A mechanistic model improves off-target predictions and reveals the physical basis of *Sp*Cas9 fidelity

Behrouz Eslami-Mossallam[1,*], Misha Klein[1,*], Constantijn v.d. Smagt[1], Koen v.d. Sanden[1], Stephen K. Jones Jr.,[2,3,4] John A. Hawkins,[2,3,4,5], Ilya J. Finkelstein[2,3,4], and Martin Depken[1,**]

[1]Kavli Institute of NanoScience and Department of BioNanoScience, Delft University of Technology, Delft 2629HZ, the Netherlands

[2] Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas 78712, USA

[3] Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas 78712, USA

[4] Center for Systems and Synthetic Biology, University of Texas at Austin, Austin, Texas 78712, USA

[5] Oden Institute for Computational Engineering and Science, University of Texas at Austin, Austin, Texas 78712, USA

* equal contribution

** correspondence: S.M.Depken@tudelft.nl

The *Sp*Cas9 endonuclease has become an important tool in gene-editing and basic science alike. Though easily programmed to target any sequence, *Sp*Cas9 also shows considerable activity over genomic off-targets. Many empirical facts regarding the targeting reaction have been established, but a comprehensive mechanistic description is still lacking—limiting fundamental understanding, our ability to predict off-target activity, and ultimately the safe adaptation of the *Sp*Cas9 toolkit for therapeutics. By mechanistically modelling the *Sp*Cas9 structure-function relationship, we simultaneously capture binding and cleavage dynamics for *Sp*Cas9 and *Sp*-dCas9 in terms of free-energies. When our model is trained on high-throughput data, we outperform state-of-the-art off-target prediction tools. Based on the biophysical parameters we extract, our model predicts the open, intermediate, and closed complex configurations described in single-molecule FRET experiments, and indicates that R-loop progression is tightly coupled to structural changes in the targeting complex. We further show that *Sp*Cas9 targeting kinetics are tuned for extended sequence specificity while maintaining on-target efficiency. Our approach can be used to characterize any other CRISPR derived nuclease, and contrasting future studies of high-fidelity variants with the *Sp*Cas9 benchmark we here provide will help elucidate the determinants of CRISPR fidelity and the path to increased specificity and efficiency in engineered systems.

The use of RNA guided DNA endonuclease CRISPR-Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats – CRISPR associated protein 9) is by now ubiquitous in the biological sciences[1,2], with applications ranging from live-cell imaging[3] and gene knockdown/overexpression[4,5], to genetic engineering[6,7] and gene therapy[8,9]. Due to its relative ease of use, *Streptococcus pyogenes* Cas9 (*Sp*Cas9) has become the most widely applied Cas9 variant to date. By programming *Sp*Cas9 with a ~ 100 nucleotide (nt) long single-guide RNA (sgRNA), the complex can differentially target DNAs based on the level of complementarity to a 20nt segment of the sgRNA[10]. While wildtype *Sp*Cas9 (Cas9 from now on) induces specific double-stranded breaks, catalytically 'dead' Cas9 (dCas9) mutants provide binding specificity without cleavage[3,5]. Apart from complimentary *on-targets*, Cas9-sgRNA also binds and cleaves partially-complementary *off-targets*[11–18]. Such off-targeting risks unwanted genomic alterations, including point mutations, large-scale deletions, and chromosomal rearrangements[19]. The potentially deleterious effects associated with such editing errors impedes wide-spread implementation of the CRISPR tool kit in human therapeutics.

Off-target risk is typically assessed based on various *in silico* off-target prediction tools. These tools come in bioinformatic[20,21], heuristic[12,14,22,23] and machine learning[24,25] flavors, but all rank genomic sites based on their own unique off-target activity measures. Although ranking approaches might predict strong off-targets, they can never tell how quickly (d)Cas9-sgRNA

differentially (binds) cuts a given fraction of available DNA when applied at a certain concentration. The lack of quantitative prediction is a problem, as *in vivo* Cas9 activity can depend heavily on exposure time, which is also an active area of development for limiting off-target activity[26].

Quantitative predictions of Cas9 activity requires a physical model, but existing physical models[22,30] assume that Cas9-sgRNA binding equilibrium is reached over the entire genome before cleavage catalysis. That binding does not always equilibrate before cleavage[31,32] can be directly inferred from the fact that binding and cleavage activities often correlate rather weakly—as we will show explicitly, and is also known from literature[33–35]. To account for this fact, we construct a comprehensive kinetic model that includes binding and cleavage reactions, and globally train it on two high-throughput *in vitro* datasets that capture each process separately[15]. Our fully parameterized model accurately predicts an independent high-throughput dataset[11], without the use of any additional fitting parameters.

As our model is fully-parameterized in terms of physical quantities, it offers many insights into biophysical mechanisms. Notoriously, dCas9 binds more off-targets than Cas9 cleaves[27,33–41]. By establishing the free-energy landscape of the targeting reaction with any off-target, we show that the difference in binding and cleavage activities stems from a (relatively) long-lived DNA-bound intermediate state. We further show that this state is tuned for both high cleavage specificity and on-target cleavage efficiency. We also connect the binding intermediate to the intermediate HNH-conformation observed in single-molecule FRET experiments[42,43], and argue that the conformational change is driven by R-loop formation. Finally, we show that our physical model outperforms the two best performing off-target prediction tools used today[12,22,44].

Results

Kinetic model simultaneously captures binding and cleavage profiles

To quantitatively describe the outcome of binding and cleavage experiments within a single physical framework, we must model the internal dynamics of the Cas9-sgRNA targeting complex, as well as the three high-throughput experimental assays we intend to use for training and validation.

The first (training) data set is produced using the NucleaSeq (nuclease digestion and deep sequencing) assay, which estimates the effective cleavage rates (k_{clv}) for a library of off-targets by monitoring the fraction of uncut DNA over time[15] (**Supplementary Information**). The second (training) data set is produced with the CHAMP (chip-hybridized association-mapping platform) assay, where we extract the effective association constant (K_{A}) over a library of off-targets exposed to dCas9-sgRNA for 10 min[15,45] (**Supplementary Information**). The third (validation) data set is sourced from published HiTS-FLIP (high-throughput sequencing-fluorescent ligand interaction profiling) experiments[11], which report the effective association rate estimated over 1500 seconds of exposure to dCas9-sgRNA at a 1nM concentration (**Supplementary Information**).

Though the experiments cover different molecules (Cas9 and dCas9), report on different quantities (cleavage rates, association constants, and association rates), and consider different variable sweeps (concentration and time), the single reaction scheme shown in **Fig. 1a** describes them all[31]. From solution, a Cas9-sgRNA complex uses protein-DNA interactions to recognize a 3nt protospacer adjacent motif (PAM) DNA sequence – canonically 5'-NGG-3' [46,47]. Binding to the PAM sequence opens the DNA double helix, and allows the first base of the target sequence to hybridize with the guide[46,47]. The double helix further denatures as the target strand forms a hybrid structure with the guide RNA, typically referred to as an R-loop[48–51]. The R-loop grows and shrinks with the length of the sgRNA-DNA hybrid, until it is either reversed and Cas9 unbinds, or it reaches completion (a 20 nt hybrid) and is cleaved. Cas9 uses its two nuclease domains (HNH and RuvC) to catalyze the cleavage of both DNA strands[52].

The most general reaction schemes for cleavage and binding are completely parameterized only when we estimate all the rates for every potential guide-target combination (**Fig. 1a**)—a prohibitively large number of parameters for any genome. To render parameter estimation tractable, we make four mechanistic model assumptions: (1) mismatch positions within the hybrid are more important than mismatch types (as can be inferred directly from data[11,15]), and all 12 mismatch types can be treated equally; (2) dCas9 differs from Cas9 only in that dsDNA bond-cleavage catalysis is completely suppressed, and all other rates can be taken to be identical between the two[42,53]; (3) a mismatch influences only the reversal of the mismatched base pairing, leaving all other rates unchanged; (4) all hybrid-bond-formation rates are equal, and independent

of complementarity. Though these four assumptions are over-simplification, they allow us to capture the basic non-equilibrium nature of the problem. We justify them *post hoc* by showing that the targeting dynamics are completely determined by even a much smaller set of effective rates.

We use the detailed-balance condition for microscopic rates (**Supplementary Information**) to define the free-energy of each state in our model (**Fig. 1b**). When extending the R-loop, both gains and losses in free-energy are possible as base-pairing interactions, protein-DNA interactions[53], and any induced conformational changes[29,42,43,52] all contribute to the stability of the Cas9-sgRNA-DNA complex. As we assume that mismatches only facilitate the reversal of the mismatched base pairs, the entire free-energy landscape will rise by a positive amount from the mismatch onwards (c.f. pink and blue free-energy landscapes in **Fig. 1b**). Our model assumptions reduce the total number of parameters to 44: the rate of PAM binding from solution (k_{on}) and its free-energy cost; a single internal forward rate (k_{f}); 20 free-energy costs dictating R-loop progression for matching guide and target; 20 free-energy penalties for mismatches at different R-loop positions; and, for Cas9, the rate at which the final cleavage reaction is catalyzed (k_{cat}) (see **Supplementary Information** for further details).

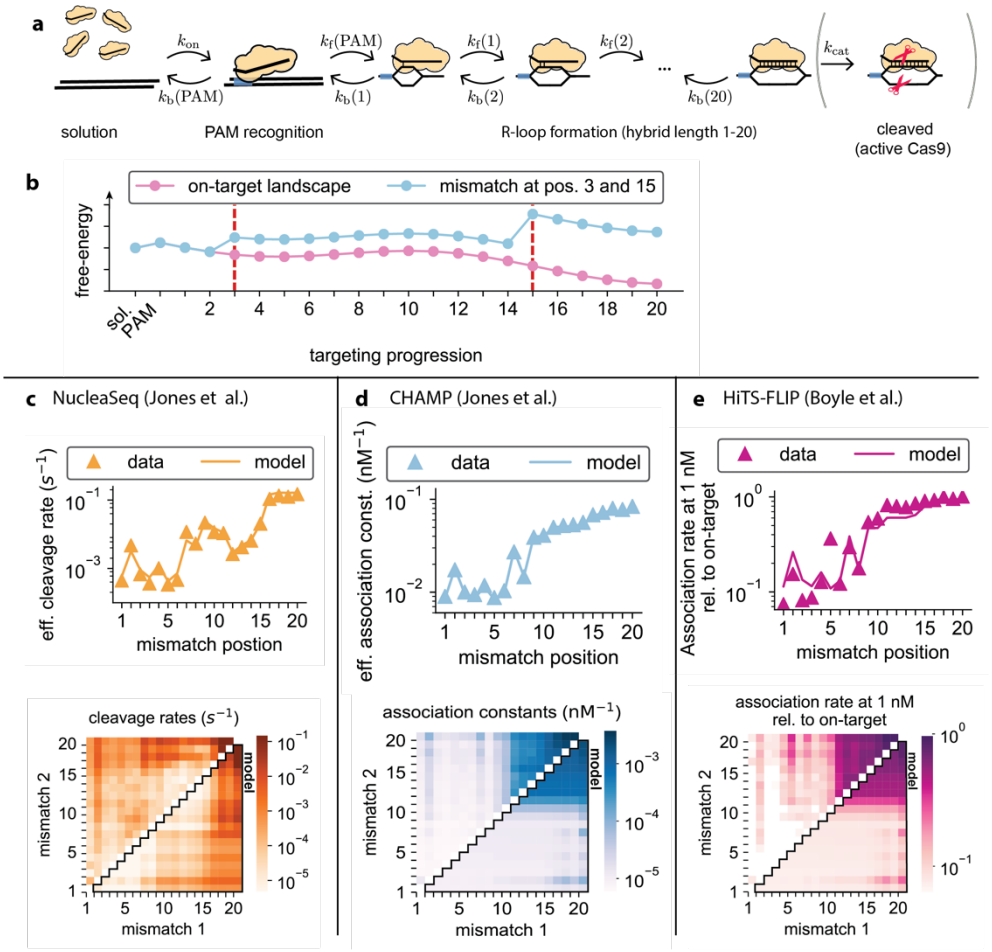

Fig. 1| A kinetic model captures both binding and cleavage data. a, Reaction schema underlying the proposed kinetic model (**Supplementary Information** for details). An available (d)Cas9-sgRNA from solution binds a DNA sequence (either on- or off-target) at its PAM site (blue rectangle) with rate k_{on}. R-loop formation then proceeds in one base-pair increments. A partially formed R-loop containing n base pairs can either extend one base pair at a rate $k_{\mathrm{f}}(n)$ or shrink one base pair at a rate $k_{\mathrm{b}}(n)$. A complete R-loop (20 base pairs) is cleavage competent, and a dsDNA break is catalyzed at a rate k_{cat}. For dCas9, cleavage catalysis is not available, and $k_{\mathrm{cat}} = 0$. **b**, Illustration of a possible free-energy landscape for Cas9-sgRNA-DNA for the on-target (pink) and an off-target with mismatches placed at positions 3 and 15 (blue). Each mismatch raises the entire free-energy landscape starting from the position where it occurs. **c**, Effective cleavage rates and **d**, effective association constants as measured by and simultaneously fitted (Supplementary Information) to the NucleaSeq and CHAMP datasets (*Jones at el.*). Off-targets with one mismatch are show on top and off-targets with two mismatches are shown at the bottom (data above and model below diagonal), both as a function of mismatch position(s). **e**, Model prediction for effective binding rates, compared to HiTS-FLIP data by *Boyle et al.*, for off targets with one mismatch (top) and two mismatches (bottom: data above and model below diagonal), as a function of mismatch position(s).

To train our model, we used independent binding (CHAMP) and cleavage (NucleaSeq) datasets collected using the same guide sequence[15] and performed a global fit based on off-targets with up to two mismatches, averaged over all mismatch types (**Supplementary Information**). Although these two datasets do not correlate well with each other directly (55%, see **Supplementary Fig. 1a**), our model reproduces effective cleavage rates (**Fig. 1c**) and effective association constants (**Fig. 1d**) with a high correlation (86% and 99%, respectively; **Supplementary Fig. 1b**,**c**). As validation, our model accurately captures a third, independent dataset of dCas9 effective association rates[11] (**Fig. 1e**) with a correlation of 97% (**Supplementary Fig. 1d**), and without the use of additional fitting parameters. Our model also predicts the CHAMP data for sequences with more than 2 mismatches, even though these were not included in the training data (**Supplementary Fig. 1e,f**). We conclude that our model accurately captures the physics needed to quantitatively translate between binding of dCas9 and cleavage by Cas9.

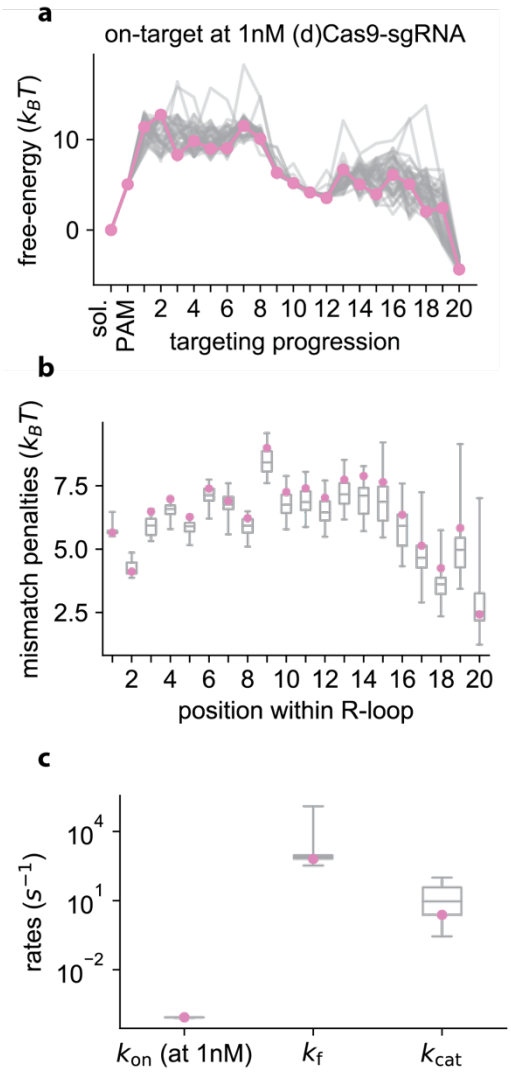

Fig. 2| Microscopic parameter estimated from NucleaSeq and CHAMP datasets. **a**, The free-energy landscape of the on-target reaction along the states shown in **Fig. 1a**. Here sol. is the solution state, PAM is the PAM-bound state and numbers indicate the number of R-loop base pairs formed. **b**, Energetic penalties for mismatches as a function of position. **c**, The three forward rates. In all panels, the global fit with the lowest chi-squared is shown in pink (**Supplementary Information**), and all nearly optimal solutions are represented in grey. For the lower two panels, the interquartile range of nearly optimal solutions are represented in the grey boxes and whiskers denote the complete range of values.

Internal R-loop states are tuned for cleavage specificity without loss of on-target efficiency

To gain mechanistic insights into the targeting reactions, we investigate the estimated free-energy landscape and kinetic parameters (**Fig. 2**), resulting from the simultaneous fit to our training datasets **Fig. 1c,d**).

Starting from the PAM bound state, the on-target free-energy (**Fig. 2a**) increases substantially when forming the first hybrid base pair, and remains relatively high until the 8th base pair is formed. This initial barrier must be bypassed before a stable binding intermediate is reached with about 11 or 12 hybridized base pairs. The free-energy landscape reveals a second barrier to forming a full R-loop (13-18 bp), and eventual cleavage. The penalty for a mismatch (**Fig. 2b**) contains contributions from both DNA-RNA base pairing and protein-nucleotide interactions. Still, the mismatch penalties remain rather constant throughout ($6\pm1\ k_BT$), with notable exceptions being positions 2, 9, 18 and 20.

If the system equilibrates between major barriers in the free-energy landscape, we expect that any change in barrier height can be compensated for by the appropriate change in forward rate (k_f) (**Supplementary Fig 2a,b**)—without effecting model predictions. Consequently, barrier heights and forward rates cannot be simultaneously determined in a partially equilibrated system, and the high variability of predicted barrier heights (**Fig. 2a**) would be explained. By directly showing that the predicted binding and cleavage profiles are indeed insensitive to changing the barrier height (**Supplementary Fig 2c,d**), as long as the forward rate is appropriately adjusted, we confirm the partial equilibration of the system. . This

insight both explains the high variance of free-energy estimates in barrier regions (**Fig. 2a**), and allows us to perform coarse-grain modeling of the system to isolate parameters that are well determined by the data.

Based on the free-energy landscapes in **Fig. 2a**, and this insight, we directly identify equilibrated states as those with free energies that are well-constrained by the fits, and use these as states in our coarse-grained model, with parameter values calculated for every mismatch configuration based on the full model's parameter values (**Supplementary Information**). We define the open (O) R-loop state as the state right after PAM recognition. The local minimum in **Fig. 2a** defines our coarse-grained intermediate (I) R-loop state with between 7 and 13 of its hybrid base pairs formed. Finally, the closed (C) R-loop and cleavage-competent state contains a fully formed hybrid. The resulting coarse-grained reaction scheme (**Fig. 3a**) captures the experimental data as well as the complete model (**Supplementary Fig. 3**).

For the on-target, the coarse-grained parameterization (**Fig. 3b-d**) reveals that the rate-limiting step against cleavage is the transition from the open to the intermediate R-loop state ($k_{\mathrm{OI}} \ll k_{\mathrm{IC}}$). If the intermediate state is entered, the closed state is typically also entered ($k_{\mathrm{IO}} \ll k_{\mathrm{IC}}$). The free-energy difference between the open and intermediate state is low (resulting in $k_{\mathrm{IO}} \approx k_{\mathrm{OI}}$), rendering the transition between the open and the intermediate state reversible. The free-energy difference between the intermediate and closed state is high ($k_{\mathrm{IC}} \gg k_{\mathrm{CI}}$), rendering the transition from an opened to closed configuration essentially irreversible, after which cleavage is nearly guaranteed ($k_{\mathrm{CI}} \ll k_{\mathrm{cat}}$).

For off-targets with a PAM-proximal seed mismatch (nt 1-8), our coarse-grained parameterization (**Fig. 3e-g**) reveals that the rate of transition from an open to intermediate state is greatly suppressed ($k_{\mathrm{IO}}^{\mathrm{on-target}} \gg k_{\mathrm{IO}}^{\mathrm{seed\ m.m.}}$), with the system otherwise behaving as it would on the on-target. In contrast, we find that a PAM-distal mismatch (nt 12-17) limits the effective rate of cleavage from the open state by dramatically reducing the intermediate to closed state transition ($k_{\mathrm{IC}} \ll k_{\mathrm{OI}}$) (**Fig. 3h-j**). The transition from binding to the intermediate state remains unaffected, though returning to the open state competes with completion of the R-loop ($k_{\mathrm{IO}} \approx k_{\mathrm{IC}}$). If the closed state is entered, cleavage is essentially guaranteed ($k_{\mathrm{CI}} \ll k_{\mathrm{cat}}$).

It is interesting to note that specificity in PAM-distal regions stems from having the barrier separating the intermediate and closed states (second barrier in **Fig. 3h** is higher than the first). Moreover, this added specificity seems tuned not to interfere with the crucial on-target cleavage efficiency (second barrier in **Fig. 3b** is lower than the first).

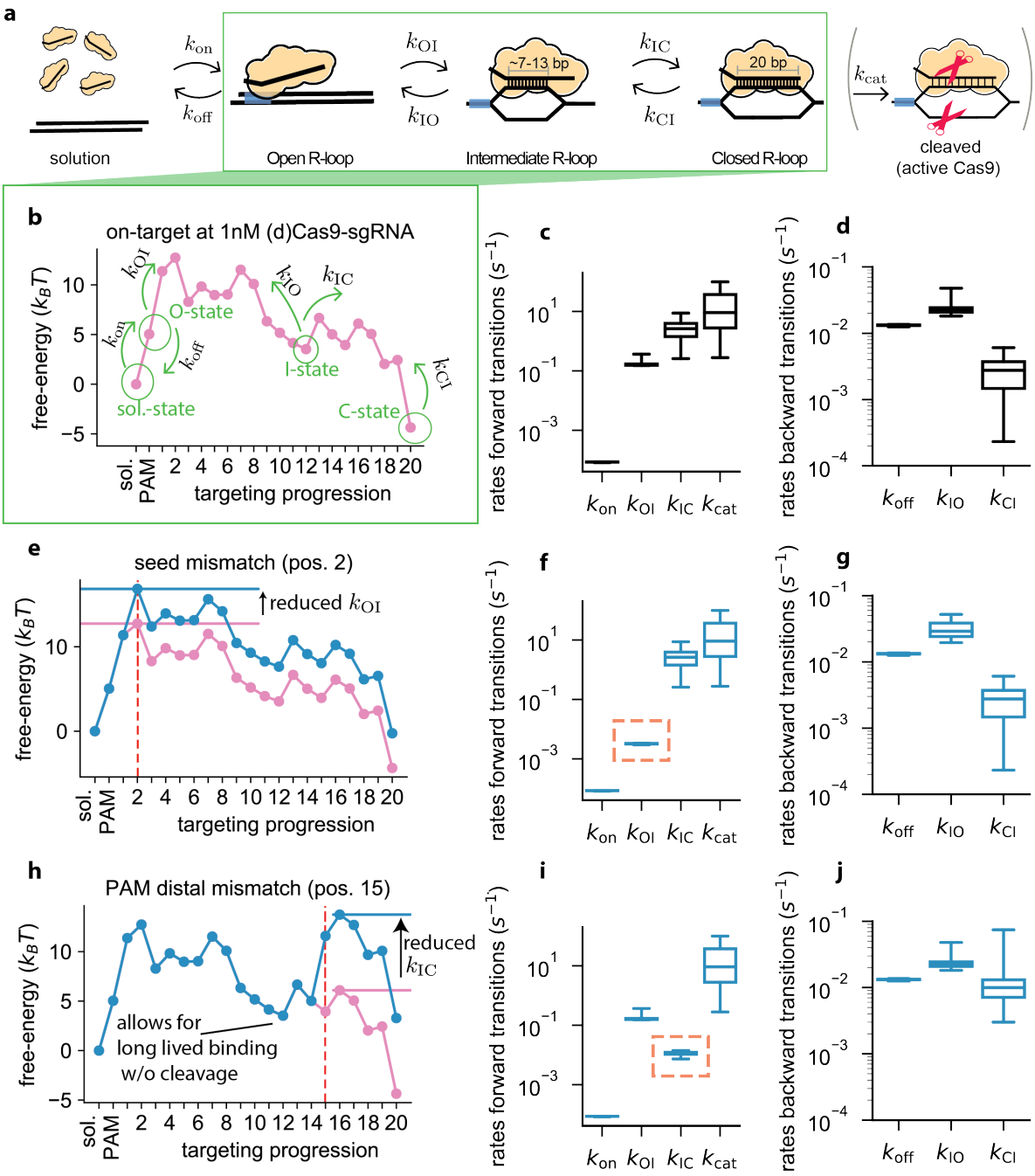

Fig. 3| Coarse-grained kinetic model fully captures bulk data. a, Coarse-grained version of the reaction scheme shown in **Fig. 1a**. Keeping the unbound and post-cleavage state, the targeting reaction is represented by just three states (open, intermediate, and closed R-loops, see **Supplementary Information** for details). **b**, Microscopic free-energy landscape for the on-target exposed to 1nM (d)Cas9-sgRNA (**Fig. 2a**) with coarse-grained states and rates indicated in green. **c**, Coarse-grained forward and **d**, backward rates associated with the landscape in **b**. **e**, Microscopic free-energy landscape for an off-target with a mismatch at position 2 exposed to 1nM (d)Cas9-sgRNA (blue), together with the on-target free-energy landscape (pink). **f, g,** Coarse-grained forward (**f**) and backward (**g**) rates associated with the landscape in **e**. **h-j**, Same as (**f-g**) for an off-target with a mismatch at position 15.

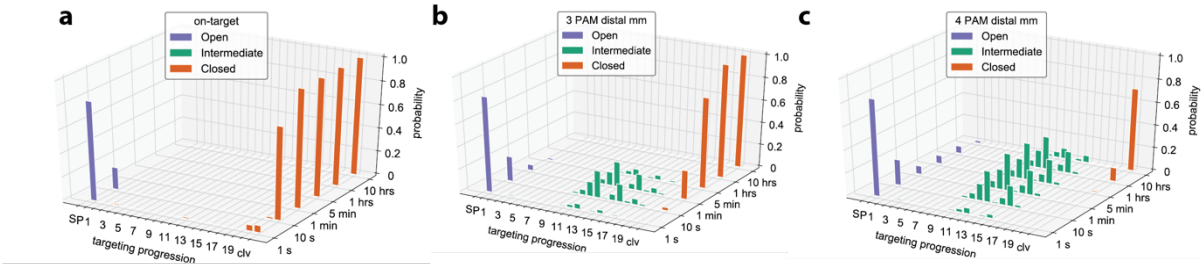

Fig. 4| The time evolution of R-loop hybridization is reminiscent of conformational dynamics. a, The evolution of the occupation probability for any of the 23 microscopic states shown in **Fig. 1a**, as a function of time when interacting with the on-target. **b**, Same as a but when interacting with an off-target with last (PAM distal) 3 base pairs mismatched. **c**, Same as a but when interacting with an off-target with the last 4 base pairs mismatched. Colors indicate the corresponding coarse-grained R-loop configuration as defined in **Fig. 3a**: open R-loop and unbound states (blue), intermediate R-loop states (green) and cleavage-competent and post-cleavage states (orange).

R-loop formation drives Cas9 conformation dynamics

We next questioned what structural properties of Cas9 might couple to hybrid formation, and give rise to the non-monotonic free-energy landscape of **Fig. 2a**. Comparisons between DNA-bound and unbound Cas9 structures have revealed that Cas9 repositions HNH and RuvC nuclease domains to catalyze cleavage[47,54,55]. Bulk FRET experiments also showed that the HNH domain rearranges itself prior to cleavage, and that RuvC activation is tightly coupled to HNH activation[52]. Moreover, single-molecule FRET studies have shown the existence of two dominant DNA-bound configurations with distinct HNH-conformations[29,42,43]. These studies also indicate that the dynamics of the HNH domain are affected by mismatches between sgRNA and DNA. As interrogating the target with the sgRNA is the only way for Cas9 to sense mismatches, we hypothesize that the HNH dynamics is directly coupled to a change in R-loop dynamics.

To test this hypothesis, we mimicked the experiments of Dagdas *et al.*[42] by calculating the time evolution of the occupancy for each microscopic states in the DNA-bound Cas9 landscape for three target sequences (**Fig. 4**). The HNH-domain completes its conformational change within seconds after Cas9 binds to on-target DNA[42]. Our model demonstrates a similar behavior for R-loop progression (**Fig. 4a**). The intermediate R-loop state (green) is visited only transiently, while the closed state (red) strongly resists unwinding of the full hybrid ($k_{\mathrm{IC}} \gg k_{\mathrm{CI}}$) (see **Fig. 3c**,**d** and **4a**). Compared to the on-target, DNA with PAM-distal mismatches reduces the intermediate closure rate (k_{IC}) and increases the time spent in the intermediate state (**Fig. 4b**), in agreement with prior observations[42]. Our model also shows how going from three to four PAM distal mismatches effectively abolishes the occupancy of the closed state at short times[42], as R-loop formation is stalled in the intermediate state (**Fig. 4c**).

In prior FRET experiments, the FRET value corresponding to the intermediate state depended on the number of mismatches introduced, which is evidence that the HNH domain adopts slightly different configurations[43]. The reported relationship between FRET values and mismatches is consistent with tight coupling of conformational change to R-loop progression in PAM distal regions, as our model predicts that going from three to four PAM-distal mismatch increases the probability of residing as a longer intermediate R-loop (**Fig. 4**). The free-energy landscape (**Fig. 2** and **3**) obtained by fitting bulk data (**Fig. 1**) thus complements structural and single-molecule data to describe how Cas9 targets matched and mismatched DNAs.

Kinetic modelling improves genome-wide off-target prediction

Having established that our model quantitatively captures all the essential physics and experimental phenomenology of the targeting process, we sought to compare the model with state-of-the-art off-target prediction tools. Current methods[12,14,20–23,25,44] generally do not directly predict measurable quantities, but rather rank genomic off-targets according to various measures of *in vivo* activity. A common off-target prediction tool uses the Cutting Frequency Determination (CFD) score[12] —a naïve-Bayes classification scheme[24] that assumes mismatches affect the relative cleavage probabilities multiplicatively. More recently, Zhang *et al.* presented a unified CRISPR (uCRISPR) score that outperforms the CFD score[22], in which cleavage probability is evaluated as proportional to the Boltzmann weight corresponding to the cleavage competent state. The assumption of multiplicative errors and the use of Boltzmann weights can both be seen as

implying binding equilibrium, which is a dubious assumption considering that binding and cleavage patterns do not match (see e.g. **Supplementary Fig. 1a**).

To test if the mechanistic nature of our model could improve off-target prediction, we collected data from sequencing-based cleavage experiments. Many experiments have reported on the cleavage activity of *Sp*Cas9-sgRNA along the human genome[27,36–39,41]. Several of these experiments used the same guide RNA, but identified different sets of cleavage sites. To perform a comprehensive evaluation, we gathered data from all experiments with the same guide to create combined test sets. Unsure of the origins of the discrepancies between reports, for each guide we separately tested against the union (sites found in any experiment) and intersection (sites found in every experiment) of the reported off-target sites (**Fig. 5** and **Supplementary Fig. 4-6**). The union of all reported off-targets maximizes the likelihood of covering low probability off-targets, while the intersection minimizes the effect of experiment-dependent biases and noise.

We tested how well our model, the CFD score, and the uCRISPR score could separate reported off-targets over the human genome. For sake of comparison, we need to collapse our dynamic description into a binary classification. We choose to separate out strong off-targets based on the predicted cleavage vs. unbinding probability once the Cas9-sgRNA has bound the PAM[31], as this reflects the steady-state cleavage rate in the low concentration limit. As we are yet to incorporate differences in PAM affinities, we only considered sequences flanked by a canonical NGG motif. **Fig. 5a** shows the resulting precision-recall (PR) curve when tested against all reported off-targets of the EMX1 guide sequence (union). As the threshold for strong off-targets is swept, PR curves display the fraction of sites that are correctly labelled as off-target (precision) against the fraction of the experimentally-identified sequences that are predicted (recall). For therapeutic genome-editing, a high recall is imperative as a false negative prediction is more harmful than a false positive one. Remarkably, our model produces higher recall values for all achievable precisions, clearly outperforming state-of-the-art CFD and uCRISPR classifying schemes for a variety of different guides (**Fig. 5a** and **Supplementary Fig. 4**).

While not trained for it, it is interesting to see that our model also does well on highly mismatched genomic off-target sites, and outperforms the leading off-target predictors also for different guides (**Supplementary Fig. 4-6**), as judged using PR-curves, receiver operating characteristic curves, and the F1-score (**Fig. 5a** and **Supplementary Fig. 4-5,** and **Fig. 5b** and **Supplementary Fig. 6,** respectively).

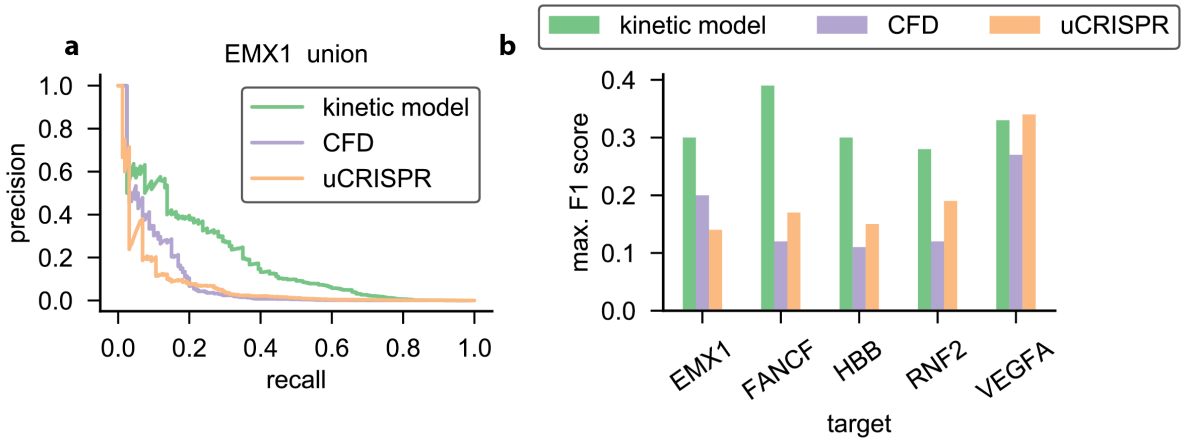

Fig. 5| Genome-wide off-target classification a, Precision recall-curves for our model (green) and predictions based on the CFD score (purple) or uCRISPR score (orange) for the EMX1 site using all experimentally identified off-targets. **b,** F1-scores for our model (green), CFD prediction tool (purple) and uCRISPR (orange), for target sites EMX1, FANCF, HBB, RNF2 and VEGFA site 1 using all experimentally identified off-targets. For each condition, the maximum obtainable F1-score along the corresponding PR-curve is displayed (see **a** and **Supplementary Fig. 4**).

Discussion

The increasing popularity of CRISPR-Cas9 for genome-editing calls for a quantitative understanding of its function and the risks involved with its application. We have built a kinetic model and trained it on high-throughput cleavage and binding propensity measurements[15]. Our bottom-up modelling approach has allowed us to decipher the microscopic free-energy landscape underlying *Sp*Cas9 target recognition (**Fig. 1-2**). Based on extracted free-energy landscapes, we found that *Sp*Cas9's kinetics are effectively dominated by transitions between the open, intermediate R-loop, and closed states (**Fig. 3**). As mismatches affect the three R-loop states similarly to the three configurational states of Cas9's nuclease domains[42,43], we propose that PAM distal R-loop formation directly couples to protein conformation (**Fig. 4**)—pointing toward the relevant structure-function relation for the most important RNA-guided nuclease in use today.

By mechanistically accounting for the kinetic nature of the targeting process, our model outperforms existing genome-wide off-target prediction tools. For simplicity and robustness, we built our model to exclude mismatch type parameters (*e.g.* G-G vs. G-A), allowing for extensive training using datasets based on a single guide sequence and off-targets containing up to two mismatches. This training does not limit the model's application as the model also improves on the detection of highly-mismatched genomic off-target sites (**Fig. 5** and **Supplementary Fig. 4-6**).

Beyond off-target prediction, our model allows us to determine the time dependence of off-target binding and cleavage, which can facilitate the design of off-target libraries in Cas9- or dCas9-based experiments. For example, a recent study by Jost *et al.*[5] demonstrated that a series of mismatched guides can be used to titrate gene expression during CRISPRa/CRISPRi. Knowing *Sp*Cas9's microscopic free-energy landscape (**Fig. 2-3**) can simplify the design of CRISPRa/CRISPRi libraries for novel gene targets. Given the possibility of inactivating wildtype Cas9 with PAM-distal mismatches in the guide[56], our model can also guide titration of the binding affinity of thus inactivated Cas9-sgRNA.

The physical insights generated by the free-energy landscapes we extract could also help rational protein-reengineering efforts aimed at producing high-fidelity Cas9 variants that maintain high on-target efficiency[27–29]. For *Sp*Cas9, we find that the barrier between the intermediate and closed states is tuned to extend the cleavage specificity beyond the seed, without affecting on-target efficiency (**Fig. 3b,h**).

Taken together, we have shown that mechanistic modelling combined with high-throughput data sets give biophysical insights into *Sp*Cas9 off-targeting, and that those insights give predictive power far beyond the training sets. *Sp*Cas9 is only one of many RNA-guided nucleases with biotechnological applications, and other CRISPR associated nucleases (such as Cas12a, Cas13 and Cas14) offer a diversified genome-engineering toolkit[15,57–62]. These nucleases can all be characterized with our approach, and it will be especially interesting to compare the free-energy landscape of our *Sp*Cas9 benchmark to that of engineered[27–29] and natural (e.g. *N. meningitides* Cas9[63]) high-fidelity Cas9 variants.

Acknowledgements

We would like to thank Kristian Blom, Diewertje Dekker, and Sonny de Jong for valuable discussions and their help during the project. Thank you also to the members of the Chirlmin Joo lab and Stan Brouns lab for valuable discussions. We also thank Evan Boyle for sharing his data and answering all our questions. B.E.M. forms part of the research program ''Crowd management: the physics of genome processing in complex environments", supported by NWO. M.K. was supported by the Netherlands Organization for Scientific Research (NWO/OCW), as part of the Frontiers in Nanoscience program. M.D. acknowledges support from the Parents in KIND program, sponsored by The Kavli Institute of Nanoscience Delft, the Department of Bionanoscience at TU Delft, and through a Spinoza Prize awarded to M. Dogterom by NWO. I.J.F. is supported by a University of Texas College of Natural Sciences Catalyst award and the Welch Foundation (F-1808). I.J.F. and S.K.J. are supported by the U.S. National Institute of Health (R01GM124141, F32AG053051).

Author contributions

B.E.M and M.K: designed and performed the research. Wrote the manuscript
K.v.d.S: and C.v.d.S performed the research.
S.K.J: provided data. Wrote manuscript
J.H: provided data. Wrote manuscript
I.J.F: provided data. Wrote manuscript
M.D: designed the research and wrote manuscript

References

1. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32**, 347–350 (2014).

2. Wang, H., La Russa, M. & Qi, L. S. CRISPR/Cas9 in Genome Editing and Beyond. *Annu. Rev. Biochem.* **85**, 227–264 (2016).

3. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).

4. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442 (2013).

5. Jost, M. *et al.* Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nat. Biotechnol.* (2020). doi:10.1038/s41587-019-0387-5

6. Niu, D. *et al.* Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science (80-.).* **1307**, eaan4187 (2017).

7. Hammond, A. *et al.* A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae. *Nat. Biotechnol.* **34**, 78–83 (2016).

8. Amoasii, L. *et al.* Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy. *Science (80-.).* **362**, 1–6 (2018).

9. Park, C. Y. *et al.* Functional Correction of Large Factor VIII Gene Chromosomal Inversions in Hemophilia A Patient-Derived iPSCs Using CRISPR-Cas9. *Cell Stem Cell* **17**, 213–220 (2015).

10. Jinek, M. *et al.* A Programmable Dual-RNA – Guided. *Science (80-.).* **337**, 816–822 (2012).

11. Boyle, E. A. *et al.* High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci.* **114**, 5461–5466 (2017).

12. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

13. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).

14. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

15. Jones Jr, S. K. *et al.* Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *BioRxiv* 1–17 (2019).

16. Kim, D., Luk, K., Wolfe, S. A. & Kim, J.-S. Evaluating and Enhancing Target Specificity of Gene-Editing Nucleases and Deaminases. *Annu. Rev. Biochem.* 1–30 (2019). doi:10.1146/annurev-biochem-013118-111730

17. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).

18. Tsai, S. Q. & Joung, J. K. Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases. *Nat. Rev. Genet.* **17**, 300–312 (2016).

19. Cullot, G. *et al.* CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nat. Commun.* **10**, 1–14 (2019).

20. Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.* **44**, W272–W276 (2016).

21. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: Fast CRISPR target site identification. *Nat. Methods* **11**, 122–123 (2014).

22. Zhang, D., Hurst, T., Duan, D. & Chen, S.-J. Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. *Proc. Natl. Acad. Sci.* **116**, 8693–8698 (2019).

23. Stemmer, M., Thumberger, T., Del Sol

Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS One* **10**, 1–11 (2015).

24. Listgarten, J. *et al.* Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat. Biomed. Eng.* **2**, 38–47 (2018).

25. Chuai, G. *et al.* DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**, 1–18 (2018).

26. Tycko, J., Myer, V. E. & Hsu, P. D. Methods for Optimizing CRISPR-Cas9 Genome Editing Specificity. *Mol. Cell* **63**, 355–370 (2016).

27. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science (80-.).* **351**, 84–88 (2016).

28. Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

29. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

30. Farasat, I. & Salis, H. M. A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLoS Comput. Biol.* **12**, 1–33 (2016).

31. Klein, M., Eslami-Mossallam, B., Arroyo, D. G. & Depken, M. Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Rep.* **22**, (2018).

32. Bisaria, N., Jarmoskaite, I. & Herschlag, D. Lessons from Enzyme Kinetics Reveal Specificity Principles for RNA-Guided Nucleases in RNA Interference and CRISPR-Based Genome Editing. *Cell Syst.* **4**, 21–29 (2017).

33. O'Geen, H., Henry, I. M., Bhakta, M. S., Meckler, J. F. & Segal, D. J. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Res.* **43**, 3389–3404 (2015).

34. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).

35. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**, 677–683 (2014).

36. Cameron, P. *et al.* Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat. Methods* **14**, 600–606 (2017).

37. Tsai, S. Q. *et al.* CIRCLE-seq: A highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).

38. Kim, D. *et al.* Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243 (2015).

39. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–198 (2015).

40. Frock, R. L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33**, 179–188 (2015).

41. Yan, W. X. *et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.* **8**, 1–9 (2017).

42. Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Sci. Adv.* **3**, 1–9 (2017).

43. Yang, M. *et al.* The Conformational Dynamics of Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET. *Cell Rep.* **22**, 372–382 (2018).

44. Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 1–12 (2016).

45. Jung, C. *et al.* Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. *Cell* **170**, 35-47.e13 (2017).

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

47. Jiang, F., Zhou, K., Gressel, S. & Doudna, J. A. A cas9 guide RNA complex preorganized for target DNA recognition. *Science (80-.).* **348**, 1477–1482 (2015).

48. Josephs, E. A. *et al.* Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic Acids Res.* **43**, 8924–8941 (2015).

49. Rutkauskas, M. *et al.* Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Rep.* **10**, 1534–1543 (2015).

50. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci.* **111**, 9798–9803 (2014).

51. Xiao, Y. *et al.* Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. *Cell* **170**, 48-60.e11 (2017).

52. Sternberg, S. H., Lafrance, B., Kaplan, M. & Doudna, J. A. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110–113 (2015).

53. Sung, K., Park, J., Kim, Y., Lee, N. K. & Kim, S. K. Target Specificity of Cas9 Nuclease via DNA Rearrangement Regulated by the REC2 Domain. *J. Am. Chem. Soc.* **140**, 7778–7781 (2018).

54. Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science (80-.).* **351**, 867–871 (2016).

55. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science (80-.).* **343**, (2014).

56. Dahlman, J. E. *et al.* Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. *Nat. Biotechnol.* **33**, 1159–1161 (2015).

57. Chen, J. S. *et al.* CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science (80-.).* **360**, 436–439 (2018).

58. Gootenberg, J. S. *et al.* Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438–442 (2017).

59. Gootenberg, J. S. *et al.* Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science (80-.).* **444**, 439–444 (2018).

60. Harrington, L. B. *et al.* Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science (80-.).* **362**, 839–842 (2018).

61. Kim, D. *et al.* Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34**, 863–868 (2016).

62. Kleinstiver, B. P. *et al.* Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **34**, 869–874 (2016).

63. Amrani, N. *et al.* NmeCas9 is an intrinsically high-fidelity genome-editing platform Jin-Soo Kim. *Genome Biol.* **19**, 1–25 (2018).