Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design

Dong Zhang[a,b,c], Travis Hurst[a,b,c], Dongsheng Duan[d,e,f,g], and Shi-Jie Chen[a,b,c,1]

[a]Department of Physics, University of Missouri, Columbia, MO 65211; [b]Department of Biochemistry, University of Missouri, Columbia, MO 65211; [c]University of Missouri Informatics Institute, University of Missouri, Columbia, MO 65211; [d]Department of Molecular Microbiology and Immunology, School of Medicine, University of Missouri, Columbia, MO 65211; [e]Department of Neurology, School of Medicine, University of Missouri, Columbia, MO 65211; [f]Department of Biomedical Sciences, College of Veterinary Medicine, University of Missouri, Columbia, MO 65211; and [g]Department of Bioengineering, University of Missouri, Columbia, MO 65211

Edited by Ken A. Dill, Stony Brook University, Stony Brook, NY, and approved March 26, 2019 (received for review December 2, 2018)

While CRISPR/Cas9 is a powerful tool in genome engineering, the on-target activity and off-target effects of the system widely vary because of the differences in guide RNA (gRNA) sequences and genomic environments. Traditional approaches rely on separate models and parameters to treat on- and off-target cleavage activities. Here, we demonstrate that a free-energy scheme dominates the Cas9 editing efficacy and delineate a method that simultaneously considers on-target activities and off-target effects. While data-driven machine-learning approaches learn rules to model particular datasets, they may not be as transferrable to new systems or capable of producing new mechanistic insights as principled physical approaches. By integrating the energetics of R-loop formation under Cas9 binding, the effect of the protospacer adjacent motif sequence, and the folding stability of the whole single guide RNA, we devised a unified, physical model that can apply to any cleavage-activity dataset. This unified framework improves predictions for both on-target activities and off-target efficiencies of spCas9 and may be readily transferred to other systems with different guide RNAs or Cas9 ortholog proteins.

RNA folding | free-energy landscape | folding stability | CRISPR | Cas9

The clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein 9 (Cas9) genome-engineering technology is a powerful tool for broad areas of biological significance, where a target DNA strand is edited via a single guide RNA (sgRNA) (1–5). Because the on-target activity (6–10) and off-target effects (10–14) of individual guide RNAs (gRNAs) widely vary, efforts to determine predictive governing parameters for Cas9 on-target activity and specificity in silico are of great interest for broadly applying this technology. In recent years, bioinformatics models derived from data-processing algorithms (6–10, 15, 16) were developed to identify key features that determine on-target activity. For instance, Chari et al. (8) developed an in vivo library-on-library methodology (sgRNA Scorer) to assess sgRNA activity and unravel underlying nucleotide sequence and epigenetic parameters, while Doench et al. (6, 10) devised sgRNA design rules (Azimuth) to create human and mouse genome-wide libraries and perform positive- and negative-selection screens. Conversely, to evaluate and score potential off-target sites, several different bioinformatics models have arisen (10, 12, 17, 18). For example, the CRISPR off-target model developed by Zhang's lab at MIT (MIT_Zhang's model) (12) and CRISPR/Cas9 target online predictor (CCTop) (18) use empirically determined scoring algorithms to quantify off-target cleavage, while Doench et al. (10) proposed the cutting frequency determination (CFD) score to calculate the off-target potential of sgRNA-DNA interactions.

While gRNA sequence plays a vital role in determining Cas9 on-target activity and off-target efficiency, many studies have indicated that these results also depend on experimental conditions (6–14, 19, 20), such as cell type, species, delivery modality, and dosage. In previous bioinformatics-based models, the algorithms used to predict on-target activity or off-target efficiency were usually empirically determined and lack a definite physical foundation. Although various features related to gRNA sequence and experimental conditions were shown to affect Cas9 on-target activity (6–10, 15, 16), a systematic analysis of the general features over mixed experimental systems is absent. One concerning aspect of previous models is that special parameters related to particular experimental measurements may be given overestimated importance. An additional issue is the possible overfitting of the experimental data, especially when only one or two datasets are used to train and test the model. Overall, discrepancies in the identified key features that determine on- and off-target cleavage efficacy from previous models inhibit deep understanding of the cleavage mechanism and preclude further optimization of gRNA design.

Here, inspired by the insights garnered from previous studies and relying on a deduced physical perspective of the CRISPR-Cas9/R-loop complex, we report a unified framework (uCRISPR) to evaluate the Cas9 on-target activity over various experimental datasets and to concurrently consider off-target effects. Compared with previous bioinformatics models, this tool is expected to foster several advances. (*i*) The uCRISPR model bridges the gap between structural experiments and biological functions for the CRISPR/Cas9 system, where a physical framework with a free-energy foundation is presented to evaluate the on- and off-target cleavage efficacy. Thus, this tool can help us gain insights into the key elements that determine Cas9 editing efficacy and optimize sgRNA design for favorable results. (*ii*) Because evaluations of Cas9 off-target effects are considered alongside on-target activity within a

Significance

Evaluation of the cleavage efficacy, including the on-target activity and off-target effects, for individual guide RNA (gRNA) in silico can help optimize application of CRISPR/Cas9 systems. Many bioinformatics models based on data-processing algorithms have been developed, but discrepancies in the identified key features that determine cleavage efficacy inhibit deep understanding of the cleavage mechanism and preclude further optimization of gRNA design. Here, we present a physical framework with rigorous free-energy analysis to bridge the gap between experimental structural studies and cleavage-efficacy evaluations. This tool simultaneously considers on-target activity and off-target effects in a unified framework, improves the prediction power in both realms for diverse spCas9 cleavage efficacy datasets, and is readily transferred to other CRISPR/Cas9 systems.

Author contributions: D.Z. and S.-J.C. designed research; D.Z. performed research; D.Z., T.H., D.D., and S.-J.C. analyzed data; and D.Z., T.H., and S.-J.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The source code for the uCRISPR algorithm and the original datasets have been deposited in GitHub, https://github.com/Vfold-RNA/uCRISPR.

[1]To whom correspondence should be addressed. Email: chenshi@missouri.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1820523116/-/DCSupplemental.

Published online April 15, 2019.

unified framework, the uCRISPR model accounts for off-target effects due to DNA recognition, sgRNA loading, DNA cleavage induced by Cas9/R-loop interactions, and free-energy variations caused by mismatches in RNA/DNA hybrids. In fact, whether the RNA/DNA hybrid contains mismatches or not, cleavage of the target DNA sequence should follow the same mechanism at both on- and off-target sites. Therefore, when free-energy fluctuations caused by mismatches are properly considered, a straightforward unified treatment of Cas9 on- and off-target cleavage efficacy is attainable. (*iii*) All of the parameters involved in the tool are physically meaningful and generalizable for a variety of experimental systems. Hence, this tool effectively avoids overestimating the importance of specific experimental conditions and recognizes general features contributing to Cas9 cleavage efficacy. (*iv*) The diversity in isolated experimental systems employed to parameterize this unified framework can average the effects from specific experimental conditions and largely reduce the risk of overtraining the model over a particular system, which allows the uCRISPR model to more reliably design gRNA in other independent experiments. Independent tests on various experimental datasets indicate that this robust tool improves predictions for both the on-target activity and off-target efficiency for the *Streptococcus pyogenes* Cas9 (spCas9). (*v*) Machine learning-based approaches often utilize rules located in a "black box" to evaluate Cas9 on-target activity, and the reason why these rules work is often not straightforward. In contrast, an analytical expression, gleaned from physics-based logic, is clearly presented here to evaluate on- and off-target cleavage efficacy. Additionally, all of the energy parameters in this expression are directly determinable, leaving only a few parameters related to the experimental conditions to be trained. Therefore, this tool can be readily applied to experiments with modified gRNA sequence lengths or other Cas9 orthologs after reparameterization.

Results and Discussion

A Free-Energy Scheme Dominates Cas9 Editing Efficacy. Previous structural studies have shed light on the molecular mechanism by which the Cas9-sgRNA complex recognizes and cleaves target DNA (21–25). Upon binding of the protospacer adjacent motif (PAM), Cas9 undergoes a conformational change, thereby triggering R-loop formation and duplex unwinding by interacting with the +1-phosphate group in the target DNA strand. Higher thermal stability of the RNA-DNA heteroduplex drives the unidirectional unzipping of the DNA duplex and formation of the RNA/DNA hybrid from the PAM-proximal end. Extensive protein–nucleic acid interactions direct the nontarget DNA strand into the protein and induce further local conformational changes in Cas9 to prepare for Cas9-mediated DNA cleavage.

Viewing the recognition and cleavage mechanism from a physical viewpoint, we conclude that the free-energy change for the formation of the R-loop structure ($\Delta G_{\mathrm{R-loop}}$) under Cas9 binding dictates Cas9 editing efficacy (Fig. 1). The free-energy change is the main driving force for unzipping the DNA duplex, forming the RNA/DNA hybrid (the R-loop structure) and facilitating Cas9 editing activity (26). According to the nearest-neighbor model for nucleic acids (27), the free-energy change is empirically determined by a set of base-stacking energy parameters for the RNA/DNA hybrid and the DNA helix. During the formation of the R-loop structure, rich protein/nucleic acid interactions between Cas9 and the R-loop initiate the unwinding of the DNA duplex (22) and induce local conformational changes in Cas9 to prepare for DNA cleavage (24). Thus, those Cas9/R-loop interactions will also affect the Cas9 editing performance, and this effect can be integrated into the free-energy change when the aforementioned base-stacking energy parameters are considered as both sequence- and position-dependent.

For on-target sites, the free-energy change $\Delta G_{\mathrm{R-loop}}$ can be simplified to a sum of canonical dinucleotide sequence ($\tilde{\mathrm{X}}\tilde{\mathrm{Y}}$) and position ($i$)-dependent energy parameters $E_i^{\mathrm{C}}(\tilde{\mathrm{X}}\tilde{\mathrm{Y}})$, where $\tilde{\mathrm{X}}$ and $\tilde{\mathrm{Y}}$

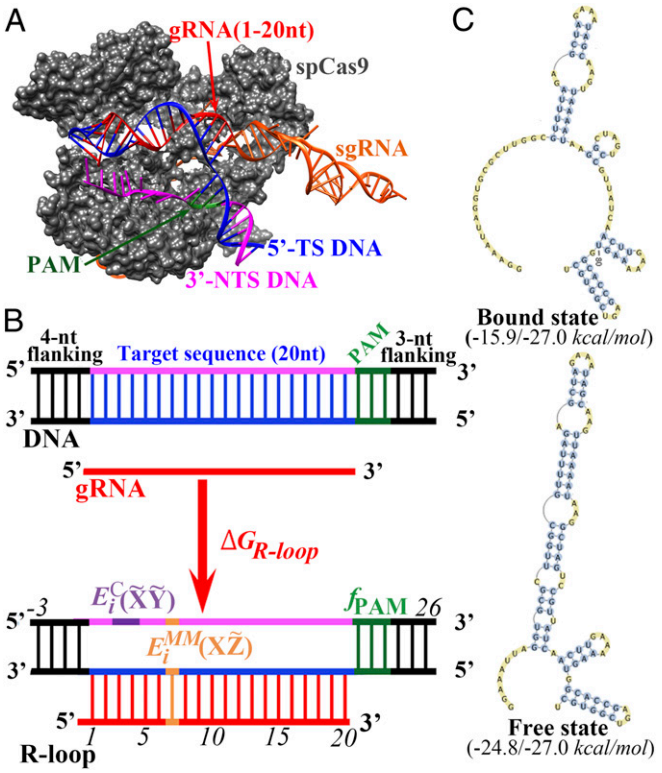

Fig. 1. Key factors considered in the unified physical framework to determine editing efficacy of Cas9. (*A*) Structural information of Cas9-sgRNA-dsDNA ternary complex (24) (Protein Data Bank ID code 5F9R). After recognition of the PAM (dark green), the first 20-nt segment of sgRNA (gRNA, red) is paired with the unwinding target DNA strand (TS DNA, blue) to form the RNA-DNA heteroduplex, and the nontarget DNA strand (NTS DNA, pink) is displaced inside the Cas9 protein (gray). The R-loop structure has complex interactions with the Cas9 protein. (*B*) Schematic illustrating shows the free-energy change for formation of the R-loop structure under Cas9 binding $\Delta G_{\mathrm{R-loop}}(E_i^{\mathrm{C}}, E_i^{\mathrm{MM}})$. The free energy associated with the formation of the Cas9/R-loop complex, along with the efficiency f_{PAM} associated with the relative activity of the PAM sequence, dominates the evaluation of the CRISPR/Cas9 editing efficacy. On-target energy parameters $E_i^{\mathrm{C}}(\tilde{\mathrm{X}}\tilde{\mathrm{Y}})$ integrate the dinucleotide sequence- and position-dependent effects from rich Cas9/R-loop interactions into the free-energy change, while mismatch energy parameters $E_i^{\mathrm{MM}}(\mathrm{X}\tilde{\mathrm{Z}})$ account for sequence- and position-dependent variations in the free-energy change caused by mismatches in the RNA/DNA hybrid. While the nucleotide at the ith position of gRNA is represented by X, the nucleotide at the ith position of the nontarget DNA strand is labeled by $\tilde{\mathrm{Y}}$ or $\tilde{\mathrm{Z}}$. (*C*) The bound state extracted from the crystal structure (21) ($\Delta G_{\mathrm{sgRNA}}^{\mathrm{bound}} = -15.9$ kcal/mol) for the whole sgRNA has a higher free energy than the optimal secondary structure in the free state ($\Delta G_{\mathrm{sgRNA}}^{\mathrm{free}} = -24.8$ kcal/mol). Selection of the bound state from the conformation ensemble ($\Delta G_{\mathrm{sgRNA}}^{\mathrm{ensemble}} = -27.0$ kcal/mol) is required for sgRNA loading and is critical for evaluation of Cas9 editing efficacy.

denote the nucleotides on the nontarget DNA strand at sites i and $i + 1$, respectively (Fig. 1*B*, *Methods*, and *SI Appendix*, Eqs. **S1**–**S4**). For off-target sites, the presence of mismatches in the RNA/DNA hybrid will break the sequence complementarity and directly alter the free-energy change during R-loop formation. Additionally, these mismatches will indirectly modify the free energy of Cas9 binding (the Cas9/R-loop interactions). Furthermore, since the RNA/DNA hybrid is unidirectionally zipped from the PAM-proximal end and because formation of the RNA-DNA heteroduplex is kinetically controlled (28, 29), the presence of mismatches in the RNA/DNA hybrid will also affect the heteroduplex-formation kinetics and cause variation in the editing efficiency of Cas9. Therefore, to account for those effects in the evaluation of the off-target efficiency, a series of sequence- and position-dependent

mismatch energy parameters $E_i^{MM}(X\tilde{Z})$ are combined with eight additional parameters to calculate the free-energy change at off-target sites. More details for the calculation of ΔG_{R-loop} can be found in *Methods* and *SI Appendix*.

Not only does the uCRISPR model properly account for the Cas9/R-loop interaction energies and mismatch energy corrections, but our unified physical framework also considers two other factors to determine the Cas9 on- and off-target cleavage efficacy. The first factor considers effects related to alternative PAM sequences (f_{PAM}). Apart from the canonical NGG PAM, some alternative PAM sequences can also lead to notable but smaller rates of spCas9 activity (10). For example, compared with the NGG PAM sequence, the NAG PAM and NGA PAM sequences have cleavage activities of about 26 and 7% (10), respectively. The second factor incorporates the selection of the bound state from the conformational ensemble of the entire sgRNA. From previous structural studies (21, 24, 30), loading of the sgRNA by the Cas9 protein may require the sgRNA to adopt a specific scaffold (Fig. 1*C*). However, for the whole sgRNA sequence, this specific bound state (native structure) is not always the most stable secondary structure in the free (unbound) form of the sgRNA. For example, as shown in Fig. 1*C*, the folding free energy for a given sgRNA in the bound state (21) ($\Delta G_{sgRNA}^{bound} = -15.9$ kcal/mol) is much higher than the free energy for the lowest free-energy structure in the free state ($\Delta G_{sgRNA}^{free} = -24.8$ kcal/mol) and the free energy for the whole conformational ensemble ($\Delta G_{sgRNA}^{ensemble} = -27.0$ kcal/mol). Therefore, the folding stability of the bound state, $\Delta G_{sgRNA}^{bound} - \Delta G_{sgRNA}^{ensemble}$, may also affect the loading of the sgRNA by Cas9 and alter the Cas9 editing activity (9). For different spacer sequences (20 nt) on the sgRNA, the free energies of the bound state may not be notably different, but the ensemble energies for the whole sgRNA $\Delta G_{sgRNA}^{ensemble}$ can significantly vary. For simplicity, we use a parameter $w\Delta G_{sgRNA}^{ensemble}$ (w is a weight coefficient) to account for the effect of sgRNA-bound structure-folding stability in the evaluation of Cas9 on-target activity and off-target efficiency.

Overall, three factors are considered to control the Cas9 editing activity in our unified physical framework: the stability of the bound conformation relative to the ensemble for the sgRNA $w\Delta G_{sgRNA}^{ensemble}$, the effect of PAM sequences f_{PAM}, and the free-energy change for the formation of R-loop structure ΔG_{R-loop}. The (unified) scoring function to evaluate the Cas9 on-target activity and off-target efficiency can be written as

$$S = f_{PAM} \cdot e^{-\left(\Delta G_{R-loop} - w\Delta G_{sgRNA}^{ensemble}\right)/k_B T}. \quad [1]$$

Here, ΔG_{R-loop} is the sum of the given energy parameters (*Methods* and *SI Appendix*, Eqs. **S1–S13**), $\Delta G_{sgRNA}^{ensemble}$ can be calculated by RNA secondary structure-folding programs, f_{PAM} has been experimentally measured (10), k_B is the Boltzmann constant, T is the temperature, and $k_BT = 0.59$ kcal/mol at T = 25 °C.

The uCRISPR Model Favorably Evaluates On-Target Activity. The deduced scoring function (Eq. **1**) in the uCRISPR algorithm was parameterized using five different experimental systems (6, 8, 10, 15) (see *Methods* and *SI Appendix* for details) and tested on other independent datasets for on-target activity evaluation. As shown in Fig. 2, the proposed uCRISPR model captures the general features that determine Cas9 on-target activity over various experimental datasets and possesses favorable performance in comparison with popular bioinformatics models. For the five training datasets (Fig. 2*A*), our uCRISPR model yields an average Spearman rank correlation (±SD) between experimental activities and predicted scores of 0.55 ± 0.07, which is comparable to that given by Azimuth (average: 0.55 ± 0.13) and an improvement over sgRNA Scorer (average: 0.34 ± 0.09). To further prove the generality of the unified physical framework, we tested the performance of the uCRISPR model on 10 additional isolated experimental datasets (9, 15, 19, 31–33) (*SI Appendix*, Table S1). As shown in Fig. 2*B*, for 8 of the 10 test cases, the uCRISPR model outperforms the other models. For the remaining two cases, uCRISPR gives an intermediate prediction, so it falls between the two popular bioinformatics models. The average Spearman rank correlations over the 10 test datasets given by sgRNA Scorer, Azimuth, and uCRISPR models are 0.18 ± 0.05, 0.26 ± 0.11, and 0.30 ± 0.10, respectively. For the evaluation of on-target activity, these results clearly demonstrate that our unified physical framework recognizes general features that control Cas9 performance over diverse experimental conditions and performs favorably in comparison with previous bioinformatics models. For statistical significance, only experimental datasets that contain more than 100 gRNAs are considered here. Extended comparisons between our uCRISPR model and other bioinformatics models over the 10 additional experimental datasets are given in *SI Appendix*, Table S1.

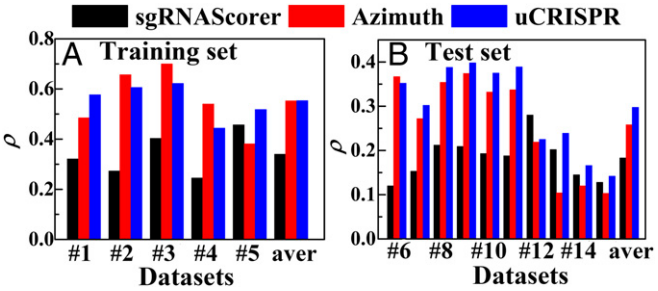

Fig. 2. Performance of on-target activity prediction. To compare with bioinformatics-based models, the Spearman rank correlations (ρ) between the experimental activities and predicted activity scores are plotted for the training set (*A*) and the test set (*B*). The experimental datasets include Wang/Xu HL60 (15) (#1, 2076 gRNA sequences), Doench MOLM13/NB4/TF1 (6) (#2, 881), Doench Mouse EL4 (6) (#3, 951), Doench A375/AZD (10) (#4, 2333), Chari 293T (8) (#5, 1234), Koike-Yusa/Xu Mouse ESC (15) (#6, 907), Hart Rpe (31) (#7, 4149), Hart Hct116-1 Lib1 (31) (#8, 4226), Hart Hct116-2 Lib1 (31) (#9, 4172), Hart HeLa Lib1 (31) (#10, 4189), Hart HeLa Lib2 (31) (#11, 3809), Varshney Zebrafish (32) (#12, 102), Gagnon Ciona (33) (#13, 111), Moreno-Mateos Zebrafish (9) (#14, 1020), and Shkumatava Zebrafish (19) (#15, 163). The columns labeled "aver" show the average Spearman rank correlations over the training and test sets, respectively. More details are given in *SI Appendix*, Table S1.

The uCRISPR Model Improves Off-Target Effect Predictions. After further parameterizing the scoring function given in Eq. **1** using 18 off-target datasets (10, 12, 34) (*Methods* and *SI Appendix*), the performance of the uCRISPR model on off-target effect prediction was validated using 51 independent experimental datasets. Not only do these datasets contain single and multiple mismatch variants of perfectly matched gRNA spacer sequence, but they also include genome-wide off-target sites (*SI Appendix*, Table S2). We used the Pearson correlation coefficient r between experimental efficiencies and predicted scores to measure the performance in evaluation of off-target effects (*SI Appendix*, Eq. **S14**). Compared with popular bioinformatics models, the uCRISPR model remarkably improves off-target effect predictions. As shown in Fig. 3*A*, the uCRISPR model gives successful predictions (with Pearson correlation $r \geq 0.5$) for about 82% of those test datasets, which is higher than the success rates found using MIT_Zhang's model, CCTop, CROP-IT, and CFD, which yield 33%, 47%, 45%, and 78%, respectively. When the success threshold is increased to higher correlations, the differences between the uCRISPR model and the four other models are even more profound. For instance, the success rates $P(r)$ with Pearson correlations $r \geq 0.7$ for MIT_Zhang's model, CCTop, CROP-IT, CFD, and uCRISPR are about 14%, 18%, 18%, 63%, and 71%, respectively. Furthermore, our uCRISPR model gives highly reliable predictions (with Pearson

correlations $r \geq 0.9$) for 17 of 51 cases, which is much better than the six cases for CFD, two cases for CCTop, one case for CROP-IT, and zero cases for MIT_Zhang's model. For the other metrics, as shown in Fig. 3*B*, the average Pearson correlations ($\bar{r} \pm$ SD) over all 51 datasets are 0.38 ± 0.28 (MIT_Zhang's model), 0.43 ± 0.28 (CCTop), 0.47 ± 0.23 (CROP-IT), 0.68 ± 0.24 (CFD), and 0.75 ± 0.24 (uCRISPR), respectively. For each model, the fraction of best predictions F_{best} was calculated according to whether the model obtained the highest Pearson correlation on each dataset (Fig. 3*B*). The uCRISPR model achieved the best predictions on about 67% of datasets, which shows significant improvement over the 4%, 13%, 6%, and 10% yielded by MIT_Zhang's model, CCTop, CROP-IT, and CFD, respectively. For the evaluation of off-target efficiency, our investigation indicates that uCRISPR performs favorably in comparison with previous bioinformatics models.

For the failed cases, where low Pearson correlations ($r < 0.4$) were found using the uCRISPR model, most have large numbers of off-target sites (>100) and multiple mismatches (>3) in the RNA/DNA hybrid, and the other bioinformatics models also give poor predictions for all of them. Those failures indicate that some factors are missed in our physical framework, such as the effects of multiple mismatches on RNA/DNA hybrid-formation kinetics and the accessibility of the target site in chromatin. With more experimental data containing multiple mismatches, the former can be resolved by integrating the formation kinetics into Eq. **1**. The latter can be settled when the (rough) 3D structure of the corresponding chromatin is available.

Insights from Energy Parameters. To evaluate the qualitative insights that may be extracted from physically meaningful uCRISPR quantities, we sought to uncover significant sequence and position features underlying the energy parameters. As shown in Fig. 4*A*, for the Cas9 on-target activity, the SD $\sigma(E_i^C)$ of the on-target energy parameters at different positions indicates that the PAM-proximal region (positions 10–20) of the R-loop structure is more sequence-sensitive than the PAM-distal region (positions 1–9), especially for the four positions (positions 17–20) neighboring the PAM. Two mechanisms may help us understand this phenomenon. First, previous structural studies (21, 24, 30) have indicated that the Cas9/R-loop interactions in the PAM-proximal region are richer than that in the PAM-distal region. Thus, the sequence importance in the PAM-proximal region will be enhanced as the role of Cas9/R-loop interactions in determining Cas9 editing activity is assumed to be sequence-dependent. Second, the kinetic importance of the free-energy change in R-loop formation decreases for nucleotides away from the PAM-proximal end because the formation of the RNA/DNA hybrid is unidirectional from the PAM-proximal end (21, 24). Therefore, the free-energy differences for various dinucleotide sequences in the PAM-proximal region are more sensitive than those in the PAM-distal region. Additionally, sequence features consistent with previous studies are also found, such as the favorability of dinucleotide sequence $\tilde{G}\tilde{G}$ at position 20 for promoting Cas9 cleavage (9), which is indicated by the minimal E_i^C at position 20 (Fig. 4*A*, *Inset* and *SI Appendix*, Table S3). Also, the dinucleotide sequence $\tilde{T}\tilde{T}$ is the least favorable (highest in the energy parameters) and is depleted at most positions (Fig. 4*A*, *Inset* and *SI Appendix*, Table S3). For off-target performance, the importance of the proximity of mismatches to PAM is seen by considering the average of mismatch energy parameters $< E_i^{MM} >$ on different positions (Fig. 4*B* and *SI Appendix*, Table S4). Position-dependent averaging of mismatch energies indicates that the mismatches in the RNA/DNA hybrid are tolerated to a greater extent in the PAM-distal region (low values of $< E_i^{MM} >$) than in the PAM-proximal region (high values of $< E_i^{MM} >$) (2, 12). As with the on-target parameters, the sensitivity to mismatch sequences in the PAM-proximal region is larger than that in the PAM-distal region, which is confirmed by the larger deviations $\sigma(E_i^{MM})$ in the PAM-proximal region (Fig. 4*B*, *Inset* and *SI Appendix*, Table S4). The integration of Cas9/R-loop interactions and kinetic effects into uCRISPR would simultaneously account for those position and sequence features.

To further understand the sequence features underlying the uCRISPR unified framework for on-target activity evaluation, the average on-target energy parameters ($< E^C > /k_BT$) for various dinucleotide sequences over positions involved in the R-loop structure (positions 1 through 19) were compared with the folding free-energy changes for the formation of the

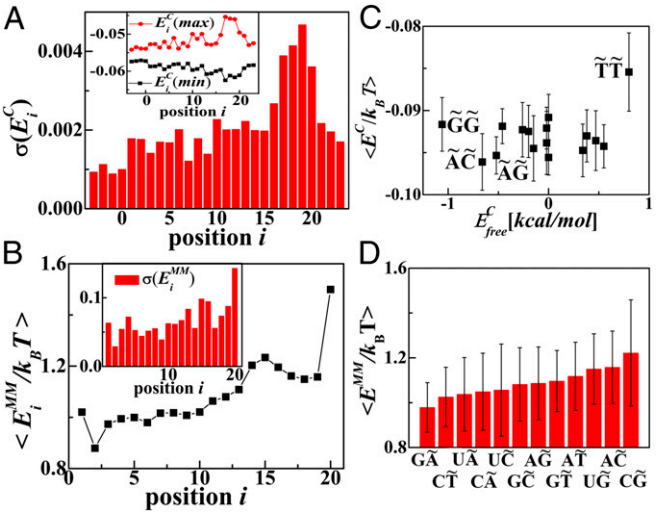

Fig. 4. Statistical analysis of the deduced energy parameters. (*A*) SDs of the on-target energy parameters $\sigma(E_i^C)$ at different positions indicate that the Cas9 on-target activity is more sequence-sensitive in the PAM-proximal region (positions 10–20) than in the PAM-distal region (positions 1–9). *Inset* shows the most-favorable $E_i^C(min)$ and least-favorable $E_i^C(max)$ dinucleotide sequence for each position. The last three columns indicate positions 21_NGGN, 24, and 25. (*B*) Average mismatch energy parameters $< E_i^{MM}/k_BT >$ at each position show that mismatches in the RNA/DNA hybrid are more tolerated in the PAM-distal region than in the PAM-proximal region. SDs $\sigma(E_i^{MM})$ in *Inset* show the mismatch sequence sensitivity at different positions. (*C*) Comparison of the averaged on-target energy parameters $< E^C/k_BT >$ for various dinucleotide sequences over positions involved in the R-loop structure (positions 1 through 19) with respect to the folding free-energy changes for formation of corresponding R-loop structures in the free state E_{free}^C. (*D*) Averaged mismatch energies $< E^{MM}/k_BT >$ over positions 1–20 for various mismatch sequence types. All error bars are SDs of the property of interest.

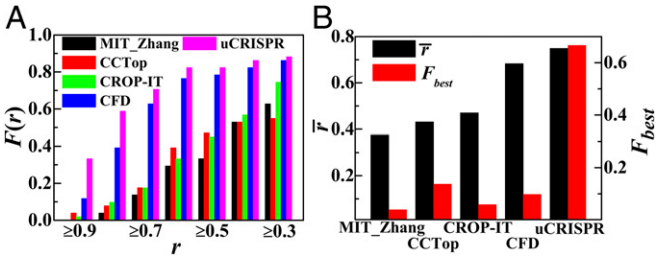

Fig. 3. Performance of off-target efficiency prediction. To compare uCRISPR with other models, the Pearson correlations r between experimental efficiencies and predicted scores were measured. (*A*) Success rates of the models with Pearson correlation meeting certain thresholds (such as ≥0.5) over all of the 51 test cases are plotted as fractional success rates $F(r)$. (*B*) Overall comparisons between different models with average correlation $\bar{r}$ (left axis) and the fraction of best prediction (highest correlation among all models for each dataset) F_{best} (right axis) over 51 test cases as metrics. More details are presented in *SI Appendix*, Table S2.

corresponding R-loop in the free state (E^{C}_{free}) (Fig. 4*C*). To some extent, $< E^{C}/k_BT >$ represents the relative importance of the related interactions (characterized by the dinucleotide sequence) in the evaluation of on-target activity. A lower value of $< E^{C}/k_BT >$ results in a lower free-energy change $\Delta G_{\mathrm{R-loop}}$ and a greater relative contribution to the on-target cleavage activity. Although the G̃G̃ dinucleotide sequence has the lowest folding free-energy change in the free state (Fig. 4*C* and *SI Appendix*, Fig. S1), the high occurrence of GG dimers in the gRNA is thought to inhibit cleavage activity because guanine-rich sequences may have a propensity to form the G-quadruplex structure that makes the gRNA less accessible for target DNA recognition (9, 16). Thus, the overall relative importance of G̃G̃ dinucleotides is low. Instead, the ÃC̃ and ÃG̃ dinucleotide sequences that have the second and third lowest folding free-energy changes in the free state contribute the most with overall high relative importance in on-target activity prediction. In contrast, the T̃T̃ dimer is the most unfavorable and has the least relative importance in a majority of positions because it has the highest folding free-energy change in the free state. The remaining cases are somewhat complicated since they are affected by both the free-energy change in the free state and the presence of rich interactions between the protein and the R-loop.

To similarly explain the mismatch sequence features involved in the unified framework for off-target efficiency evaluation, the averaged mismatch energy parameters ($< E^{MM}/k_BT >$) over positions 1–20 for various mismatch sequence types are shown in Fig. 4*D*. In general, $< E^{MM}/k_BT >$ represents the (positive) folding free-energy correction during R-loop structure formation caused by a related mismatch in the RNA/DNA hybrid, and a lower value of $< E^{MM}/k_BT >$ means the related mismatch is more tolerated for off-target efficiency evaluation. Since the mismatch type GÃ (G:T̃-Ã) contains a wobble-like base pair interaction G:T̃ in the RNA/DNA hybrid, which is more energetically favorable than other noncanonical base pairs, it is more well tolerated than other mismatch types and is seen to be the most-tolerated mismatch sequence at most positions (*SI Appendix*, Table S4).

Performance on New Systems. To further check the efficacy of the present unified framework on totally new on- and off-target systems that are not in any of the above datasets used for training or testing, we applied the uCRISPR model to evaluate three totally new on-target systems (35, 36) and 12 isolated off-target datasets (37, 38), and the results are given in *SI Appendix*, Fig. S3. For the three new on-target systems, the uCRISPR model provides better predictions than the other two popular models (*SI Appendix*, Fig. S3*A*). The average Spearman rank correlations over these datasets given by sgRNA Scorer, Azimuth, and uCRISPR are 0.138, 0.163, and 0.214, respectively. For the 12 isolated off-target datasets, uCRISPR yields the best predictions for nine cases, while the CFD model gives the best predictions for the remaining three cases (*SI Appendix*, Fig. S3*B*). The average Pearson correlations for those 12 datasets given by CCTop, CROP-IT, MIT_Zhang, CFD, and uCRISPR are 0.127, 0.229, 0.392, 0.490, and 0.591, respectively. In all, for those totally new systems, our uCRISPR algorithm provides overall better predictions than the current state-of-the-art approaches in both on- and off-target editing-efficacy evaluation.

Conclusion

Since Cas9 editing efficacy varies for different gRNA sequences, cell cultures, and experimental conditions, a systematic analysis of the general features that determine Cas9 on- and off-target cleavage efficacy over diverse experimental systems is essential to optimize sgRNA design. Based on the molecular mechanism for Cas9-mediated DNA cleavage, we explored the relationship between structural/energetics information and biological functions and developed a unified physical framework to generally examine the experimental on-target activities over varied datasets and simultaneously evaluate off-target effects. The proposed unified physical framework considers three factors to evaluate the Cas9 editing efficacy: the free-energy change of formation for the R-loop structure under Cas9 binding, the bound-state (native-structure) selection from the conformational ensemble for the whole sgRNA, and the relative activities of alternative PAMs.

As a common set of energy parameters with physical definitions is used to analytically express the on-target activity over diverse experimental systems, the general features underlying sgRNA efficacy are well captured and interpreted by the unified framework, while effects from special experimental conditions are averaged. Compared with previous machine learning-based approaches, the uCRISPR model uses fewer energetic parameters with well-defined meanings and sundry experimental systems for parameterization to largely reduce the risk of overfitting. When required, the number of energy parameters involved in the unified framework can be further reduced without losing much in terms of performance (*SI Appendix*, Fig. S2). Moreover, our ability to consider both on- and off-target cleavage efficacy in the same framework allows us to better account for variations caused by mismatches in off-target site predictions that affect target DNA recognition, sgRNA loading, and kinetic RNA/DNA hybrid formation. Tests on numerous isolated datasets show that our unified physical framework improves predictions for both on-target activity and off-target efficiency. Taken together, these aspects suggest that our unified physical framework can facilitate the design of optimal gRNA with high activity and specificity for use with genome-engineering technology.

In general, a physically grounded approach, such as the uCRISPR model, can provide mechanistic insights to understand on- and off-target CRISPR Cas9 cleavage activity. When the functional mechanism of systems like CRISPR Cas9 are available, physical approaches are more reliable, transparent, and informative than unprincipled applications of data-processing methods, such as machine-learning. In comparison with previous bioinformatics-based machine-learning approaches, our physical approach yields better results and provides insights into the importance of specific interactions that should be considered in the evaluation of cleavage efficacy, which lays down a foundation for further optimization of sgRNA design. At this stage, to establish a universal physical approach, only the position-dependent sequence information is considered in the uCRISPR unified framework, while effects from other specific experimental conditions are averaged and indirectly integrated into the model. Some specific experimental conditions, such as cell type and concentrations, may affect the kinetics of the R-loop formation and/or the interaction pattern between the Cas9 protein and R-loop structure. Thus, to better capture the effects from specific experimental conditions and to improve performance, the present model to evaluate the on/off-target cleavage efficacy may require alteration to achieve optimal predictive power.

Recently, modified gRNAs with longer or shorter spacers were studied to improve Cas9 specificity (39–41), and chemical modifications (42, 43) such as 2′-*O*-methylation, 3′phosphorothioation, and 2′-fluorination for the gRNAs were reported to alter CRISPR-Cas genome-editing efficacy. Additionally, bulge loops caused by insertions or deletions in target DNA compared with gRNA in CRISPR/Cas9 systems were shown to affect off-target activity (44). Furthermore, a broad array of Cas9 orthologs (45–47) or new CRISPR proteins (48, 49) have been engineered as genome-editing tools. However, few theoretical efforts have attempted to uncover factors contributing to Cas9 activity and specificity for those systems. Because of the fundamental physical nature of our scoring function, the general approach described here provides a method to overcome these limitations, and straightforward application of our framework to those systems is possible after reparameterization.

Methods

Calculation of Free-Energy Change. According to the nearest-neighbor model for the free-energy calculation of nucleic acids, the free-energy change for R-loop formation can be roughly reduced to the sum of a series of dinucleotide-energy parameters (base-stacking interaction energies) with the effects from Cas9/R-loop interactions included (*SI Appendix*, Eqs. **S1–S10**). For off-target sites, the energies for noncanonical base-stacking interactions containing mismatches are decomposed to reduce the number of unknown energy parameters (*SI Appendix*, Eqs. **S5–S10**). The ensemble energy $\Delta G^{\text{ensemble}}_{\text{sgRNA}}$ for the whole sgRNA with a specific gRNA sequence is calculated using the RNAstructure package (50), in which only the non–cross-linked structures are considered. Cross-linked structures, such as the pseudoknots as shown in *SI Appendix*, Fig. S5, can be predicted by, for example, the Vfold model (51), but to facilitate computational speed, they are not included here.

Parameterization of the uCRISPR Model. Assuming the on- and off-target scores given by the uCRISPR algorithm (Eq. **1**) linearly scale with the experimentally determined activity (A_{exp}), we let

$$A_{\text{exp}} = aS + b, \qquad [2]$$

where a and b are factors that depend on experimental systems. Assuming we have predefined values for b and that the total number of unique data points used to parameterize the model over diverse experimental systems is much larger than the number of unknown parameters, all of the energy parameters involved in the uCRISPR model can be directly determined using singular value decomposition (52) (*SI Appendix*, Eqs. **S11–S13**).

Experimental Datasets and Scoring Functions. All of the experimental datasets used to train and test the present algorithm are from previous public works, and a complete separation between the training and testing data sets was made (*SI Appendix*). For off-target effect predictions by CCTop (18), CROP-IT (17), and MIT_Zhang's model (12), we rebuilt the scoring functions in C++ based on the descriptions in corresponding articles. For the CFD (10) scoring function, we used the code provided in the publication.

Availability. The source code for the uCRISPR algorithm to evaluate on-target activity and off-target efficiency and the original datasets to train and test the model are all available at rna.physics.missouri.edu/uCRISPR/index.html (53). A web server located at the same website is under construction.

ACKNOWLEDGMENTS. This research was supported by NIH Grants GM063732 (to S.-J.C.), GM117059 (to S.-J.C.), and AR-69085 (to D.D.); the National Science Foundation Graduate Research Fellowship Program under Grant 1443129 (to T.H.); Department of Defense Grant MD150133 (to D.D.); and the Jackson Freel DMD Research Fund (D.D.).

1. Mali P, et al. (2013) RNA-guided human genome engineering via Cas9. *Science* 339:823–826.
2. Cong L, et al. (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* 339:819–823.
3. Hsu PD, Lander ES, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 157:1262–1278.
4. Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* 32:347–355.
5. Wright AV, Nuñez JK, Doudna JA (2016) Biology and applications of CRISPR systems: Harnessing nature's toolbox for genome engineering. *Cell* 164:29–44.
6. Doench JG, et al. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* 32:1262–1267.
7. Wang T, Wei JJ, Sabatini DM, Lander ES (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 343:80–84.
8. Chari R, Mali P, Moosburner M, Church GM (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* 12:823–826.
9. Moreno-Mateos MA, et al. (2015) CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods* 12:982–988.
10. Doench JG, et al. (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* 34:184–191.
11. Fu Y, et al. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* 31:822–826.
12. Hsu PD, et al. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* 31:827–832.
13. Ran FA, et al. (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* 154:1380–1389.
14. Guilinger JP, Thompson DB, Liu DR (2014) Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat Biotechnol* 32:577–582.
15. Xu H, et al. (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Res* 25:1147–1157.
16. Wong N, Liu W, Wang X (2015) WU-CRISPR: Characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol* 16:218.
17. Singh R, Kuscu C, Quinlan A, Qi Y, Adli M (2015) Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Res* 43:e118.
18. Stemmer M, Thumberger T, Del Sol Keyer M, Wittbrodt J, Mateo JL (2015) CCTop: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS One* 10:e0124633.
19. Haeussler M, et al. (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* 17:148.
20. Tycko J, Myer VE, Hsu PD (2016) Methods for optimizing CRISPR-Cas9 genome editing specificity. *Mol Cell* 63:355–370.
21. Nishimasu H, et al. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156:935–949.
22. 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.
23. Jiang F, Doudna JA (2015) The structural biology of CRISPR-Cas systems. *Curr Opin Struct Biol* 30:100–111.
24. Jiang F, et al. (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* 351:867–871.
25. Jiang F, Doudna JA (2017) CRISPR-Cas9 structures and mechanisms. *Annu Rev Biophys* 46:505–529.
26. Xu X, Duan D, Chen S-J (2017) CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: From physical mechanism to off-target assessment. *Sci Rep* 7:143.
27. Tinoco I, Jr, et al. (1973) Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol* 246:40–41.
28. Singh D, Sternberg SH, Fei J, Doudna JA, Ha T (2016) Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat Commun* 7:12778.
29. Klein M, Eslami-Mossallam B, Arroyo DG, Depken M (2018) Hybridization kinetics explains CRISPR-Cas off-targeting rules. *Cell Rep* 22:1413–1423.
30. Jiang F, Zhou K, Ma L, Gressel S, Doudna JA (2015) STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* 348:1477–1481.
31. Hart T, et al. (2015) High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 163:1515–1526.
32. Varshney GK, et al. (2015) High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res* 25:1030–1042.
33. Gagnon JA, et al. (2014) Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One* 9:e98186, and erratum (2014) 9:e106396.
34. Tsai SQ, et al. (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* 33:187–197.
35. Horlbeck MA, et al. (2016) Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife* 5:e19760.
36. Labuhn M, et al. (2018) Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Res* 46:1375–1385.
37. Tsai SQ, et al. (2017) CIRCLE-seq: A highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat Methods* 14:607–614.
38. Listgarten J, et al. (2018) Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat Biomed Eng* 2:38–47.
39. Chen B, et al. (2013) Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* 155:1479–1491.
40. Dang Y, et al. (2015) Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. *Genome Biol* 16:280.
41. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* 32:279–284.
42. Hendel A, et al. (2015) Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat Biotechnol* 33:985–989.
43. Yin H, et al. (2017) Structure-guided chemical modification of guide RNA enables potent non-viral in vivo genome editing. *Nat Biotechnol* 35:1179–1187.
44. Lin Y, et al. (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res* 42:7473–7485.
45. Ran FA, et al. (2015) In vivo genome editing using Staphylococcus aureus Cas9. *Nature* 520:186–191.
46. Mali P, et al. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol* 31:833–838.
47. Müller M, et al. (2016) Streptococcus thermophiles CRISPR-Cas9 systems enable specific editing of the human genome. *Mol Ther* 24:636–644.
48. Zetsche B, et al. (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* 163:759–771.
49. Shmakov S, et al. (2015) Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol Cell* 60:385–397.
50. Reuter JS, Mathews DH (2010) RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11:129.
51. Xu X, Zhao P, Chen S-J (2014) Vfold: A web server for RNA structure and folding thermodynamics prediction. *PLoS One* 9:e107504.
52. Press W, Teukolsky S, Vetterling W, Flannery B (2007) Singular value decomposition. *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ Press, New York), 3rd Ed, pp 65–75.
53. Zhang D, Hurst T, Duan D, Chen S-J (2019) Data from "uCRISPR - Unified energetics analysis to evaluate the Cas9 on-target activity and off-target effects." GitHub. Available at https://github.com/Vfold-RNA/uCRISPR). Deposited February 10, 2019.