavage reaction has been proposed [48], where the Cas9 protein recognizes sgRNA to form the Cas9-sgRNA complex, which then recognizes the PAM in the double-stranded DNA (dsDNA) target genome [49,50]. The PAM for Cas9 is generally a 5'-NGG-3' motif that can be recognized by the Cas9 through water-mediated hydrogen bonds in the minor groove of the dsDNA duplex [13,47], and the binding of the Cas9-sgRNA complex to the dsDNA at the PAM site results in local destabilization of the adjacent DNA base pairs, initiating the unwinding of the target dsDNA from the PAM-proximal end [48]. If the downstream

DNA sequence is complementary to the sgRNA spacer segment, the target strand of the unwound dsDNA hybridizes with the RNA strand, initiating a zipping-like reaction toward the PAM-distal end of the dsDNA. The process results in the formation of an R-loop with a 20-base pair (bp) RNA-DNA hybrid duplex and a single-stranded (ss) nontarget DNA strand [51] [see Figs. 1(a)–1(c)]. In most previous theories, the process was tied to the thermal equilibrium properties of the system, and cleavage occurred after the formation of the R-loop. As a result, the efficiency was related to the equilibrium population of the R-loop [52]. However, such theories are not consistent with off-target cleavage data, where for some systems, a less stable RNA-DNA duplex could cause a high cleavage efficiency. Moreover, recent kinetic experiments further challenged the correlation between Cas9-sgRNA-dsDNA thermal equilibrium binding affinity and off-target cleavage efficiency, with distinct tolerance for mismatches in the PAM-proximal and -distal regions [48,53,54]. The results of these kinetic experiments suggest that cleavage efficiency may be kinetically controlled rather than determined by binding equilibrium [53].

A general kinetic transition network was proposed based on experimental data [48], which suggests that the overall kinetic step for the Cas9-sgRNA complex binding to DNA and the subsequent formation of the R-loop is the slow kinetic step (e.g., $1.8\text{--}3.0\text{ s}^{-1}$ for some on-target binding) [48]. Depending on the solution environment and the concentration of the Cas9-sgRNA complex, the formation of the R-loop would eventually become the rate-limiting step in the ki-

netic network [48]. As part of a complex multiprocess kinetic network that involves the Cas9, sgRNA, and DNA association, dissociation, and structural changes and the possible degradation of Cas9-sgRNA in solution, the R-loop folding and unfolding kinetics as the critical kinetic steps in the network rather than the equilibrium stability would play an important role to determine the final outcome of the cleavage. Previous kinetic models [55,56] account only for the position rather than sequence dependence of kinetic parameters [56] and lack the full sequence dependence for the data-derived kinetic rates [53] and, thus, may not be accurate to capture the full kinetic mechanism.

Here, we leverage a large-scale cleavage database, an ensemble machine learning algorithm, kinetic Monte Carlo simulations, and free energy landscape analysis to explore the kinetic mechanism of CRISPR-Cas9 gene editing and to develop an accurate predictive model for cleavage efficiency. Different from previous thermodynamic and machine learning studies, our approach focuses on the R-loop kinetic folding of CRISPR-Cas9 to capture the underlying kinetic mechanism and its relationship with cleavage efficiency:

(1) From the large-scale experimental data, the ensemble learning identifies the R-loop unfolding rate as the primary determinant of CRISPR cleavage efficiency. This finding leads to a significant improvement in predictive accuracy for both on-target and off-target cleavage outcomes.

(2) Kinetic simulations and energy landscape analyses of large-scale data reveal a three-state kinetic mechanism for the R-loop, a novel physical scheme that cannot be inferred from current end-to-end machine learning approaches or thermal equilibrium analysis. The kinetic finding helps explain the weak correlation observed in experiments between thermal equilibrium binding affinity and off-target cleavage efficiency—a weak correlation that thermodynamic models have failed to account for.

These findings provide a biophysical foundation of potential new strategies for improving the accuracy and efficiency of CRISPR-Cas9 gene editing and the selection of cleavage targets.

II. RESULTS AND DISCUSSION

In the present study, large-scale simulations, combined with data-driven learning and kinetic analysis, suggest that the folding and unfolding kinetics of the R-loop—and the associated targeting process—follow a two-step (three-state) reaction pathway: folding of the seed region resulting in the formation of a kinetic intermediate state (I), followed by subsequent folding of the R-loop into the cleavage active state (AS). The intermediate state I corresponds to an R-loop containing $L = 10, 11, 12$ hybrid base pairs formed from the PAM-proximal site, whereas the active state (AS) comprises an ensemble of R-loop structures with at least $L \geq 17$ base pairs (formed from the PAM-proximal site; see Fig. 1). The I state represents a local minimum on the folding free energy landscape (see Fig. 3), while the AS corresponds to the critical state required to initiate cleavage by the RuvC/HNH domains of Cas9. Notably, the AS identified through our kinetic profiling differs from the previously proposed fully formed R-loop with $L = 20$ hybrid base pairs [44,52]. Our finding aligns

more closely with experimental observations [48,53,54]. In the following sections, we describe the detailed framework of our kinetic modeling approach, where the unfolding rate of the AS emerges as a key kinetic signature governing the cleavage efficiency.

A. Supervised learning indicates the unfolding rate of the active state as the most important determinant for cleavage efficiency

1. Kinetic determinant for cleavage efficiency

We identify the key feature determining the cleavage efficiency from a dataset comprising 14 538 data points (see Table S1 of the Supplemental Material (SM) [57]). Each data point is represented as $\{[S], \eta\}$, where the input data $[S] = \{S_{\text{target}}, s_1, s_2, \dots, s_{20}\}$ denotes the 20-nt target DNA sequence S_{target} and the guide RNA sequence with s_n represents the nucleotide at position $+n$ ($n = 1, 2, \dots, 20$) from the PAM, and the output η represents the experimentally measured cleavage efficiency. For each $[S]$, based on the uCRISPR energy parameters [44] and the Vfold RNA folding model [58–60], we employ kinetic Monte Carlo (KMC) simulations to predict the kinetic and thermodynamic properties of the R-loop system. The results provide a feature dataset, where the original data can be represented in the form $\{[F], \eta\}$, with the input data $[F] = \{f_1, f_2, \dots, f_m\}$ the m selected features for the corresponding $[S]$. These features $[F]$ include the folding free energy $\Delta G(R_L)$ (equal to the free energy difference between an R-loop structure R_L and the unfolded state $U = R_0$, where R_L is the R-loop with $L (= 1, 2, \dots, 20)$ RNA-DNA hybridized base pairs formed), the equilibrium population $P(R_L)$ of R_L , and the folding and unfolding rates $k_{R_L}^{(f)}$ and $k_{R_L}^{(u)}$ of R_L from and to the unfolded state U (see Methods).

Experiments have shown that the cleavage efficiency is sensitive to mismatches in the seed region ($L \leq 10$) and is tolerant of those in the PAM-distal positions [48,53,54]. Our KMC simulations indicate that the states R_{17} , R_{18} , R_{19} , and R_{20} with variable (cleavage-insensitive) PAM-distal ($L \geq 17$) base pairs occupy more than 95% of the total population in equilibrium and each state occupies a similar population (see Fig. S1(b) of the SM [57]). Therefore, in our feature space $[F]$, in addition to the unfolded state U , we define two special states: $I^* = R_{10}$ and $AS = R_{L \geq 17}$ (the active state). The various thermodynamic and kinetic features of the different R-loop states serve as candidate predictors for CRISPR cleavage efficiency. Figure 2(a) shows an example of ten selected features.

To evaluate the importance of each feature using the feature database $\{[F], \eta\}$, we employ the random forest (RF), a supervised machine learning algorithm that utilizes an ensemble of decision trees. Each tree in the ensemble is constructed from different choices of decisions, culminating in predictions of cleavage efficiency outcomes. The algorithm trains these decision trees so that, given a set of input feature values, the RF algorithm predicts the cleavage efficiency accurately with minimal ambiguity. RF leverages the random subspace method (feature bootstrap aggregation), allowing it to effectively select features from a diverse dataset while mitigating bias and overfitting. This approach enhances the robustness of feature selection across various data scenarios [61,62].

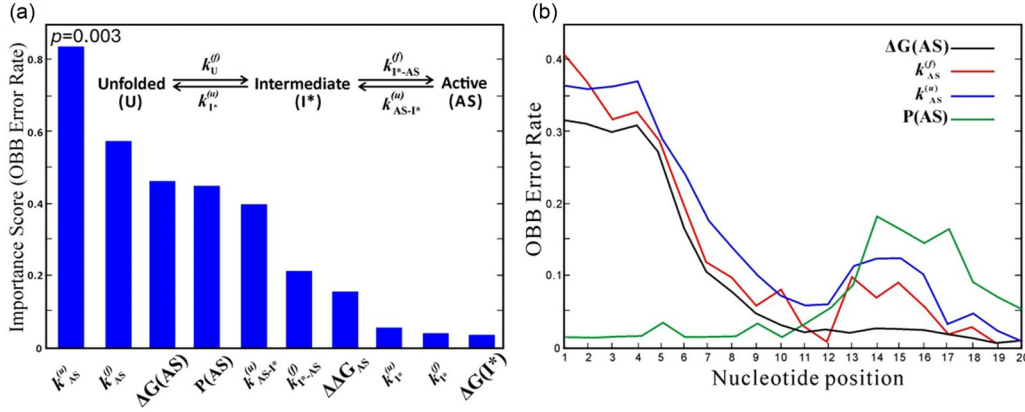


FIG. 2. (a) The importance score [out-of-bag error rate (OER)] for ten candidate feature variables. Here $k_x^{(u)}$, $k_x^{(f)}$, $\Delta G(x)$, and $k_{x_1 \rightarrow x_2}^{(u/f)}$ denote the unfolding rate ($x \rightarrow U$), the folding rate ($U \rightarrow x$), the folding free energy (relative to U) of state x , and the rate of transition from x_1 to x_2 , respectively, $\Delta \Delta G_{AS}$ is the stability of AS relative to the intermediate state I^* , $\Delta G(AS) - \Delta G(I^*)$, and $P(AS)$ is the equilibrium population of the AS. The p -values for $k_{AS}^{(u)}$ in the statistical test are labeled accordingly. (b) The position dependence of OER for the different features, where the energy parameter for a given position is treated as the input data and the result of a selected feature is the output data.

The learning power of a decision tree is measured by its test results on the out-of-bag (OOB) dataset $D_{OOB}([F])$, i.e., the data not included in the training of the decision tree. The importance of an input feature is determined through the calculation of the OOB error (OE) of an OOB dataset as the difference between the decision-tree-predicted output result, such as cleavage efficiency, and the experimentally determined value, averaged over an ensemble of trees (i.e., the forest). Second, we calculate $\overline{OE}$, the mean OE averaged over the $D_{OOB}([F])$ dataset. Third, we determine the OER of an input feature by randomly shuffling its values among the different data entries in $D_{OOB}([F])$ and compute the change of $\overline{OE}$. A larger OER indicates a higher importance. For details, see the Random Forest Approach section in the (SM) [57]. As shown in Fig. 2(a), the AS unfolding rate $k_{AS}^{(u)}$ is the most important feature to determine cleavage efficiency. To further validate the random forest feature ranking results, we perform further extended statistical tests across the training datasets. Permutation tests are performed for the highest ranked kinetic feature $k_{AS}^{(u)}$ and the four kinetically important nucleotide positions. The p -value is 0.003 for $k_{AS}^{(u)}$ and 0.007, 0.004, 0.001, and 0.002 for positions 13, 14, 15, and 16, respectively. The p -values are significantly lower than 0.05, suggesting a high statistical significance of the tested features. Friedman tests yield $p = 7 \times 10^{-15}$ for the kinetic or thermodynamic features and $p = 6 \times 10^{-14}$ for the nucleotide positions, respectively, suggesting the robustness of the importance ranks. See the Random Forest Approach section in the SM for details of the statistical test. In the following sections, we discuss the physical mechanism and the applicability of using $k_{AS}^{(u)}$ as the kinetic predictor for cleavage efficiency.

2. The critical nucleotides

An R-loop structure is stabilized not only by the hybrid base pairing and stacking between the sgRNA and the target strand of the DNA (tDNA) but also by the interactions between Cas9 and the nucleic acids, especially with the

nontarget strand of the DNA (ntDNA). These interactions influence the importance of the nucleotides at different positions for folding and unfolding kinetics, thereby affecting cleavage efficiency. To identify the most crucial nucleotides, we further apply RF to the dataset of the AS unfolding rate: $\{[S], k_{AS}^{(u)}\}$ with the sequence $[S]$ of the target DNA and the guide RNA as the input data and $k_{AS}^{(u)}$, the unfolding rate of AS, as the output data. To account for the influence from the interaction energies, we convert $[S]$ into the corresponding energy parameters: $\{[E], k_{AS}^{(u)}\}$, where $[E] = \{E_1, E_2, \dots, E_{20}\}$ (as the input data) denotes the (position-dependent) energy parameters with E_n being the energy parameter set for position $+n$ ($= 1, 2, \dots, 20$). Physically, E_n is the free energy change for the disruption of the dsDNA base pair followed by the formation of the DNA-RNA hybrid base pair at position n along with the interaction between Cas9 and the nontarget DNA strand. E_n is dependent on the chemical identity of the nucleotide at n and the neighboring sites. We compute the OER for each position n by randomly shuffling the E_n values among the different sequences in the dataset.

We can apply the OER computation above to investigate the position importance of other features. As shown in Fig. 2(b), the OER for the equilibrium folding stability ΔG_{AS} of the R-loop active state shows three phases: a slow decrease from +1 to +5, a steep decrease from +6 to +10, and a plateau from +11 to +20. The result indicates a monotonically decreasing importance of nucleotides as the position moves away from the PAM site. In contrast, the unfolding rate $k_{AS}^{(u)}$ of the R-loop active state shows a drastically different behavior: the OER decreases continually to a low trough in the seed region $L = [+1, +10]$, then rises to a relatively high plateau from +13 to +16 before decreasing in the region from +17 to +20. The result suggests the critical importance of nucleotides from +13 to +16 for the R-loop unfolding kinetic rate and hence the CRISPR-Cas9 cleavage efficiency. Furthermore, the position-dependent OER profile leads to two important regions for determining the cleavage efficiency: the seed region from +1 to +10 and the kinetically important distal region from +13 to +16. Since RNA-DNA hybridization

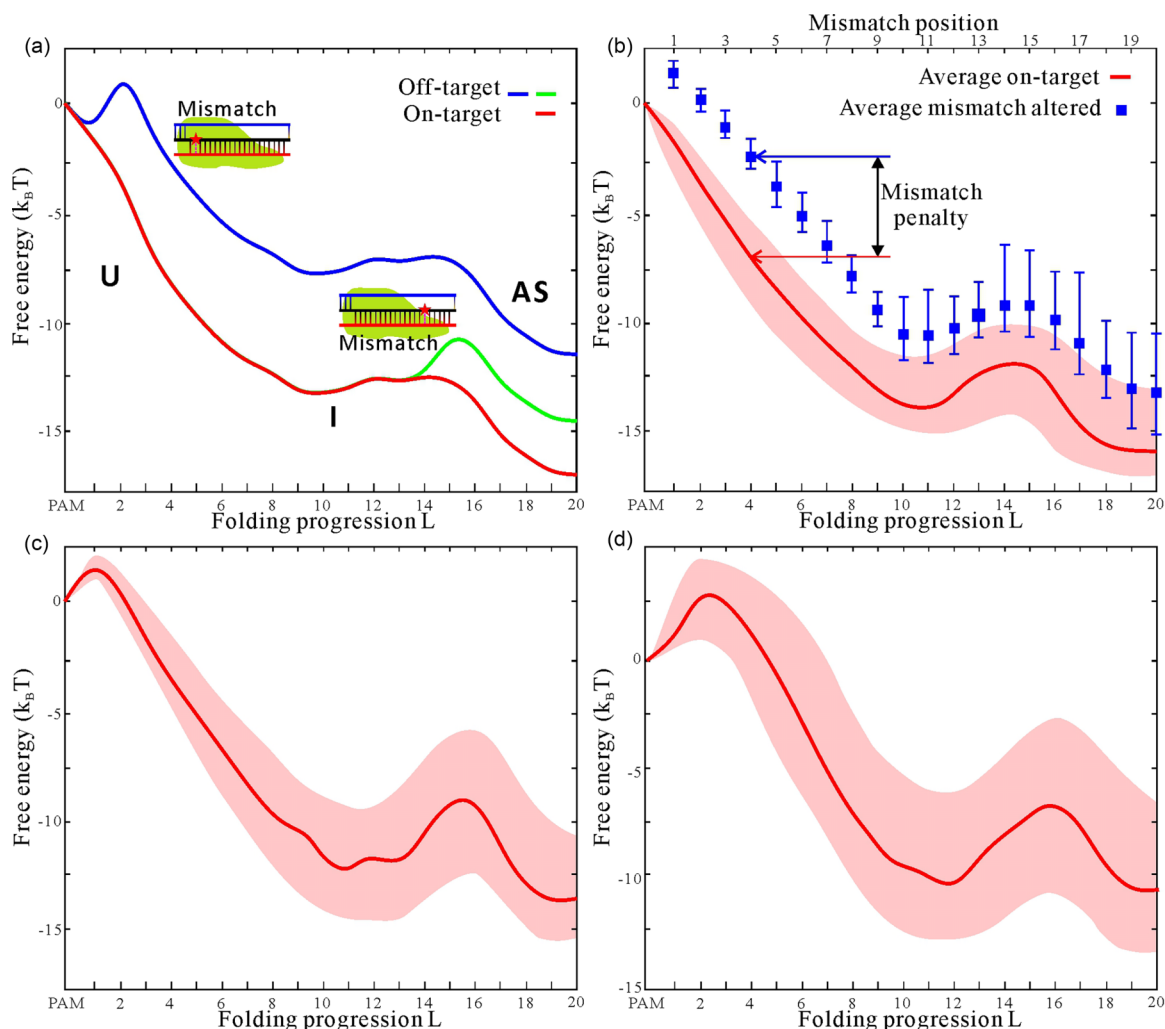


FIG. 3. The free energy landscape of the R-loop for (a) the on-target sequence EMX1 target 1 [28,40,68] (red line) and two off-target cases with a single mismatch in PAM-proximal (blue line) and PAM-distal (green line) regions [28,40,68], (b) the 7148 on-target cases (Table S1 [57]) [7,26,32], and EMX1 off-target with (c) a single mismatch [7,26,32] and (d) two consecutive mismatches [7,26,32]. Here a folding progression is characterized by L , the number of hybrid base pairs extended from the PAM site (Fig. 1). We use R_L to denote the R-loop structure for a given L . A red line and pink shading represent the average free energy profile and the landscape fluctuations for the different sequences in the datasets, respectively. The blue dots represent the average free energy increase (penalty) for a single mismatch at the given position, and the error bars indicate the variations for the different mismatch types.

occurs unidirectionally from the PAM-adjacent seed region to the PAM-distal region, the folding of the seed region is crucial for initiating and promoting R-loop formation. Consequently, destabilizing mismatches within the seed region can have a detrimental impact on R-loop stability.

Although previous studies have suggested a potential correlation between cleavage efficiency and the equilibrium thermal stability of the seed region [53], further mutation experimental data negate such a correlation for the PAM-distal region, particularly for the mutations in the [+13 to +16] region [53]. The data contradict the equilibrium stability-based results and highlight the significance of nonequilibrium kinetic effects in the distal region, particularly in the [+13 to +16] region. The structure and molecular dynamic simulations of the system show that the ntDNA segment between positions +13 and +16 is highly flexible and Cas9 may undergo conformational changes upon the formation of the

R-loop in this region [9–11,63–66]. The results suggest the possibility of weakened interactions in the region from +13 to +16 and hence a kinetic barrier for the R-loop folding in this region. For these kinetically important nucleotides, we further refine the energy parameters using a genetic algorithm (GA) guided by the correlation between $k_{AS}^{(ii)}$ and the experimental cleavage efficiency data (see details in Methods).

B. Energy landscape analysis of on- and off-target cleavages

Using the refined uCRISPR energy parameters above [44,67], we investigate the free energy landscape along the progression of the R-loop formation from the PAM-proximal ($L = 0$) to the PAM-distal ($L = 20$) end. To understand the distinctive kinetics for on- and off-target cleavages, we perform a large-scale energy landscape computation for the 7148 on-target and 7390 off-target cases (see Fig. 3 for the results). For on-target cleavage, as shown in Fig. 3(a), the R-loop

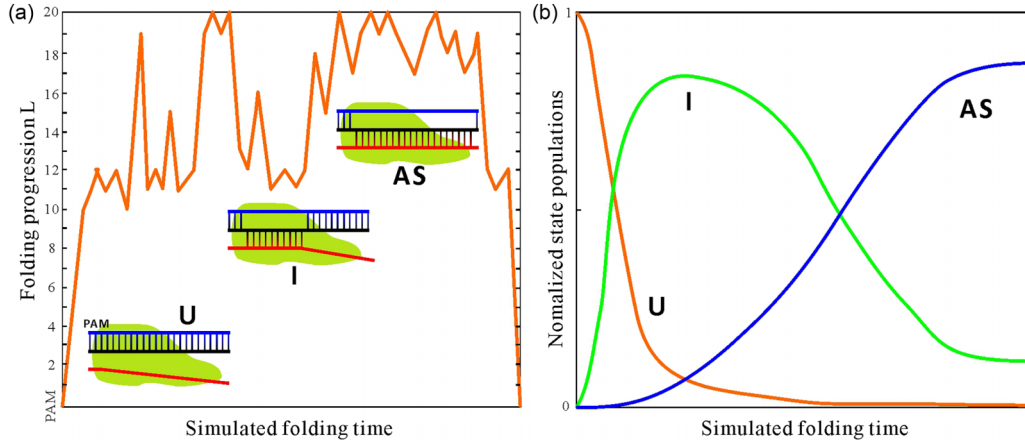


FIG. 4. (a) R-loop folding trajectory of an on-target sequence (EMX1 target 1 [28,40,68]) with schematic illustrations of the structures. The vertical axis shows the DNA-RNA duplex length L of the R-loop (Fig. 1). (b) The population kinetics of the unfolded state U, the intermediate state I, and the active state AS.

folding follows a downhill energy landscape for the formation of the seed region from $L = 0$ to $L = 10$, followed by an energy barrier crossing between +13 and +16 before reaching the AS. For off-target cleavage, the non-Watson-Crick (mismatch) base pairs between the sgRNA and the DNA target are less stable than the corresponding Watson-Crick (canonical) base pairs. As a result, the free energy of an off-target sequence is elevated at the positions where mismatches occur (see the bumps in the energy landscape in Fig. 3(a) and in Fig. S5 of the SM [57]). Such distorted energy landscapes would lead to distinct kinetics due to the formation or change of the energy barriers.

C. A multistate kinetic mechanism of R-loop folding and unfolding

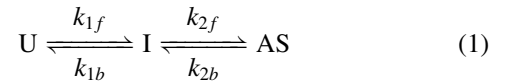
The KMC trajectory for the folding of a typical R-loop [e.g., the EMX1 target 1 gene [28,40,68] in Fig. 4(a)] shows three characteristic states for on-target systems: the unfolded state U ($=R_0$), where the Cas9-sgRNA complex has bound to the PAM site of the DNA, but the hybridization between sgRNA and DNA has not yet occurred; the partially folded state I with $L = 10$ –12 hybridized base pairs formed; and the folded state AS (active state) with $L = 17$ –20 hybridized base pairs formed. Here, the intermediate state I and the cleavage active state AS correspond to the two energy minima on the energy landscape [see Figs. 3(b)–3(d)], and state I is effectively an updated version of the aforementioned I^* state. We note that the R-loop has two kinetic exits: either going forward to the cleaved state through AS or returning to the pre-PAM-binding state through U. Therefore, we explicitly include AS and U in the kinetic model, and adding the intermediate I, we have a three-state kinetic model.

We note that the exact energy landscape for the R-loop can vary for the different on- and off-target scenarios, and off-target cases can be further complicated by additional free energy barriers due to mismatch free energy penalties (see Fig. 3(a) and Fig. S6 of the SM [57]). To account for such case-specific energy landscape features, we use KMC simulations to determine the kinetics for each individual case.

Large-scale energy landscape calculations for the 7148 on-target cases and 7390 off-target cases with single [Fig. 3(c)] or tandem double mismatches [Fig. 3(d)] show similar average energy landscape profiles. Off-target sequences typically exhibit a lower stability of the R-loop and consequently a lower cleavage efficiency. However, as shown in Fig. 3(b), certain on-target R-loops can be less stable (with higher free energy) than off-target R-loops containing mismatches in the PAM-distal region. We note that the majority of genome-wide off-target effects involve single-mismatched off-target sites [28].

D. Two-step reaction mechanism and the cleavage efficiency

As shown by the populational kinetics in Fig. 4(b) for the three states, the R-loop folding and unfolding can be described as a two-step process: the formation or disruption of the seed region (the transition between U and I) and the folding and unfolding of the cleavage active state AS (the transition between I and AS), respectively:



Here k_{1f} , k_{2f} , k_{1b} , and k_{2b} can be obtained from KMC [Eqs. (5) and (6) in the Methods section].

To further validate the two-step folding and unfolding reactions model, we compare the predicted cleavage rate with experimental data. As shown in Fig. 5(a), we construct the complete kinetic network by including the (initial) pre-PAM binding (unbound) state of the Cas9-sgRNA complex and the DNA target and the (final) cleaved state. DNA cleavage occurs at an experimentally determined rate constant k_{cut} [53] once the AS is formed. We perform KMC simulations to predict the R-loop folding and unfolding kinetics for an on-target cleavage case and off-target cases with various single mismatches. These simulations explicitly sample all the microstates of the system instead of the simplified five macrostates in Fig. 5(a). As shown in Figs. 5(b) and 5(c), starting from the initial unbound state, the simulated kinetics for the uncleaved population ($U + I + SA$) aligns with the experimental data for both

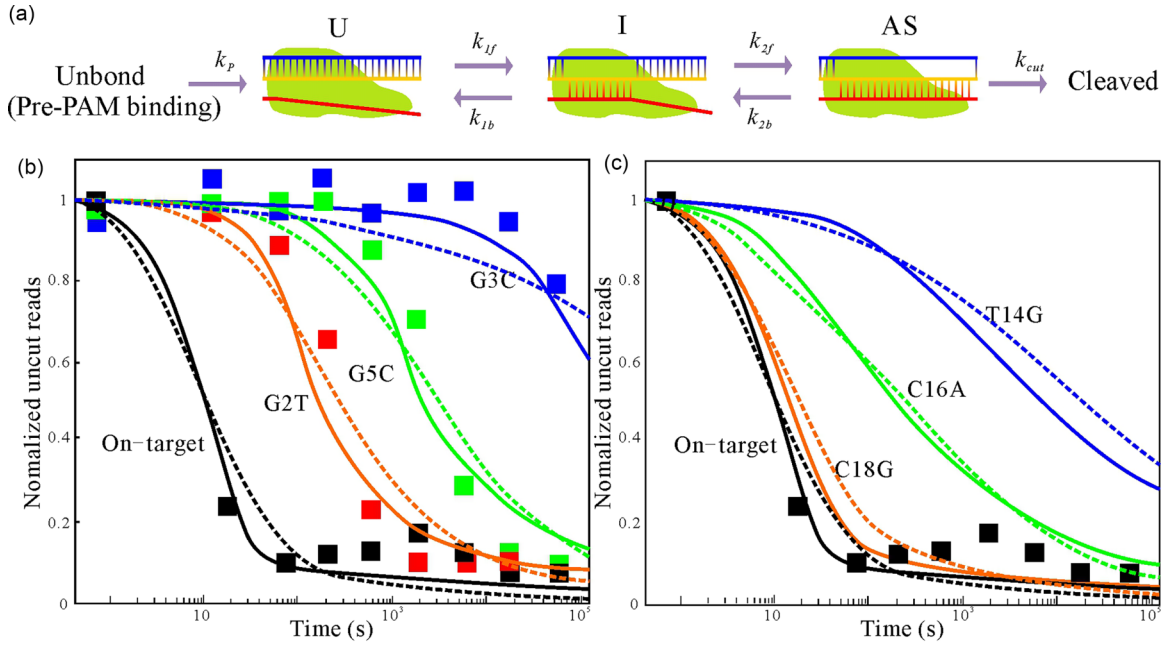


FIG. 5. (a) A simplified kinetic model for Cas9 cleavage efficiency, including the binding between the Cas9-sgRNA complex and the DNA at the PAM site, progression of the R-loop formation involving states U, I, and AS, and the cleavage. [(b) and (c)] Simulated cleavage fraction of an on-target sequence (20-GACGCATAAAGATGAGACGC-1) and a corresponding off-target system with a single mismatch in (b) the PAM-proximal region and (c) the PAM-distal region as a function of reaction time. The symbols (squares) represent the experimentally measured data [53]. Solid and dashed lines are the results from the KMC simulations based on the complete ensemble of microstates and on the simplified two-step reaction network in (a), respectively.

on-target and off-target cases. The result supports the validity of the definition of the AS.

In comparison, we also run simulations for the five-state network [Fig. 5(a)] with the interstate transition rate constants precalculated by KMC. As shown in Figs. 5(b) and 5(c), both the simplified five-state model and the simulation with the complete ensemble of microstates yield similar results for the on-target sequence. These results suggest that our two-step reaction model may have successfully captured the key kinetic features of the on-target cleavage mechanism. The unfolding rate $k_{AS}^{(u)}$ of the AS based on the two-step reaction can therefore serve as a valid approximation for on-target cases. For the off-target cases, while the KMC simulations of the five-state network have shown some consistency with the original simulations, discrepancies between the results of the two kinetic networks become more pronounced for sequences with lower cleavage rates [e.g., G3C and T14G, as shown in Figs. 5(b) and 5(c)]. As a result, the simplified kinetic model may not perform well for some off-target cases due to the increased complexity of the energy landscape due to mismatches at different positions, and we need to run KMC simulations with the full ensemble of microstates.

E. A kinetic predictor for cleavage efficiency

The DNA-Cas9/sgrRNA system undergoes a dual-direction reaction: the AS state can either move forward to the cleaved state through the cleavage reaction at rate k_{cut} (Fig. 5) or backward to the unfolded state U, followed by the possible dissociation of Cas9-sgRNA from the DNA ($k > 1 \text{ s}^{-1}$) and the subsequent degradation of the sgRNA, leaving the DNA

uncleaved [48]. Here the effective rate k_{cut} is a combined result of multiple processes including the conformational changes of the Cas9 protein ($k < 3 \times 10^{-1} \text{ s}^{-1}$) and metal ion binding (k depends on ion concentration), and can be slow [48]. The cleavage efficiency for the dual-direction system is correlated to the residence time of AS, which is determined by the escape rate. The rate of the precleavage escape from the cleavage active state AS is effectively determined by the unfolding rate $k_{AS}^{(u)}$ of AS. Therefore, the physical mechanism of the system supports the unfolding rate $k_{AS}^{(u)}$ as the principal kinetic predictor for cleavage efficiency, a conclusion drawn from the RF supervised ensemble learning based on the experimental cleavage data.

The energy landscape of on-target systems [Fig. 3(b)] shows two minima, AS and I. The system, including the initial state U, leads to two-step kinetics. The slow rates k_{lb} and k_{cut} compared to k_{2f} and k_{2b} [Eq. (1)] suggest a preequilibration between I and AS and hence the formation of an effective macrostate $AS^* = I + AS$. The unfolding rate $k_{AS^*}^{(u)}$ (from AS^* to U) is given by

$$k_{AS^*}^{(u)} \simeq k_{lb} \frac{k_{2b}}{k_{2f} + k_{2b}}. \quad (2)$$

Here $k_{AS}^{(u)} = (1 + K)k_{AS^*}^{(u)} \simeq k_{AS^*}^{(u)}$, where the $I \rightleftharpoons AS$ equilibrium constant K is small for on-target cases. We note that the initial calculation for $k_{AS}^{(u)}$ using Eq. (6) ignores the presence of the intermediate state I. The two-step kinetic process involving the intermediate I above would provide a more accurate estimation for the unfolding rate for on-target systems.

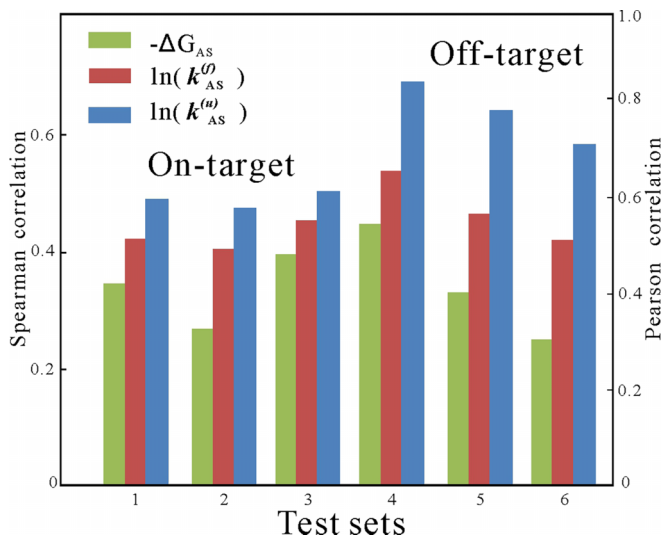


FIG. 6. Overview of the correlation between the experimentally measured cleavage efficiencies and the folding free energy, and folding and unfolding rates. Labels 1–3 denote the on-target datasets (1, Koike-Yusa/Xu Mouse ESC [32]; 2, Hart Rpe [37]; 3, Hart Hct116-lib1 [37]) and labels 4–6 denote off-target datasets (4, Doench-CD33 part [7]; 5, Tsai-VEGFA site 2 [40]; 6, Kescu-sgRNA 6 [33]).

For off-target cases, as shown in Fig. 3, the mismatched base pairs could cause significant energy barriers at positions of the mismatches, resulting in sequence-specific alteration of the energy landscape and kinetics. In particular, a high energy barrier induced by a mismatch in the kinetically important region between +13 and +16 would lower the exchange rate between I and AS and invalidates the preequilibrium approximation used in Eq. (2). Therefore, for off-target systems, there exists no general analytic rate formula such as Eq. (2) and the unfolding rate of the AS can only be determined from (KMC) simulations for the given sequence.

Considering the impact of the PAM sequence [26], we use the following scoring function (predictor) for the cleavage efficiency, $\eta^{(\text{pred})} = \ln(f_{\text{PAM}} k_{\text{AS}}^{(u)})$, where f_{PAM} accounts for the effects of noncanonical PAM sequences (5'-NNN-3') on the cleavage efficiency [26]. The cleavage efficiency is also influenced by experimental and/or biological conditions, and accounting for such effects is challenging. Therefore, the scoring function is more meaningful for comparing different sequences or targets for a given biological system.

To illustrate the interpretability of the importance rankings obtained from the random forest approach, we calculated the correlation between experimentally measured cleavage efficiencies and the top three ranked indicators—AS folding free energy, folding rate, and unfolding rate. Figure 6 presents representative examples from six datasets, including both on- and off-target cases across training and test sets. Among these datasets, $k_{\text{AS}}^{(u)}$ consistently shows the highest correlation with cleavage efficiency, which is consistent with its top ranking in the random forest feature importance analysis. Additional Wilcoxon signed-rank tests for $k_{\text{AS}}^{(u)}$ and $k_{\text{AS}}^{(f)}$ across the training datasets within our random forest approach yielded a p -value of 3×10^{-5} , indicating high reliability of the kinetic predictor for cleavage efficiency.

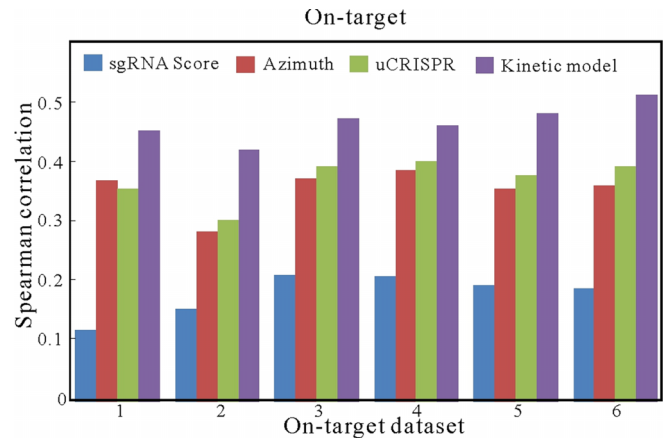


FIG. 7. Performance of on-target activity predictions compared with sgRNA Scorer [24], Azimuth [26], and uCRISPR [44]. Datasets 1–6 are Koike-Yusa/Xu Mouse ESC [32], Hart Rpe [37], Hart Hct116-1 lib1 [37], Hart Hct116-2 lib1 [37], Hart HeLa lib 1 [37], and Hart HeLa lib 2 [37].

To test the accuracy of the kinetic predictor $\eta^{(\text{pred})}$ for cleavage efficiency, we first test the kinetic scoring approach on six on-target datasets, each containing more than 900 sgRNA sequences. Since the ranking of cleavage efficiency among on-target sequences is the most relevant factor to be considered for on-target site selection, we calculate the Spearman rank correlation between the kinetic predictor $\eta^{(\text{pred})}$ and the experimentally measured quantification of cleavage efficiency. The results are compared to two other bioinformatics-based models, sgRNA Scorer [24] and Azimuth [26], as well as the thermodynamics-based model uCRISPR [44]. As shown in Fig. 7, for all six test datasets, the kinetic approach outperforms the other models, achieving an average Spearman correlation of 0.47. In comparison, sgRNA Scorer, Azimuth, and uCRISPR yield lower correlations of 0.18, 0.33, and 0.36, respectively. Therefore, the kinetics-based evaluation of on-target activity significantly improves the predictions, including those from thermal stability-based predictions. The conclusion is corroborated by further extensive testing on large datasets (see results in Table S3 of the SM [57]).

We then evaluate the model on off-target systems by calculating the Pearson correlation coefficient between the kinetic predictor and experimental measurements in order to assess the model's ability to predict the overall off-target effect. The test system contains 51 independent experimental off-target datasets (a total of 17 099 target-gRNA sequence pairs), which include both single and multiple mismatches across a wide range of genome-wide off-target sites. As shown in Fig. 8, the kinetic model shows notable improvements in the overall average Pearson correlation, 0.82 ± 0.17 , compared to 0.38 ± 0.28 , 0.43 ± 0.28 , 0.47 ± 0.23 , 0.68 ± 0.24 , and 0.75 ± 0.24 for the four popular bioinformatics-based models MIT_Zhang [28], CCTop [29], CROP-IT [34], and CFD [26] and the thermodynamics-based uCRISPR model [44], respectively. Further detailed results are described in Table S4 of the SM [57]. Compared to uCRISPR [44], the number of failed cases in the current model (Pearson correlation < 0.4) is

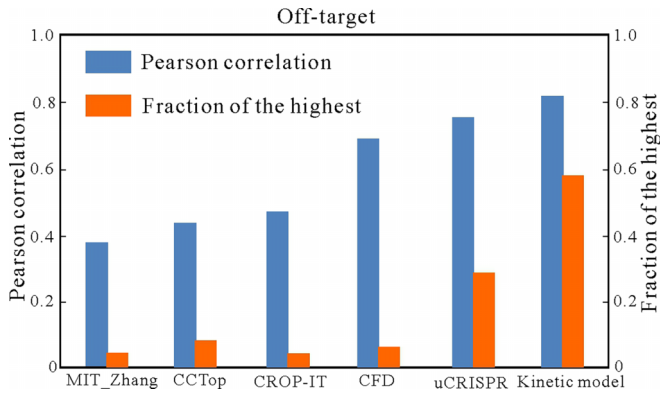


FIG. 8. Overall comparisons between different models with the average correlation (blue bars) and the fraction of the best predictions with the highest correlations (orange bars) for 51 off-target datasets.

significantly reduced. Such a substantial improvement for the large-scale test datasets, which contain multiple mismatches, further supports our kinetic mechanism for off-target cleavage efficiency. Friedman tests were performed to compare the algorithms for the on-target and off-target test sets, respectively. Using the chi-squared cumulative distribution function, we calculated the p -values of the nonparametric statistical tests, which yielded $p = 6.4 \times 10^{-11}$ for on-target test sets and $p = 5.4 \times 10^{-15}$ for off-target test sets. These results support the reliability of our kinetic model.

III. CONCLUSIONS

Predicting cleavage efficiency for on-target sequences and related off-target events is a critical step in sgRNA design. Kinetics-based predictions offer a powerful new strategy by enhancing on-target and mitigating off-target cleavage. A comprehensive kinetic profile of R-loop folding and unfolding for both on- and off-target cases can also aid in the mechanistic interpretation of gene-editing data and in optimizing experimental design to improve overall gene-editing outcomes.

Leveraging large-scale data for cleavage efficiencies and combining supervised learning and physics-based kinetic modeling, we identify the critical feature that determines the CRISPR-Cas9 efficiency for on- and off-target cleavages. Furthermore, we develop a kinetic model that can demonstrate the step-by-step transitions of the R-loop folding and unfolding progression, allowing us to trace the kinetic trajectories of the system. By considering the nontarget DNA strand interacting with the Cas9 RuvC and HNH domains at key positions, the kinetic model, compared with other approaches, acquires notably improved predictions for CRISPR-Cas9 cleavage efficiencies for both on- and off-target systems. Furthermore, our analyses of the great variety of sequences in the databases of on- and off-target cleavage sites reveal the following kinetic mechanisms in Cas9-sgRNA targeting and cleaving:

(1) The folding and unfolding of the R-loop act as a two-step reaction process with a hybridized seed region formed as an intermediate.

(2) Cleavage can be initiated after the AS with 17 continuous base pairs formed from the PAM end is formed.

(3) The unfolding rate of the AS shows the most significant correlation with the experimentally measured cleavage efficiencies.

In contrast to the traditional thermodynamic analysis of the CRISPR-Cas9 system, which assumes a two-state transition without considering the formation of the possible intermediates, the kinetic process follows a two-step (three-state) reaction process. The first step in R-loop folding (seed-region folding) essentially determines the binding affinity of Cas9-sgRNA to dsDNA, and the second step initiates cleavage as induced by RuvC and HNH domains. This mechanism may explain why binding affinity and cleavage efficiency weakly correlate *in vitro* and *in vivo*, especially for off-target sequences. Moreover, the strong tolerance of mismatches in the distal region suggests that the AS may contain multiple hybridized structures. The strong correlation between states with at least 17 continuous base pairs formed in the R-loop and experimentally measured cleavage efficiencies further supports this mechanism.

Recent atomic-level structural studies [69–72] investigating the interactions between target DNA and the Cas9 HNH domain have revealed that mismatches at nucleotide positions 12–17 exhibit distinct interactions, and in modulating Cas9 cleavage activation [69,70]. Our study reaches consistent conclusions, identifying positions 13–16 as the kinetically significant positions. Unlike previous studies that rely on single-sequence structural models, our conclusion is based on a large dataset encompassing more than 14 000 on- and off-target cases, enabling us to draw more general conclusions about the kinetic mechanisms. More importantly, our simulations capture R-loop folding kinetics on a case-by-case basis, explicitly incorporating the sequence context for each target. As a result, the model developed here integrates local mismatch effects with the on-target sequence, allowing for more robust and broadly applicable predictions. However, one limitation of our approach is the lack of detailed atomic-scale structural information, especially for some critical competitions between nucleotide-nucleotide and nucleotide-protein interactions, which reshape the R-loop folding kinetics by integrating nucleotide-protein interactions into the free energy landscape of DNA-RNA hybrid strand displacement [73,74]. To bridge this gap, future extensions of the kinetic model will aim to incorporate atomic-scale structural kinetics.

IV. METHODS

A. R-loop folding kinetics

A functional CRISPR-Cas9 nuclease complex requires the R-loop to be properly folded and positioned in the reaction channel between the Cas9 RuvC and HNH domains before the target DNA is cleaved (see Fig. 1(d) and Fig. S2(a) of the SM [57]) [64,65]. The R-loop is formed through the hybridization of the single-stranded guide RNA (sgRNA) with the target strand (TS) upon the unwinding of the double-stranded DNA (dsDNA) [see Figs. 1(a)–1(c)], which is an essential kinetic step in nuclease activation [48,51]. Experimental results suggest that the R-loop folding kinetics is closely related to

cleavage efficiency [48,53,54]. We use a sequential zipping-unzipping-like kinetic model to describe the kinetic process of the RNA-DNA hybrid duplex and the subsequent folding and unfolding of the R-loop (see Figs. 1(a)–1(c) and Fig. S1(a) of the SM [57]).

Folding states and kinetic moves. The base-pair exchange between the dsDNA and the sgRNA-DNA hybrid progresses from the PAM-proximal region to the PAM-distal region [see Fig. 1(a)]. The zipping pathway can be described by 20 states, defined by the number of RNA-DNA hybrid base pairs formed. Consistent with previous studies, the (PAM-adjacent) first 10 base pairs constitute the seed region, while the remaining 10 base pairs make up the PAM-distal region. A basic kinetic move of R-loop folding (or unfolding) is defined by a single base pair switching from (or to) dsDNA to (or from) the sgRNA-DNA hybrid [75–78]. Specifically, an R-loop state with N base pairs (denoted as state R_N) has two kinetically adjacent states $R_{N\pm 1}$. The time evolution of the population for each state in a given system is described by the master equation

$$\frac{dp_i(t)}{dt} = \sum_j [k_{ji}p_j(t) - k_{ij}p_i(t)], \quad (3)$$

where k_{ij} is the rate constant for a kinetic move from state i to state j and $p_i(t)$ represents the fractional populations of state i .

Rate constants. We use the conventional Metropolis rule to calculate the rate constant for a kinetic move:

$$k_{i \rightarrow j} = \min(k_0, k_0 e^{-\Delta G_{ji}/k_B T}), \quad (4)$$

where $\Delta G_{ji} = G_j - G_i$ is the free energy difference between the states before and after the exchange of base pairs between dsDNA and the sgRNA-DNA hybrid, k_0 is the attempt frequency (set according to previous studies [79]), k_B is the Boltzmann constant, and T is the temperature. The free energies of the dsDNA and the sgRNA-DNA hybrid system are calculated from the nearest-neighbor model [67] and the uCRISPR model [44], respectively.

Kinetic Monte Carlo simulation. To understand the microscopic trajectories of R-loop folding and the folding and unfolding pathways of the R-loop, we perform the Kinetic Monte Carlo (KMC) simulations [79] based on the full ensemble of the 20 states as the microstates. The kinetic moves in KMC are simulated in three steps:

(1) Construct the total outflow transition rate as $k_{\text{total}} = \sum_n k_{pn}$, where k_{pn} is the rate constant for the transition from the current state p to one other adjacent state n .

(2) Select the next state q such that $\sum_{n=1}^q k_{pn}/k_{\text{total}} \geq \rho_1 \geq \sum_{n=1}^{q-1} k_{pn}/k_{\text{total}}$ for a randomly generated number ρ_1 between 0 and 1.

(3) Update the system time from t to $t - \ln(\rho_2)/k_{\text{total}}$, where ρ_2 is another random number between 0 and 1.

By repeating the above kinetic moves, starting from an initial state of the R-loop, KMC generates a folding trajectory. Averaging over an ensemble of simulated trajectories yields the R-loop folding-unfolding pathway and the free energy landscape along the pathway.

Populational kinetics and overall transition rate. We perform multiple (e.g., 5000) KMC simulations to generate an

ensemble of folding trajectories. Statistically, the average over a sufficiently large number of trajectories from the same initial state provides the population kinetics as well as the folding pathway (see Fig. S1(b) [57]). More importantly, we estimate the overall transition rate from state A to state B as

$$k_{AB} = \frac{1}{\langle \text{FPT}_{AB} \rangle}, \quad (5)$$

where FPT_{AB} denotes the first-passage time from state A to state B in a KMC trajectory. For the reverse process from B to A, if the kinetic barrier is high, using $k_{BA} = 1/\langle \text{FPT}_{BA} \rangle$ to compute the rate is not feasible due to poor sampling. In that case, we estimate the corresponding “apparent” rate from the following equation:

$$k_{BA}^{\text{app}} = k_{AB} e^{\Delta G_{BA}/k_B T}, \quad (6)$$

where $\Delta G_{BA} = G_B - G_A$ is the free energy difference between the two states. The free energy for a state is computed as an ensemble of microstates S accessible to the state: $G = -k_B T \ln(\sum_S e^{-G_S/k_B T})$, where G_S is the free energy of microstate S . These transition rates for an R-loop can be denoted as $k_x^{(u)}$ and $k_x^{(f)}$, where x is an R-loop state. The rate $k_x^{(f)}$ for the folding process can be computed using Eq. (5) from KMC simulations, and the rate $k_x^{(u)}$ for unfolding, an uphill process on the energy landscape, is calculated using Eq. (6).

B. Position-dependent parameters

In the presence of the Cas9 protein, the nucleotide-based energy parameters and hence the rate constants are position dependent. Starting from the uCRISPR parameters [44], we refine the energy parameters for the identified critical (most sensitive) positions.

Energy parametrization. The free energy of state N , with n base pairs zipped from the PAM side in the RNA-DNA hybrid duplex, is calculated as the sum of the base-pair stacking free energies of the RNA-DNA hybrid:

$$\Delta G_N = \sum_{i=1}^n E_i \left(\begin{smallmatrix} 5'XY3' \\ 3'X'Y'5' \end{smallmatrix} \right). \quad (7)$$

Here, $E_i(\begin{smallmatrix} 5'XY3' \\ 3'X'Y'5' \end{smallmatrix})$ is the energy of the stack of base pair $(\begin{smallmatrix} X \\ X' \end{smallmatrix})$ at position i and $(\begin{smallmatrix} Y \\ Y' \end{smallmatrix})$ at position $i - 1$.

There are $4^4 = 256$ energy parameters for all the possible $(\begin{smallmatrix} XY \\ X'Y' \end{smallmatrix})$ stacks at each position of the sequence (positions 1–20), including $4^2 = 16$ energy parameters for on-target sequences (canonical base-pair stacking) and 240 energy parameters for off-target sequences (noncanonical base-pair stacking). Considering the 20 positions in the R-loop, we have $20 \times 16 = 320$ position-dependent energy parameters for canonical base stacks and $20 \times 240 = 4800$ parameters for noncanonical (mismatched) base stacks. We use the uCRISPR approach [44], which is based on the equilibrium thermal stability of the folded CRISPR-Cas9 system, to assign initial values to the energy parameters in Eq. (7).

On-target parameters. The energy parameters for the on-target sequences are obtained through singular value decomposition (SVD) based on five training datasets: Wang/Xu

HL60/KBM7 [32], Doench MOLM13/NB4/TF1 [7], Doench A375/AZD [26], Doench Mouse EL4 [7], and Koike-Yusa/Xu Mouse ESC [32] (see Table S1 [57] for details). The quantification of gRNA on-target efficiency is consistent within the corresponding ($\log_2$ -transformed) experimental datasets used for training. This process yields the fitted on-target energy parameters for $E_i(XY)$ for DNA and $E_i(\overline{X'}\overline{Y'})$ for RNA.

Off-target parameters. We decompose the off-target energy parameters at position i :

$$E_i\left(\frac{5'XY3'}{3'X'Y'5'}\right) = w_D E_i(XY) + w_R E_i(\overline{X'}\overline{Y'}) + E_i\left(\frac{Y}{\overline{Y}}\right). \quad (8)$$

Here $E_i(XY)$ and $E_i(\overline{X'}\overline{Y'})$ are the energy parameters for the dinucleotides $5'-XY-3'$ and $5'-\overline{X'}\overline{Y'}-3'$, respectively. Here, w_D and w_R are the position-independent weight coefficients, and $E_i\left(\frac{Y}{\overline{Y}}\right)$ is the position-dependent noncanonical base-pairing energy. $5'-\overline{X'}\overline{Y'}-3'$ and $\overline{Y}$ denote the dinucleotide complementary to $3'-\overline{X'}\overline{Y'}-5'$ and the nucleotide complementary to Y , respectively. Accounting for all the possible non-canonical base-pairing types, the energy term $E_i\left(\frac{Y}{\overline{Y}}\right)$ has $4 \times 3 = 12$ possibilities at position i . Therefore, the above decomposition reduces the total number of parameters from 4800 (240 possible mismatched base-pair stacks $\times$ 20 positions) to 242, including the two position-independent weight coefficients.

Similar to the on-target parameters, the 242 off-target parameters can be determined using SVD with 18 off-target training sets (see Table S1 [57] for details).

C. Refinement of energy parameters for the kinetically critical positions

Kinetically critical positions. The RF analysis revealed critical (the most sensitive) base-pair positions (+13 to +16) for determining the unfolding rate ($k_{AS}^{(u)}$) of the AS. We intro-

duce a novel term ($\Delta G_{i,\text{DNA-Cas9}}$) to explicitly account for the Cas9-DNA interaction:

$$\Delta G_i\left(\frac{5'XY3'}{3'X'Y'5'}\right) = -\Delta G_{i,\text{DNA}}(Y) + \omega_i \Delta G_{i,\text{DNA-Cas9}}(Y) + \Delta G_{i,\text{RNA-DNA}}\left(\frac{5'XY3'}{3'X'Y'5'}\right), \quad (9)$$

where ω_i is the position-dependent weight factor for the critical positions $i = 13, 14, 15, 16$, $\Delta G_{i,\text{DNA}}(Y)$ is the free energy of DNA base-pair stacking at position i , and $\Delta G_{i,\text{RNA-DNA}}\left(\frac{5'XY3'}{3'X'Y'5'}\right)$ is the free energy of RNA-DNA hybrid base stacking at position i .

Parameter optimization. We use a genetic algorithm (GA), a highly efficient, simplexlike, self-guided search approach, to optimize the ω_i 's. The objective is to maximize the correlation between $k_{AS}^{(u)}$ and experimentally measured cleavage efficiency in the training datasets. The refinement is guided by the correlation between $k_{AS}^{(u)}$, calculated using Eq. (2), and the experimental cleavage efficiency for the on-target parameters (see Table S5 [57] for the refined parameters, Fig. 6 for the Spearman and Pearson correlation coefficients, and Fig. S4 [57] for the importance score). To further validate the optimized parameters, we perform 10-fold cross validation. The convergence of ω_i and the correlations (see Fig. S7 [57]) suggest the robustness of the optimized parameters.

ACKNOWLEDGMENTS

This research was supported by NSF Grant No. CHE-2154924 (to S.-J.C.); NIH Grants MIRA R35-GM134919 (to S.-J.C.) and No. R01-NS131416 (to D.D.). We thank Dr. Dong Zhang and Dr. Travis Hurst for valuable discussions.

DATA AVAILABILITY

The data that support the findings of this article are openly available [80].

-
- [1] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church, RNA-guided human genome engineering via Cas9, *Science* **339**, 823 (2013).
 - [2] B. Chen, L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G.-W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, *et al.*, Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system, *Cell* **155**, 1479 (2013).
 - [3] T. Wang, J. J. Wei, D. M. Sabatini, and E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system, *Science* **343**, 80 (2014).
 - [4] P. D. Hsu, E. S. Lander, and F. Zhang, Development and applications of CRISPR-Cas9 for genome engineering, *Cell* **157**, 1262 (2014).
 - [5] A. V. Wright, J. K. Nuñez, and J. A. Doudna, Biology and applications of CRISPR systems: Harnessing nature's toolbox for genome engineering, *Cell* **164**, 29 (2016).
 - [6] J. D. Sander and J. K. Joung, CRISPR-Cas systems for editing, regulating and targeting genomes, *Nat. Biotechnol.* **32**, 347 (2014).
 - [7] J. G. Doench, E. Hartenian, D. B. Graham, Z. Tothova, M. Hegde, I. Smith, M. Sullender, B. L. Ebert, R. J. Xavier, and D. E. Root, Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation, *Nat. Biotechnol.* **32**, 1262 (2014).
 - [8] J. Sun, J. Wang, D. Zheng, and X. Hu, Advances in therapeutic application of CRISPR-Cas9, *Briefings Funct. Genomics* **19**, 164 (2020).
 - [9] F. Jiang and J. A. Doudna, The structural biology of CRISPR-Cas systems, *Curr. Opin. Struct. Biol.* **30**, 100 (2015).
 - [10] F. Jiang, D. W. Taylor, J. S. Chen, J. E. Kornfeld, K. Zhou, A. J. Thompson, E. Nogales, and J. A. Doudna, Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage, *Science* **351**, 867 (2016).
 - [11] F. Jiang and J. A. Doudna, CRISPR-Cas9 structures and mechanisms, *Annu. Rev. Biophys.* **46**, 505 (2017).
 - [12] D. Singh, S. H. Sternberg, J. Fei, J. A. Doudna, and T. Ha, Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9, *Nat. Commun.* **7**, 12778 (2016).

- [13] F. Jiang, K. Zhou, L. Ma, S. Gressel, and J. A. Doudna, A Cas9–guide RNA complex preorganized for target DNA recognition, *Science* **348**, 1477 (2015).
- [14] Y. Lin, T. J. Cradick, M. T. Brown, H. Deshmukh, P. Ranjan, N. Sarode, B. M. Wile, P. M. Vertino, F. J. Stewart, and G. Bao, CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences, *Nucleic Acids Res.* **42**, 7473 (2014).
- [15] F. A. Ran, L. Cong, W. X. Yan, D. A. Scott, J. S. Gootenberg, A. J. Kriz, B. Zetsche, O. Shalem, X. Wu, K. S. Makarova, *et al.*, *In vivo* genome editing using *Staphylococcus aureus* Cas9, *Nature (London)* **520**, 186 (2015).
- [16] Y. Dang, G. Jia, J. Choi, H. Ma, E. Anaya, C. Ye, P. Shankar, and H. Wu, Optimizing sgRNA structure to improve CRISPR–Cas9 knockout efficiency, *Genome Biol.* **16**, 280 (2015).
- [17] Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio, and J. K. Joung, Improving CRISPR–Cas nuclease specificity using truncated guide RNAs, *Nat. Biotechnol.* **32**, 279 (2014).
- [18] A. Hendel, R. O. Bak, J. T. Clark, A. B. Kennedy, D. E. Ryan, S. Roy, I. Steinfeld, B. D. Lunstad, R. J. Kaiser, A. B. Wilkens, *et al.*, Chemically modified guide RNAs enhance CRISPR–Cas genome editing in human primary cells, *Nat. Biotechnol.* **33**, 985 (2015).
- [19] H. Yin, C.-Q. Song, S. Suresh, Q. Wu, S. Walsh, L. H. Rhym, E. Mintzer, M. F. Bolukbasi, L. J. Zhu, K. Kauffman, *et al.*, Structure-guided chemical modification of guide RNA enables potent non-viral *in vivo* genome editing, *Nat. Biotechnol.* **35**, 1179 (2017).
- [20] P. Mali, J. Aach, P. B. Stranges, K. M. Esvelt, M. Moosburner, S. Kosuri, L. Yang, and G. M. Church, Cas9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering, *Nat. Biotechnol.* **31**, 833 (2013).
- [21] M. Müller, C. M. Lee, G. Gasiunas, T. H. Davis, T. J. Cradick, V. Siksnys, G. Bao, T. Cathomen, and C. Mussolino, *Streptococcus thermophilus* CRISPR–Cas9 systems enable specific editing of the human genome, *Mol. Ther.* **24**, 636 (2016).
- [22] B. Zetsche, J. S. Gootenberg, O. O. Abudayyeh, I. M. Slaymaker, K. S. Makarova, P. Essletzbichler, S. E. Volz, J. Joung, J. Van Der Oost, A. Regev, *et al.*, Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR–Cas system, *Cell* **163**, 759 (2015).
- [23] S.-J. Chen, Minimizing off-target effects in CRISPR–Cas9 genome editing, *Cell Biol. Toxicol.* **35**, 399 (2019).
- [24] R. Chari, P. Mali, M. Moosburner, and G. M. Church, Unraveling CRISPR–Cas9 genome engineering parameters via a library-on-library approach, *Nat. Methods* **12**, 823 (2015).
- [25] M. A. Moreno-Mateos, C. E. Vejnar, J.-D. Beaudoin, J. P. Fernandez, E. K. Mis, M. K. Khokha, and A. J. Giraldez, CRISPRscan: Designing highly efficient sgRNAs for CRISPR–Cas9 targeting *in vivo*, *Nat. Methods* **12**, 982 (2015).
- [26] J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, *et al.*, Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9, *Nat. Biotechnol.* **34**, 184 (2016).
- [27] Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung, and J. D. Sander, High-frequency off-target mutagenesis induced by CRISPR–Cas nucleases in human cells, *Nat. Biotechnol.* **31**, 822 (2013).
- [28] P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases, *Nat. Biotechnol.* **31**, 827 (2013).
- [29] M. Stemmer, T. Thumberger, M. del Sol Keyer, J. Wittbrodt, and J. L. Mateo, CCTOP: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool, *PLoS One* **10**, e0124633 (2015).
- [30] F. A. Ran, P. D. Hsu, C.-Y. Lin, J. S. Gootenberg, S. Konermann, A. E. Trevino, D. A. Scott, A. Inoue, S. Matoba, Yi Zhang, *et al.*, Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity, *Cell* **154**, 1380 (2013).
- [31] J. P. Guilinger, D. B. Thompson, and D. R. Liu, Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification, *Nat. Biotechnol.* **32**, 577 (2014).
- [32] H. Xu, T. Xiao, C.-H. Chen, W. Li, C. A. Meyer, Q. Wu, Di Wu, Le Cong, F. Zhang, J. S. Liu, *et al.*, Sequence determinants of improved CRISPR sgRNA design, *Genome Res.* **25**, 1147 (2015).
- [33] N. Wong, W. Liu, and X. Wang, Wu-CRISPR: Characteristics of functional guide RNAs for the CRISPR/Cas9 system, *Genome Biol.* **16**, 218 (2015).
- [34] R. Singh, C. Kucsu, A. Quinlan, Y. Qi, and M. Adli, Cas9-chromatin binding information enables more accurate CRISPR off-target prediction, *Nucleic Acids Res.* **43**, e118 (2015).
- [35] M. Haeussler, K. Schönig, H. Eckert, A. Eschstruth, J. Mianné, J.-B. Renaud, S. Schneider-Maunoury, A. Shkumatava, L. Teboul, J. Kent, *et al.*, Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR, *Genome Biol.* **17**, 148 (2016).
- [36] J. Tycko, V. E. Myer, and P. D. Hsu, Methods for optimizing CRISPR–Cas9 genome editing specificity, *Mol. Cell* **63**, 355 (2016).
- [37] T. Hart, M. Chandrashekar, M. Aregger, Z. Steinhart, K. R. Brown, G. MacLeod, M. Mis, M. Zimmermann, A. Fradet-Turcotte, S. Sun, *et al.*, High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities, *Cell* **163**, 1515 (2015).
- [38] G. K. Varshney, W. Pei, M. C. LaFave, J. Idol, L. Xu, V. Gallardo, B. Carrington, K. Bishop, M. Jones, M. Li, *et al.*, High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9, *Genome Res.* **25**, 1030 (2015).
- [39] J. A. Gagnon, E. Valen, S. B. Thyme, P. Huang, L. Ahkmetova, A. Pauli, T. G. Montague, S. Zimmerman, C. Richter, and A. F. Schier, Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs, *PLoS ONE* **9**, e98186 (2014).
- [40] S. Q. Tsai, Z. Zheng, N. T. Nguyen, M. Liebers, V. V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A. J. Iafrate, L. P. Le, *et al.*, GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases, *Nat. Biotechnol.* **33**, 187 (2015).
- [41] M. A. Horlbeck, L. A. Gilbert, J. E. Villalta, B. Adamson, R. A. Pak, Y. Chen, A. P. Fields, C. Y. Park, J. E. Corn, M. Kampmann, *et al.*, Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation, *elife*, **5**, e19760 (2016).
- [42] M. Labuhn, F. F. Adams, M. Ng, S. Knoess, A. Schambach, E. M. Charpentier, A. Schwarzer, J. L. Mateo, J.-H. Klusmann, and D. Heckl, Refined sgRNA efficacy prediction improves

- large- and small-scale CRISPR–Cas9 applications, *Nucleic Acids Res.* **46**, 1375 (2018).
- [43] S. Q. Tsai, N. T. Nguyen, J. Malagon-Lopez, V. V. Topkar, M. J. Aryee, and J. K. Joung, Circle-seq: A highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets, *Nat. Methods* **14**, 607 (2017).
- [44] D. Zhang, T. Hurst, D. Duan, and S.-J. Chen, Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design, *Proc. Natl. Acad. Sci. USA* **116**, 8693 (2019).
- [45] J. Listgarten, M. Weinstein, B. P. Kleinstiver, A. A. Sousa, J. K. Joung, J. Crawford, K. Gao, L. Hoang, M. Elibol, J. G. Doench, *et al.*, Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs, *Nat. Biomed. Eng.* **2**, 38 (2018).
- [46] H. Nishimasu, F. A. Ran, P. D. Hsu, S. Konermann, S. I. Shehata, N. Dohmae, R. Ishitani, F. Zhang, and O. Nureki, Crystal structure of Cas9 in complex with guide RNA and target DNA, *Cell* **156**, 935 (2014).
- [47] C. Anders, O. Niewoehner, A. Duerst, and M. Jinek, Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease, *Nature (London)* **513**, 569 (2014).
- [48] A. T. Raper, A. A. Stephenson, and Z. Suo, Functional insights revealed by the kinetic mechanism of CRISPR/Cas9, *J. Am. Chem. Soc.* **140**, 2971 (2018).
- [49] D. Jones, C. Unoson, P. Leroy, V. Curic, and J. Elf, Kinetics of dCas9 target search in *Escherichia coli*, *Biophys. J.* **112**, 314A (2017).
- [50] A. Lapinaite, G. J. Knott, C. M. Palumbo, E. Lin-Shiao, M. F. Richter, K. T. Zhao, P. A. Beal, D. R. Liu, and J. A. Doudna, Dna capture by a CRISPR–Cas9–guided adenine base editor, *Science* **369**, 566 (2020).
- [51] S. H. Sternberg, B. LaFrance, M. Kaplan, and J. A. Doudna, Conformational control of DNA target cleavage by CRISPR–Cas9, *Nature (London)* **527**, 110 (2015).
- [52] X. Xu, D. Duan, and S.-J. Chen, CRISPR–Cas9 cleavage efficiency correlates strongly with target–sgRNA folding stability: From physical mechanism to off-target assessment, *Sci. Rep.* **7**, 143 (2017).
- [53] S. K. Jones, J. A. Hawkins, N. V. Johnson, C. Jung, K. Hu, J. R. Rybarski, J. S. Chen, J. A. Doudna, W. H. Press, and I. J. Finkelstein, Massively parallel kinetic profiling of natural and engineered CRISPR nucleases, *Biophys. J.* **120**, 138A (2021).
- [54] M.-S. Liu, S. Gong, H.-H. Yu, K. Jung, K. A. Johnson, and D. W. Taylor, Engineered CRISPR/Cas9 enzymes improve discrimination by slowing DNA cleavage to allow release of off-target DNA, *Nat. Commun.* **11**, 3576 (2020).
- [55] M. Klein, B. Eslami-Mossallam, D. G. Arroyo, and M. Depken, Hybridization kinetics explains CRISPR–Cas off-targeting rules, *Cell Rep.* **22**, 1413 (2018).
- [56] B. Eslami-Mossallam, M. Klein, C. V. D. Smagt, K. V. D. Sanden, S. K. Jones Jr., J. A. Hawkins, I. J. Finkelstein, and M. Depken, A kinetic model predicts SpCas9 activity, improves off-target classification, and reveals the physical basis of targeting fidelity, *Nat. Commun.* **13**, 1367 (2022).
- [57] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/rtxh-6fdl> for supporting information, which includes Refs. [81–94].
- [58] S. Cao and S.-J. Chen, Structure and stability of RNA/RNA kissing complex: With application to HIV dimerization initiation signal, *RNA* **17**, 2130 (2011).
- [59] S. Cao and S.-J. Chen, Free energy landscapes of RNA/RNA complexes: With applications to snRNA complexes in spliceosomes, *J. Mol. Biol.* **357**, 292 (2006).
- [60] X. Xu, P. Zhao, and S.-J. Chen, Vfold: A web server for RNA structure and folding thermodynamics prediction, *PLoS ONE* **9**, e107504 (2014).
- [61] L. Breiman, Bagging predictors, *Mach. Learn.* **24**, 123 (1996).
- [62] T. K. Ho, A data complexity analysis of comparative advantages of decision forest constructors, *Pattern Anal. Appl.* **5**, 102 (2002).
- [63] G. Palermo, Y. Miao, R. C. Walker, M. Jinek, and J. A. McCammon, Striking plasticity of CRISPR–Cas9 and key role of non-target DNA, as revealed by molecular simulations, *ACS Cent. Sci.* **2**, 756 (2016).
- [64] G. Palermo, Structure and dynamics of the CRISPR–Cas9 catalytic complex, *J. Chem. Inf. Model.* **59**, 2394 (2019).
- [65] G. Palermo, Y. Miao, R. C. Walker, M. Jinek, and J. A. McCammon, CRISPR–Cas9 conformational activation as elucidated from enhanced molecular simulations, *Proc. Natl. Acad. Sci. USA* **114**, 7260 (2017).
- [66] G. Palermo, C. G. Ricci, I. Rivalta, V. S. Batista, and J. A. McCammon, An allosteric signaling governs the CRISPR–Cas9 function, *Biophys. J.* **116**, 485A (2019).
- [67] D. H. Turner and D. H. Mathews, NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure, *Nucleic Acids Res.* **38**, D280 (2010).
- [68] B. P. Kleinstiver, V. Pattanayak, M. S. Prew, S. Q. Tsai, N. T. Nguyen, Z. Zheng, and J. K. Joung, High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects, *Nature (London)* **529**, 490 (2016).
- [69] C. G. Ricci, J. S. Chen, Y. Miao, M. Jinek, J. A. Doudna, J. A. McCammon, and G. Palermo, Deciphering off-target effects in CRISPR–Cas9 through accelerated molecular dynamics, *ACS Cent. Sci.* **5**, 651 (2019).
- [70] B. P. Mitchell, R. V. Hsu, M. A. Medrano, N. T. Zewde, Y. B. Narkhede, and G. Palermo, Spontaneous embedding of DNA mismatches within the RNA: DNA hybrid of CRISPR–Cas9, *Front. Mol. Biosci.* **7**, 39 (2020).
- [71] L. Nierzwicki, K. W. East, U. N. Morzan, P. R. Arantes, V. S. Batista, G. P. Lisi, and G. Palermo, Enhanced specificity mutations perturb allosteric signaling in CRISPR–Cas9, *Elife*, **10**, e73601 (2021).
- [72] E. Skeens, S. Sinha, M. Ahsan, A. M. D’Ordine, G. Jogl, G. Palermo, and G. P. Lisi, High-fidelity, hyper-accurate, and evolved mutants rewire atomic-level communication in CRISPR–Cas9, *Sci. Adv.* **10**, ead11045 (2024).
- [73] H. Liu, F. Hong, F. Smith, J. Goertz, T. Ouldrige, M. M. Stevens, H. Yan, and P. Šulc, Kinetics of RNA and RNA:DNA hybrid strand displacement, *ACS Synth. Biol.* **10**, 3066 (2021).
- [74] E. J. Ratajczyk, J. Bath, P. Šulc, J. P. K. Doye, A. A. Louis, and A. J. Turberfield, Controlling DNA–RNA strand displacement kinetics with base distribution, *Proc. Natl. Acad. Sci. USA* **122**, e2416988122 (2025).
- [75] W. Zhang and S.-J. Chen, Exploring the complex folding kinetics of RNA hairpins: II. effect of sequence, length, and misfolded states, *Biophys. J.* **90**, 778 (2006).
- [76] S. Cao and S.-J. Chen, Biphasic folding kinetics of RNA pseudoknots and telomerase RNA activity, *J. Mol. Biol.* **367**, 909 (2007).

- [77] P. Zhao, W.-B. Zhang, and S.-J. Chen, Predicting secondary structural folding kinetics for nucleic acids, *Biophys. J.* **98**, 1617 (2010).
- [78] X. Xu, T. Yu, and S.-J. Chen, Understanding the kinetic mechanism of RNA single base pair formation, *Proc. Natl. Acad. Sci. USA* **113**, 116 (2016).
- [79] X. Xu and S.-J. Chen, Kinetic mechanism for the conformational switch between bistable RNA hairpins, *Biophys. J.* **102**, 646a (2012).
- [80] <https://rna.physics.missouri.edu/kCRISPR/>.
- [81] C. Huai, G. Li, R. Yao, Y. Zhang, Mi Cao, L. Kong, C. Jia, H. Yuan, H. Chen, D. Lu, *et al.*, Structural insights into DNA cleavage activation of CRISPR-Cas9 system, *Nat. Commun.* **8**, 1375 (2017).
- [82] B. Farboud and B. J. Meyer, Dramatic enhancement of genome editing by CRISPR/Cas9 through improved guide RNA design, *Genetics* **199**, 959 (2015).
- [83] X. Ren, Z. Yang, J. Xu, J. Sun, D. Mao, Y. Hu, S.-J. Yang, H.-H. Qiao, X. Wang, Q. Hu, *et al.*, Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in *Drosophila*, *Cell Rep.* **9**, 1151 (2014).
- [84] X. Liu, A. Homma, J. Sayadi, S. Yang, J. Ohashi, and T. Takumi, Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system, *Sci. Rep.* **6**, 19675 (2016).
- [85] S. Gandhi, M. Haeussler, F. Razy-Krajka, L. Christiaen, and A. Stolfi, Evaluation and rational design of guide RNAs for efficient CRISPR/Cas9-mediated mutagenesis in *Ciona*, *Dev. Biol.* **425**, 8 (2017).
- [86] X. Wu, D. A. Scott, A. J. Kriz, A. C. Chiu, P. D. Hsu, D. B. Dadon, A. W. Cheng, A. E. Trevino, S. Konermann, S. Chen, *et al.*, Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells, *Nat. Biotechnol.* **32**, 670 (2014).
- [87] L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, and W. A. Lim, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression, *Cell* **152**, 1173 (2013).
- [88] S. W. Cho, S. Kim, Y. Kim, J. Kweon, H. S. Kim, S. Bae, and J.-S. Kim, Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases, *Genome Res.* **24**, 132 (2014).
- [89] C. Long, J. R. McAnally, J. M. Shelton, A. A. Mireault, R. Bassel-Duby, and E. N. Olson, Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA, *Science* **345**, 1184 (2014).
- [90] R. L. Frock, J. Hu, R. M. Meyers, Y.-J. Ho, E. Kii, and F. W. Alt, Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases, *Nat. Biotechnol.* **33**, 179 (2015).
- [91] D. Kim, S. Bae, J. Park, E. Kim, S. Kim, H. R. Yu, J. Hwang, J.-I. Kim, and J.-S. Kim, Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells, *Nat. Methods* **12**, 237 (2015).
- [92] D. Kim, S. Kim, S. Kim, J. Park, and J.-S. Kim, Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq, *Genome Res.* **26**, 406 (2016).
- [93] I. M. Slaymaker, L. Gao, B. Zetsche, D. A. Scott, W. X. Yan, and F. Zhang, Rationally engineered Cas9 nucleases with improved specificity, *Science* **351**, 84 (2016).
- [94] Y. Liu, Z. Chen, A. He, Y. Zhan, J. Li, L. Liu, H. Wu, C. Zhuang, J. Lin, Q. Zhang, *et al.*, Targeting cellular mRNAs translation by CRISPR-Cas9, *Sci. Rep.* **6**, 29652 (2016).