Sequence specificity in DNA binding is determined by association rather than dissociation

Authors:

Emil Marklund[1], Guanzhong Mao[1], Sebastian Deindl[1*], Johan Elf[1*]

Affiliations:

[1]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Box 596, 75124, Uppsala, Sweden

*Correspondence to: sebastian.deindl@icm.uu.se, johan.elf@icm.uu.se

Abstract:

Sequence-specific binding of proteins to DNA is essential for accessing genetic information. Here, we derive a simple equation for target-site recognition, which uncovers a previously unrecognized coupling between the macroscopic association and dissociation rates of the searching protein. Importantly, this relationship makes it possible to recover the relevant microscopic rates from experimentally determined macroscopic ones. We directly test the equation by observing the binding and unbinding of individual *lac* repressor (LacI) molecules during target search. We find that LacI dissociates from different target sequences with essentially identical microscopic dissociation rates. Instead, sequence specificity is determined by the efficiency with which the protein recognizes different targets, effectively reducing its risk of being retained on a non-target sequence. Our theoretical framework also accounts for the coupling between off-target binding

and unbinding of the catalytically inactive Cas9 (dCas9), showing that the binding pathway can be obtained from macroscopic data.

One Sentence Summary:

Association and dissociation rates are anti-correlated for reactions that include a nonspecific probing step.

Main Text:

Sequence-specific recognition and binding of DNA target sites by proteins such as polymerases, DNA-modifying enzymes, and transcription factors are essential for gene expression and regulation across all kingdoms of life (*1*). The textbook explanation for this sequence dependence of binding posits that favorable hydrogen bonding interactions between the protein and particular DNA sequences result in prolonged binding times (*2*). Consequently, the rate of protein dissociation would depend on the DNA sequence, while the association rate would be invariant with respect to sequence. Indeed, the rate of protein association with DNA has often been assumed to be sequence-independent (*3*–*6*). However, single-molecule measurements have shown that when a protein scans the DNA for binding sites, the association rate does depend on the sequence *(7)*, and that different target sequences can be bypassed with distinct probabilities (*8*). These differences have been ascribed to differences in the probability of recognition when the protein is centered on the target sequence (*7*). It is unknown whether the rate of binding imposes constraints on the rate of dissociation, beyond the fact that the ratio of association and dissociation rates is

necessarily dictated by the free energy difference between the free and bound states. So far, the recovery of microscopic parameters from experimentally obtained macroscopic ones has remained an unsolved problem, limiting our understanding of how sequence-specific binding is achieved on the microscopic level.

To explore the limits of the association and dissociation rates, we considered the standard model (*9*), according to which a protein has a non-specific testing mode where it is bound nonspecifically to DNA (Fig. 1A). In the testing state, the protein can either specifically bind the target with probability p_{tot}, or dissociate into solution with probability 1-p_{tot}. When the association process is modeled as a three-state (specifically bound, nonspecifically bound, and dissociated) continuous time Markov chain, the effective macroscopic target association and dissociation rates (k_a and k_d) relate to each other according to (see Supplementary Text for derivation)

$$k_a = k_{on,max} - \frac{k_{on,max}}{k_{off,\mu}} k_d, \qquad (1)$$

where $k_{on,max}$ is the association rate given by a searching protein that binds the target upon every non-specific encounter (p_{tot}=1), and $k_{off,\mu}$ is the rate of microscopical dissociation from the bound state into the nonspecifically bound searching mode. This equation implies that the association and dissociation rates are inherently coupled, and linearly anti-correlated if binding sites exhibit identical microscopical dissociation rates, since $k_{on,max}$ does not depend on the specific sequence. The linear relationship between k_a and k_d described by Eq. 1 is implicitly parameterized by the probability of binding rather than dissociating from the non-specifically bound state, p_{tot}, such that an increase in p_{tot} causes an increase in k_a, and a corresponding decrease in k_d (Fig. 1B). This anti-

correlation can be intuitively understood by acknowledging that a decrease in the number of target site encounters required for successful binding must, in turn, result in a corresponding increase in the number of dissociation attempts needed for macroscopic dissociation from the target (Fig. 1C). Most importantly, Eq. 1 makes it possible to obtain a microscopic parameter, such as $k_{\mathrm{off},\mu}$, from macroscopically measurable parameters, such as k_{a} and k_{d}.

To experimentally test the anti-correlation between association and dissociation rates, we measured the kinetics with which a prototypical DNA-binding protein, the transcription factor LacI, binds to its natural operator sites using single-molecule fluorescence colocalization. We surface-immobilized a Cy5-labeled DNA construct containing a natural *lacO* operator site (O_1, O_2, or O_3) and used total-internal-reflection fluorescence microscopy to monitor individual DNA molecules (Fig. 1D). Upon addition of LacI labeled with Cy3 distal from the DNA binding domain, we monitored the appearance and disappearance of well-defined spots with co-localized fluorescence emission from both Cy3 and Cy5 (Fig. 1E). The Cy3 label has previously been shown to affect neither the specific nor the nonspecific DNA binding (*8*) (Labeling efficiency: 84.5 %; see also Supplementary Text and Table S1). Few DNA molecules featured co-localized LacI-Cy3 spots in control experiments with Cy5-labeled DNA constructs lacking an operator site (11% and 3% at 1 and 100 mM NaCl, in contrast to >65%, >60% and >20% at 1, 100 and 200 mM NaCl for DNA with an O_1 operator; Fig. S1), indicating that the Cy3 spots represent complexes of LacI-Cy3 specifically bound to the operator with only a minor contribution from nonspecific binding of LacI-Cy3 to DNA or to the surface.

To implement conditions that give rise to a range of different association and dissociation rates, we varied the salt concentration in our experiments (Fig. 1F) since changes in salt concentration

are expected to affect the time that LacI spends nonspecifically bound to DNA while sliding along it (*9*, *10*). This in turn would change the number of operator encounters per nonspecific association, such that p_{tot} is expected to increase with decreasing salt concentration. We note that the k_a values measured for each salt titration should be interpreted as being merely proportional to the true bimolecular association rate constants since the exact concentration of active LacI needed for normalization can vary between salt titration repeats due to differences in the extent of protein surface adsorption, protein stability, and pipetting errors. Nevertheless, we obtain a reproducible and anti-correlated relationship between the measured k_a and k_d values for each salt titration and operator, consistent with the notion that p_{tot} varies, while $k_{off,\mu}$ remains constant for each operator when changing the salt concentration (Fig. 1G; see also Fig. S2-4).

To test if the specificity of LacI-binding to different operators is due to differences in microscopic association or dissociation, we fit Eq. 1 to the experimentally determined k_a and k_d values (colored lines in Fig. 1G), yielding $k_{off,\mu}$ for each operator as the k_d-intercept of each k_a *versus* k_d line (Fig. 1H). Surprisingly, the estimates obtained for $k_{off,\mu}$ are very similar for all operators.

Analogously, we can also estimate $k_{off,\mu}$ for the different operators (Fig. 2A) from existing *in vivo* estimates of k_a and k_d ((*7*, *11–13*), Fig. 2B). $k_{on,max}$ has been measured independently *in vivo* (*7*), and $k_{off,\mu}$ can therefore be calculated as the only unknown in Eq. 1. Consistent with what we found in our *in vitro* experiments, the $k_{off,\mu}$ estimates obtained from *in vivo* data are very similar for all operators (Fig. 2C). Even though the K_D value of O_2 exceeds that of O_1 more than 4-fold, and that of O_{sym} 20-fold, these operators exhibit essentially the same $k_{off,\mu}$ *in vivo*, as they all fall on the same k_a *versus* k_d line in Fig. 2B. The differences in K_D observed for the different operators can thus be explained predominantly by differences in target-site recognition (p_{tot}) and the numerous

microscopic re-associations that the protein undergoes before every successful macroscopic dissociation from a strong operator.

The agreement between the experimental data and our simple model suggests that LacI binding dynamics can be captured with one kinetic barrier, the height of which differs for different operators; i.e., it is more favorable for LacI to bind to certain operators than to others when sliding by, but the rate of escaping from the specifically bound state does not depend on the sequence (Fig. 2D). By recognizing that mutations along the binding pathway can be seen as energetic barriers for binding, our theoretical framework can also be used to dissect the binding path in more complex, sequential binding mechanisms. Accordingly, one would first mutate a binding sequence in several different ways, measure the resulting macroscopic rates k_a and k_d, and then determine which sector of the (k_a,k_d)-space the different mutations fall into (Fig 3A). Assuming that the native sequence has the highest k_a value and that mutations introduce a rate-limiting step, the sectors will be ordered according to the position of the mutations along the reaction pathway. Thus, rate-limiting steps closer to the bound state will result in fewer rebinding events, leading to an increase in k_d for the same value of k_a.

To demonstrate the use of this method, we apply it to high-throughput association and dissociation data available for dCas9 binding to off-target, mismatch mutants ((*14*), Fig. 3). dCas9 is guided by an RNA (gRNA) when it binds DNA, with a reaction coordinate for the testing of recognition that is already well-established (*15–17*). The gRNA binds the DNA by base pairing in a sequential manner starting from a seed sequence, and then continuing hybridization at base pairs more distal from the seed. When we plot linear fits to the single-base mismatch data, the resulting slopes and k_d-intercepts indicate a binding pathway that is well-aligned with the order of basepairs in the guide

RNA sequence, starting at the seed and then moving further into the sequence (Fig. 3A). When we group the sequence into 3-nucleotide groups and plot the lines corresponding to single- and double-base mismatches within each group, the k_d-intercepts show a binding pathway that is identical to the order of the base pairs in the guide RNA sequence (Fig 3B). Based on these groups, we have colored all the single and double mismatch mutations in the k_a *versus* k_d plots of Fig. 3. The resulting rainbow-colored pattern, the order of which corresponds to the order in the guide RNA sequence, demonstrates how the k_a *versus* k_d plot maps the mutations onto the reaction coordinate (Fig 3C).

In conclusion, the efficiency of target-site recognition not only is crucial for determining protein-DNA association rates but also plays an equally important role in determining how long proteins remain bound to their targets. In the case of the *lac* repressor, we have shown that the efficiency of target-site recognition (p_{tot}) - and *not* how long the protein remains in the bound state - causes the differences in binding strength observed for different sequences. This behavior may represent an evolutionary adaptation to facilitate fast search by minimizing the risk of the protein being retained on sequences that resemble the actual operators. The coupling between association and dissociation rates holds for all bimolecular association-dissociation processes adhering to detailed balance, where a step of rapid testing for molecular recognition precedes the strong binding of a target. Our theoretical result is therefore very likely to be generally applicable to a wide range of kinetic systems in addition to the ones investigated here, including processes that do not involve protein-DNA interactions.

References and Notes:

1. W. Gilbert, B. Müller-Hill, The lac operator is DNA. *Proc. Natl. Acad. Sci. U. S. A.* **58**, 2415–2421 (1967).

2. R. Milo, R. Phillips, *Cell Biology by the Numbers* (Garland Science, 2015).

3. A. Grönlund, P. Lötstedt, J. Elf, Transcription factor binding kinetics constrain noise suppression via negative feedback. *Nat. Commun.* **4**, 1864 (2013).

4. D. L. Jones, R. C. Brewster, R. Phillips, Promoter architecture dictates cell-to-cell variability in gene expression. *Science*. **346**, 1533–1536 (2014).

5. M. Z. Ali, V. Parisutham, S. Choubey, R. C. Brewster, Inherent regulatory asymmetry emanating from network architecture in a prevalent autoregulatory motif. *eLife*. **9** (2020), , doi:10.7554/elife.56517.

6. M. Morrison, M. Razo-Mejia, R. Phillips, Reconciling kinetic and thermodynamic models of bacterial transcription. *PLoS Comput. Biol.* **17**, e1008572 (2021).

7. P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, J. Elf, The lac Repressor Displays Facilitated Diffusion in Living Cells. *Science*. **336** (2012), pp. 1595–1598.

8. E. Marklund, B. van Oosten, G. Mao, E. Amselem, K. Kipper, A. Sabantsev, A. Emmerich, D. Globisch, X. Zheng, L. C. Lehmann, O. G. Berg, M. Johansson, J. Elf, S. Deindl, DNA surface exploration and operator bypassing during target search. *Nature*. **583** (2020), pp.

858–861.

9. O. G. Berg, R. B. Winter, P. H. Von Hippel, Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*. **20** (1981), pp. 6929–6948.

10. P. C. Blainey, A. M. van Oijen, A. Banerjee, G. L. Verdine, X. S. Xie, A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5752–5757 (2006).

11. H. G. Garcia, R. Phillips, Quantitative dissection of the simple repression input–output function. *of the National Academy of Sciences* (2011) (available at https://www.pnas.org/content/108/29/12173.short).

12. R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, R. Phillips, The transcription factor titration effect dictates level of gene expression. *Cell*. **156**, 1312–1323 (2014).

13. P. Hammar, Walldé n M, Fange D, Persson F, Baltekin O, Ullman G, Leroy P, Elf J. 2014. Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. *Nat. Genet.* **46**, 405–408.

14. E. A. Boyle, J. O. L. Andreasson, L. M. Chircus, S. H. Sternberg, M. J. Wu, C. K. Guegler, J. A. Doudna, W. J. Greenleaf, High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5461–5466 (2017).

15. S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, J. A. Doudna, DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*. **507**, 62–67 (2014).

16. X. Wu, D. A. Scott, A. J. Kriz, A. C. Chiu, P. D. Hsu, D. B. Dadon, A. W. Cheng, A. E. Trevino, S. Konermann, S. Chen, R. Jaenisch, F. Zhang, P. A. Sharp, Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).

17. M. D. Szczelkun, M. S. Tikhomirova, T. Sinkunas, G. Gasiunas, T. Karvelis, P. Pschera, V. Siksnys, R. Seidel, Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9798–9803 (2014).

18. K. Kipper, N. Eremina, E. Marklund, S. Tubasum, G. Mao, L. C. Lehmann, J. Elf, S. Deindl, Structure-guided approach to site-specific fluorophore labeling of the lac repressor LacI. *PLoS One*. **13**, e0198416 (2018).

19. S. Deindl, X. Zhuang, Monitoring conformational dynamics with single-molecule fluorescence energy transfer: applications in nucleosome remodeling. *Methods Enzymol.* **513**, 59–86 (2012).

20. A. Sabantsev, R. F. Levendosky, X. Zhuang, G. D. Bowman, S. Deindl, Direct observation of coordinated DNA movements on the nucleosome during chromatin remodelling. *Nat. Commun.* **10**, 1720 (2019).

21. J.-C. Olivo-Marin, Extraction of spots in biological images using multiscale products. *Pattern Recognit.* **35**, 1989–1996 (2002).

22. B. M. Sadler, A. Swami, Analysis of multiscale products for step detection and estimation. *IEEE Trans. Inf. Theory*. **45**, 1043–1051 (1999).

23. D. Garcia, Robust smoothing of gridded data in one and higher dimensions with missing values. *Comput. Stat. Data Anal.* **54**, 1167–1178 (2010).

24. M. Lindén, V. Ćurić, A. Boucharin, D. Fange, J. Elf, Simulated single molecule microscopy with SMeagol. *Bioinformatics*. **32**, 2394–2395 (2016).

25. M. Klein, B. Eslami-Mossallam, D. G. Arroyo, M. Depken, Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Rep.* **22**, 1413–1423 (2018).

26. R. B. Winter, P. H. von Hippel, Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli repressor--operator interaction: equilibrium measurements. *Biochemistry*. **20**, 6948–6960 (1981).

27. J. Elf, G.-W. Li, X. S. Xie, Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*. **316**, 1191–1194 (2007).

Acknowledgments: We thank Otto Berg, Måns Ehrenberg, Jakub Wiktor, David Fange, Irmeli Barkefors, and Daniel Jones for discussions.

Funding: KAW **(**2016.0077 & 2019.0439 to JE; 019.0306 to SD), VR (2016-06213 to JE, 2020-06459 to EM), ERC (StG, 714068 to SD; AdG, 885360 to JE); **Author contributions:** JE and EM conceived the study; EM derived Eq. 1; SD, EM, and MG designed the experiments; MG performed the experiments; EM analyzed the data; EM, SD and JE interpreted results; EM, JE, and SD wrote the paper; **Competing interests:** Authors declare no competing interests; **Data and materials availability:** All raw data and analysis codes will be made available upon request.

List of Supplementary Materials:

References 18-27 are only cited in the Supplementary Materials.

Materials and Methods

Supplementary Text

Figures S1 to S5

Tables S1 to S2

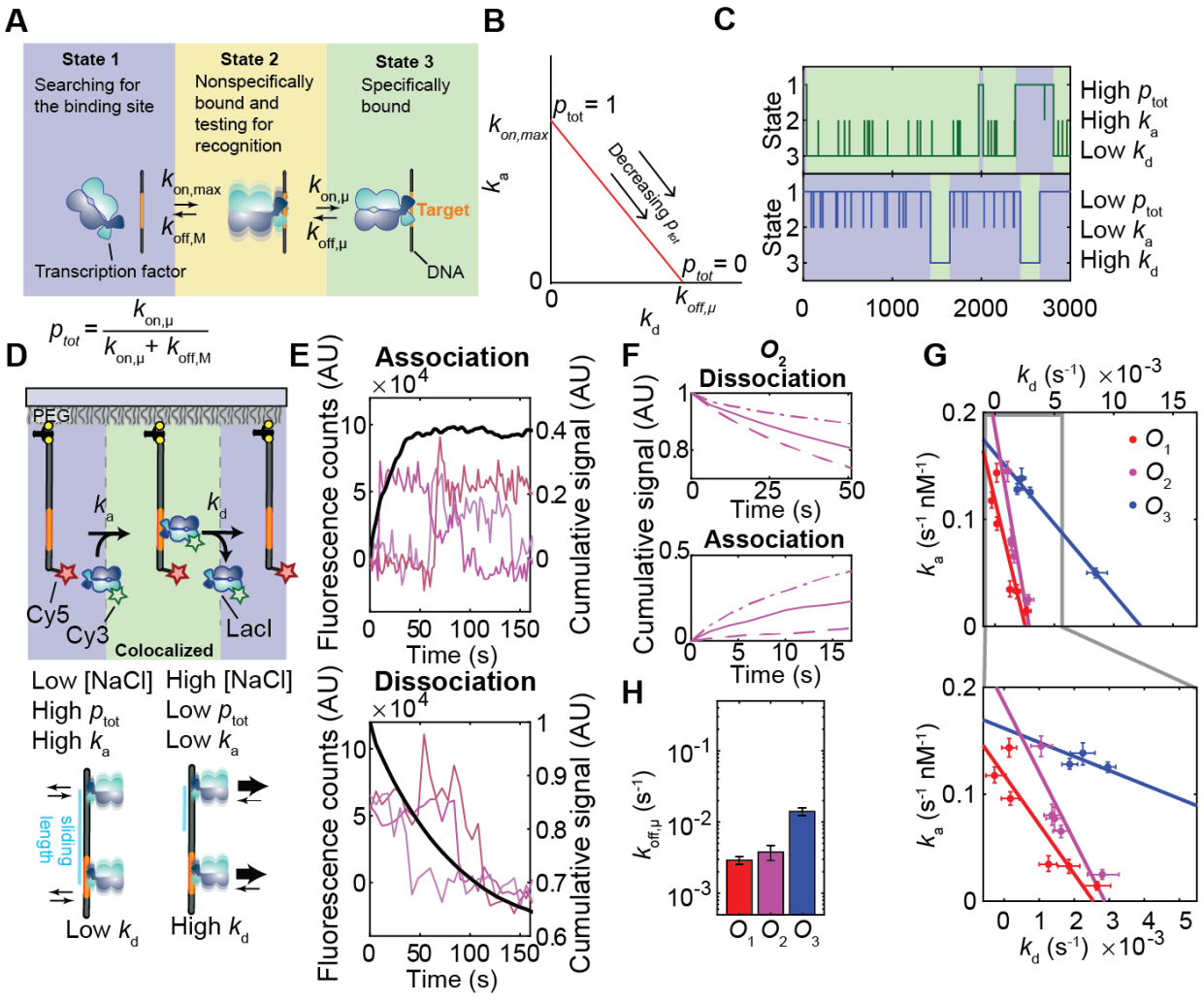

Fig. 1. Bimolecular association and dissociation rates are inherently anti-correlated due to target-site probing. (**A**) Schematic of the kinetic model describing protein-DNA binding. (**B**) The effective rate constants for the association to (k_a) and dissociation from (k_d) the target site are coupled according to Eq. 1. This relationship becomes anti-correlated and linear when $k_{off,\mu}$ is constant and p_{tot} changes (red line). (**C**) Example traces from stochastic simulations sampling the association, dissociation, and nonspecific binding with target-site probing. When p_{tot} is high (top), the search times become short ($1/k_a$, blue areas) and the binding time long ($1/k_d$, green areas). When p_{tot} is low (bottom), the search times become long and the binding times short. (**D**)

Single-molecule colocalization measurements detect association and dissociation for LacI binding to its operators (left) and the predicted effect on association and dissociation rates of changing the salt concentration (right). (**E**) Example single-molecule traces showing binding to and unbinding from the O_2 operator at 100 mM NaCl (colored lines) and the normalized association and dissociation curves (black lines) obtained after summing 667 and 773 traces for the association and dissociation experiment, respectively. a.u., arbitrary units. (**F**) Normalized association and dissociation curves for O_2 and 1 mM (dashed dotted), 100 mM (solid), and 200 mM NaCl (dashed). (**G**) Measured k_a and k_d values for the three *lac* operators, and fits to Eq. 1 for each repeat (colored lines). The salt concentrations used for the different experiments are in the range 1-250 mM supplemented NaCl for O_1 and O_2, and 1-100 mM supplemented NaCl for O_3 (Fig. S1). Error bars are 68% confidence intervals obtained after bootstrapping the fluorescence traces in each experiment. (**H)** Microscopic dissociation rates $k_{off,\mu}$ for the different operators, estimated as the k_d-intercepts of the fits to Eq. 1. Error bars are standard errors from n = 2 salt titrations for each operator.

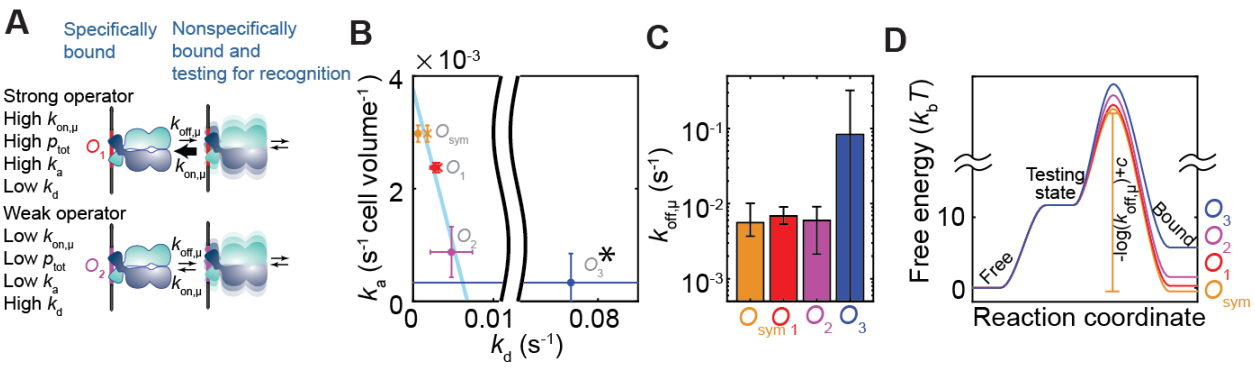

Fig. 2. Different DNA targets exhibit similar microscopic dissociation rates but different recognition probabilities. (**A**) Predicted effect on the association and dissociation rates if $k_{on,\mu}$ were different but $k_{off,\mu}$ identical for the different operators. (**B**) Experimental single-molecule target-site association rates k_a (*7*) plotted against the dissociation rates k_d for the different *lac* operators. For the crosses, k_d was directly measured by single-molecule imaging (*13*). For the dots, k_d was calculated as $K_D \times k_a$, where the equilibrium constant K_D was measured via the repression ratio of gene expression (*11*, *12*). Cyan line, best fit of Eq. 1 to the O_{sym}, O_1 and O_2 data. * Due to the large error in the k_d estimate for O_3 (68% CI: [-0.04,0.20] s^{-1}), it has been excluded from the fit. Error bars are standard errors, obtained by propagating the errors from the experiments. (**C**) Microscopic dissociation rates $k_{off,\mu}$ for the different operators, estimated from the *in vivo* data using Eq. 1. Error bars are 68 % confidence intervals, obtained by propagating the errors from the experiments. The confidence interval for is O_3 is [-0.05,0.32] s^{-1}, making this $k_{off,\mu}$ estimate an upper bound. (**D**) Energy landscapes (a putative rather than true reaction coordinate is shown) for the transition from free (State 1) to bound (State 3) states for the different operators, as determined by the measured K_D and $k_{off,\mu}$ values. The activation energy on the transition path between the testing state and bound state is not uniquely determined, but the

differences in activation energies between the different operators are. The activation energy is equal to -log($k_{\mathrm{off},\mu}$) + c, where c is the same constant for all operators (see Supplementary Text).

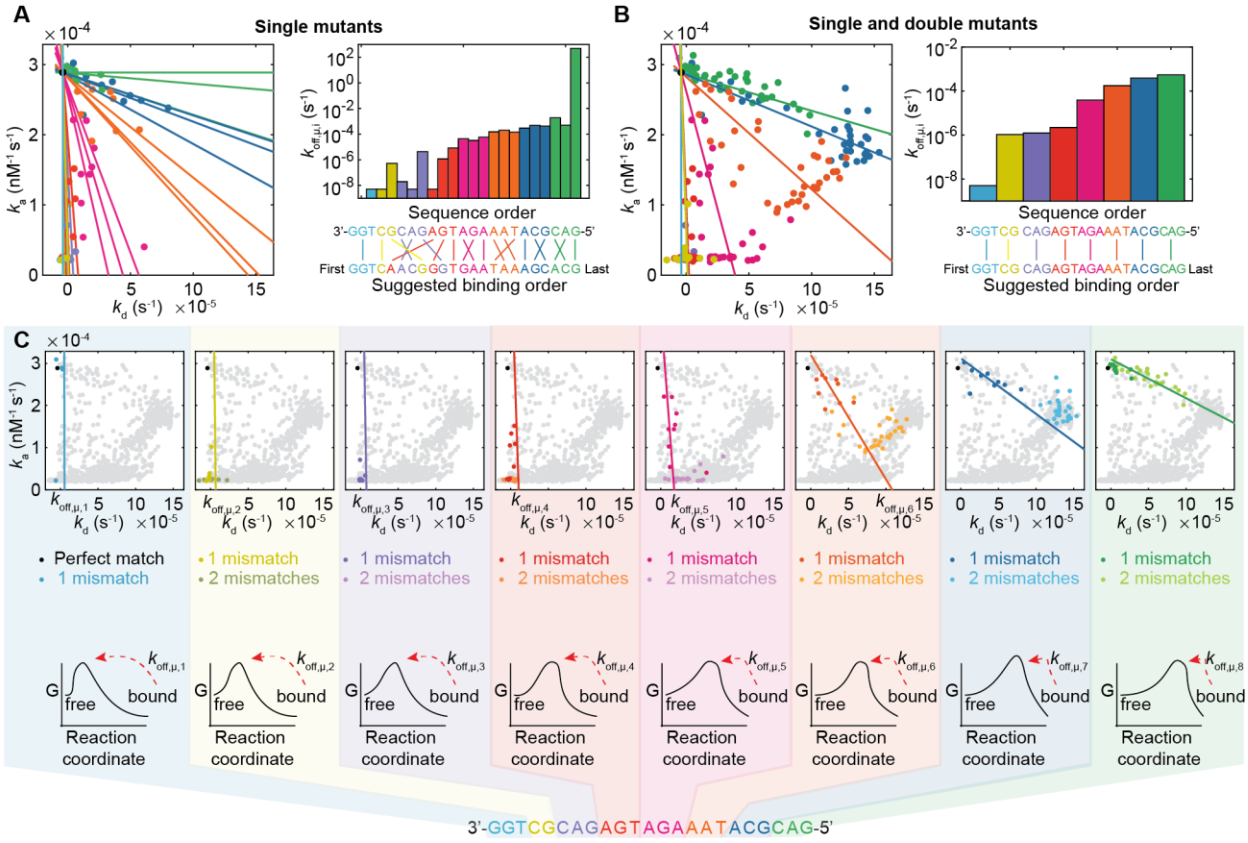

Fig. 3. The target-site recognition model describes the coupling between dCas9 off-target binding and unbinding. Triplets along the sequence are indicated by different colors. (**A**) Measured association and dissociation rates for dCas9 and different off-target sequences (single-site mutant data from (*14*)) (left). Linear fits to the data for each mutated base yield the estimated effective microscopic dissociation rate $k_{\mathrm{off},\mu,\mathrm{i}}$ (i.e. k_{d}-intercept) associated with each nucleotide, indicating the binding pathway (right). (**B**) Measured association and dissociation rates for dCas9 and different off-target sequences with single and double mismatches (*14*) (left), and estimated

effective microscopic dissociation rate $k_{off,\mu,i}$ associated with each 3-nucleotide DNA region, indicating the binding pathway (right). (**C**) Measured association and dissociation rates for dCas9 and different off-target sequences (grey dots; single and double mismatch mutant data from (*14*)), where data for mismatches occuring in specific gRNA regions are highlighted in colors. In (**A**) and (**B**), colored lines are individual fits of Eq. 1 to the single, or the single and double mismatch data, respectively. In (**C**), colored lines are representations of a global fit of a model with an 8-state sequential recognition to all the data. Here each line shows the model-predicted effect of mutations in one specific gRNA triplet region (see Supplementary Text and Fig. S5). Note that some of the data points in the plots are not highlighted in colors. These data points represent sequences with mutations in two different gRNA regions. In the right panels of (**A**) and (**B**) estimated $k_{off,\mu,i}$ values smaller than 5×10^{-9} s^{-1} have been rounded up to this value.

Supplementary Materials for

Sequence specificity in DNA binding is determined by association rather than dissociation

Emil Marklund, Guanzhong Mao, Sebastian Deindl*, Johan Elf*

*Correspondence to: sebastian.deindl@icm.uu.se, johan.elf@icm.uu.se

This PDF file includes:

Material and Methods

Expression, purification and fluorophore labelling of the *lac* repressor

Cy3 labelled *lac* repressor dimer (LacI-Far-2) was prepared according to previously published methods (*8*, *18*), with a Cy3 introduced distal from the DNA binding domain of LacI. Briefly, the protein contains a C-terminal 6xHis-tag for affinity purification purposes, and the C-terminal tetramerization domain has been removed. A cysteine for labeling was introduced at amino acid position 312. All cysteines found in the wild-type protein were removed from the sequence, except for the solvent-excluded cysteine in the monomer-monomer interface required to maintain an intact dimer (*18*).

DNA constructs for single-molecule fluorescence co-localization measurements

Double-stranded DNA constructs that contained operator sites as indicated, a backbone-incorporated Cy5 fluorophore attached to position 5 of a dT base via a 6-carbon linker (Integrated DNA Technologies), and an end-positioned biotin moiety were generated by annealing and ligating a set of overlapping, complementary oligonucleotides. High-performance liquid chromatography (HPLC)-purified oligonucleotides were mixed at equimolar concentrations in 50 mM Tris pH 8.0, 100 mM KCl, 1 mM EDTA, annealed with a temperature ramp (95–3°C), ligated with T4 DNA ligase (New England Biolabs), and purified by polyacrylamide gel electrophoresis (PAGE). Successful ligation was confirmed by denaturing PAGE.

Single-molecule colocalization microscopy

Biotinylated and fluorophore-labeled DNA constructs were surface-anchored onto PEG-coated quartz microscope slides through biotin-streptavidin linkage (*19, 20*). Cy3 and Cy5 dyes were excited with 532 nm Nd:YAG and 638 nm diode lasers, respectively, and fluorescence emissions from the two fluorophores were detected using a custom-built prism-based TIRF microscope, filtered with ZET532NF (Chroma) and NF03-642E (Semrock) notch filters, spectrally separated by 635 nm (T635lpxr) and 760 nm (T760lpxr) dichroic mirrors (Chroma), and imaged onto the separate regions of an Andor iXon Ultra 888 electron multiplying charge-coupled device (EMCCD) camera. Imaging was carried out in imaging buffer containing 20 mM K_2HPO4:KH_2PO4 pH 7.4, 1 mM β-Mercaptoethanol, 0.05 mM EDTA, 100 μg/ml acetylated BSA (Promega), 10% (v/v) glycerol, 10% (w/v) glucose, 0.01% Tween 20 (v/v), 2 mM Trolox to reduce photoblinking of the dyes (Rasnik et al), an enzymatic oxygen scavenging system (composed of 800 μg/ml glucose oxidase and 50 μg/ml catalase), as well as 1-300 mM NaCl (as indicated). LacI was introduced by infusing the sample chamber with imaging buffer supplemented with 0.5 nM LacI using a syringe pump (Harvard Apparatus). During image acquisition, a laser exposure time of 1 s was used. A frame rate of 0.5 Hz was used when detecting association and measuring k_a, except for one of the salt titration repeats for O_3 (cyan crosses, Fig. S2) where a 1 Hz frame rate was used. Directly following the association experiment, a movie at 1/6 Hz was collected for detecting dissociation and measuring k_d , except for one of the salt titration repeats for O_3 (green crosses, Fig. S2) where the association movie at 1 Hz frame rate was used for estimating dissociation. For one of the O_1 salt titrations (orange and brown crosses in Fig. S2) an additional movie at 1/12 Hz was captured directly after the 1/6 Hz movie. For this salt titration, k_d values were estimated individually from the 1/12 Hz (brown

crosses, Fig. S2) and 1/6 Hz (orange crosses Fig. S2) movie, while the same 0.5 Hz movies were used for estimating the corresponding k_a. All k_a and k_d were estimated from individual flow experiments except for the estimates for O_3 at 15 mM and 35 mM NaCl. For these two data points k_a and k_d were estimated from two flow experiments each, captured directly after each other.

Analysis of single-molecule colocalization measurements

The data were analyzed by summing the Cy3 fluorescence intensities within a 7 x 7-pixel square for each frame and each Cy5 dot. An à trous wavelet decomposition was used for dot detection in the Cy5 images (*21*). Dots were detected through scale-dependent standard deviation thresholding in the second wavelet plane with a threshold of three standard deviations, where the standard deviation was estimated by the median absolute deviation method (*22*). Dot centers were localized by calculating the weighted centroid from the pixel regions obtained from dot detection. Background intensities for pixels were estimated by a 2D moving average of each Cy3-fluorescence image, with exclusion of outliers by assigning them lower weights when calculating the average (*23*), so as not to include pixels corresponding to fluorescent dots when calculating the moving average. The fluorescence counts for each trace and frame were calculated as the difference between the raw fluorescence signal and the local background in the Cy3 image. Cumulative fluorescence curves were obtained by aligning and summing regions of single-molecule traces that had a current startpoint of the region corresponding to either low (association) or high (dissociation) fluorescence (Fig 2E) values. For the association curves, the trace regions being aligned and summed over started in a three-frame window just after LacI was introduced into the flow channel, and ended at the end of the movie. For the dissociation curves,

the trace regions being aligned and summed over started at any point in time more than 200 s before the end of the movie, and ended 200 s after the start point. Fluorescence traces were classified as 'low fluorescence' or 'high fluorescence' if the current count was below 20,000 or above 35,000, respectively. A threshold was set such that all counts above 50,000 were set to this value. Traces were excluded from the analysis if they had no counts higher than 50,000 within a 12-s moving-average window, indicating that no long-lived LacI binding occurred. To not bias results by including experiments with very weak binding, higher proportion of non-specific binding events compared to specific binding events, or higher proportion of binding to the glass surface, individual experiments were excluded from further analysis if they had a fraction of DNA spots with a binding event that was lower than 15% (See Fig. S1 for comparisons between operator and non-operator DNA). In total this excluded three data points at 200 and 250 mM NaCl from further analysis. For each experiment, the association and dissociation rates were estimated from the initial slopes of the cumulative curves. To account for photobleaching, we performed calibration dissociation experiments at different laser exposure times (fractional laser on times compared to the frame rate), and subtracted the constant contribution due to photobleaching from each k_d estimate (Fig. S3). The entire analysis pipeline was validated by performing stochastic simulations of binding and dissociation and simulated microscopy (*24*). This analysis method returned essentially the same association and dissociation rates as those that were put into the simulations (Fig. S4, Table S2). The simulations were performed with a number of DNA molecules that matches the average number of surface-immobilized DNA molecules found per field of view in the experiments (1400 Cy5 dots), and with imaging conditions mimicking the experiments in terms of level of background and fluorescence counts when LacI was bound. A frame rate of 0.5 Hz was used when simulating microscopy images for

k_a estimation, and a frame rate of 1/6 Hz was used when simulating microscopy images for k_d estimation. The reported values for $k_{off,\mu}$ in Fig. 1H are mean ± standard error of mean (s.e.m.), where each sample was obtained by fitting Eq. 1 to data from individual salt titrations, and $n = 2$ independent titration experiments for each operator.

Analysis and regression of *in vivo* association and dissociation measurements

Measured values and associated errors for association rates (*7*), dissociation rates (*13*), and equilibrium constants (*11*, *12*) for *lac* repressor binding to different operators were obtained from their respective references. In (*11*, *12*) equilibrium data was measured as the fold-change of repression, and is reported in binding energies of the repressor to its operator. With the model used in (*12*), these binding energies can be recalculated to equilibrium constants via

$$K_D = N_{genome}/exp(-\Delta\varepsilon), \tag{2}$$

where $N_{genome} = 5\times10^6$ base pairs is the size of the *E. coli* genome, and $\Delta\varepsilon$ is the binding energy to the operator as defined in (*12*). In (*11*) and (*12*), the fold-change is measured for the LacI tetramer and dimer, respectively. The model fit of the data in (*11*) is shown to perfectly describe the data in (*12*), demonstrating that these constructs have identical binding energies. Thus, we can use the values of the binding energies and the associated errors from (*11*) when estimating K_D for the LacI dimer from Eq. 2. Association and dissociation rates are here reported in units per cell volume, and since ~ 4 *lac* repressor molecules were searching for the target sites in (*7*), all association rates were divided by 4. The association rates used here are taken from the

additive model fit of all the data in (*7*), that is, they are extracted from the strains JE13, JE12, JE117, JE118, JE116, JE101 and JE104 containing different combinations of the O_{sym}, O_1, O_2, and O_3 operators. The reported values for $k_{off,\mu}$ were obtained by evaluating Eq. 1 with the experimentally estimated values for (k_a,k_d,$k_{on,,max}$), where $k_{on,max}= k_a/p_{tot}$, where k_a and p_{tot} are the values reported for O_1 in (*7*). Error bars for $k_{off,\mu}$ are 68 % confidence intervals obtained by resampling (k_a,k_d, $k_{on,,max}$) with the errors reported in (*7*, *11–13*), while assuming that the errors for all reported values are normally distributed.

Analysis and regression of dCas9 association and dissociation measurements

All data were taken from (*14*). Dissociation rates were taken from the dataset with chase, and association rates were obtained after combining the 1 nM and 10 nM datasets as described in (*14*). The colored lines in Fig. 3A and B were acquired by fitting Eq. 1 to the experimental data by minimizing the squared deviation between the model predicted (k_a,k_d) and the experimentally obtained (k_a,k_d), while constraining the line to go through the data point corresponding to the perfectly complementary sequence. In Fig. 3C colored lines are representations of a global fit of a model with an 8-state sequential recognition to all data (See Supplementary Text and Fig. S5).

Supplementary Text

Derivation of Equation 1 and the coupling between macroscopic association and dissociation

The mean first passage times for transitions between the different states of a continuous time Markov chain are given by

$$t_{ij} = t_{i,k\neq i} + \sum_{k\neq j} p_{i,k} t_{k,j} \quad , \tag{3}$$

where *i*, *j* and *k* are state indices, $t_{i,j}$ is the mean first passage time to transition from state *i* to state *j*, $t_{i,k\neq i}$ is the mean time to exit from state *i* into any of the other states in the model, and $p_{i,j}$ is the probability to transition to state *j* given that the model is currently in state *i*.

We now consider the three-state model shown in Fig. 1A of the main text and evaluate Eq. 3 for all possible transitions in the model, which gives two linear systems of equations

$$\begin{pmatrix} -1 & 1 \\ 1-p_{2,3} & -1 \end{pmatrix} \begin{pmatrix} t_{1,3} \\ t_{2,3} \end{pmatrix} = \begin{pmatrix} -t_{1,k\neq 1} \\ -t_{2,k\neq 2} \end{pmatrix} \qquad \text{(for } j=3\text{)} \tag{4}$$

and

$$\begin{pmatrix} -1 & p_{2,3} \\ 1 & -1 \end{pmatrix} \begin{pmatrix} t_{2,1} \\ t_{3,1} \end{pmatrix} = \begin{pmatrix} -t_{2,k\neq 2} \\ -t_{3,k\neq 3} \end{pmatrix} \qquad \text{(for } j=1\text{)}, \tag{5}$$

where we have used the fact that $p_{2,1} = 1 - p_{2,3}$. We now solve Eq. 4 and 5 for $1/t_{1,3}$ and $1/t_{3,1}$, which are the sought macroscopic association and dissociation rates. We then obtain

$$k_a[R] = \frac{1}{t_{1,3}} = \frac{p_{2,3}}{t_{1,k\neq 1} + t_{2,k\neq 2}} \tag{6}$$

and

$$k_d = \frac{1}{t_{3,1}} = \frac{1-p_{2,3}}{t_{3,k\neq 3} + t_{2,k\neq 2}} \quad , \tag{7}$$

where [R] is the concentration of the searching protein. We now solve for $p_{2,3}$ (= p_{tot}) in both Eq. 6 and 7, and equate the resulting expressions, which gives

$$k_a[R](t_{1,k\neq1} + t_{2,k\neq2}) = 1 - k_d(t_{3,k\neq3} + t_{2,k\neq2}). \quad (8)$$

We then assume that $t_{2,k\neq2} << t_{1,k\neq1}$ and that $t_{2,k\neq2} << t_{3,k\neq3}$, i.e. that the time spent nonspecifically bound is much shorter than both the time spent specifically bound, and the time spent dissociated from the relevant DNA region, so that the $t_{2,k\neq2}$ terms in Eq. 8 are negligible. We now solve for k_a[R], which gives

$$k_a[R] = \frac{1}{t_{1,k\neq1}} - \frac{t_{3,k\neq3}}{t_{1,k\neq1}}k_d \quad . \quad (9)$$

Using the same notations as in the main text, where $t_{1,k\neq1} = 1/ k_{on,max}[R]$ and $t_{3,k\neq3} = 1/ k_{off,\mu}$, yields the final expression,

$$k_a = k_{on,max} - \frac{k_{on,max}}{k_{off,\mu}}k_d \quad . \quad (10)$$

Extension of the model to handle multiple states of testing

The modeling framework can also be extended and used to determine the reaction mechanisms of binding and unbinding, if presented with measured kinetic rates of many binding site mutants. To achieve this, we extended the model to feature $N = n + 2$ number of states, and n number of testing states that the protein has to proceed through sequentially to reach the specifically bound

state, and where each testing state represents a group of nucleotides in the target sequence. The generalized model is thus

$$S_1 \underset{1-p_{2,3}}{\overset{k_{on,max}}{\rightleftharpoons}} S_2 \underset{1-p_{3,4}}{\overset{p_{2,3}}{\rightleftharpoons}} \cdots \underset{1-p_{i,i+1}}{\overset{p_{i-1,i}}{\rightleftharpoons}} S_i \underset{1-p_{i+1,i+2}}{\overset{p_{i,i+1}}{\rightleftharpoons}} \cdots \underset{1-p_{N-1,N}}{\overset{p_{N-2,N-1}}{\rightleftharpoons}} S_{N-1} \underset{k_{off,\mu}}{\overset{p_{N-1,N}}{\rightleftharpoons}} S_N. \quad (11)$$

The model is parameterized by two rates $k_{on,max}$ and $k_{off,\mu}$, defined by the diffusion-limited association time and the time spent specifically bound, along with n probabilities $p_{i,i+1}$, defining how likely it is for the protein to transition from one testing state to the next state in the sequential binding. The two linear equation systems obtained after evaluating Eq. 3 for the state transitions in the model are now

$$\begin{pmatrix} -1 & 1 & 0 & 0 & 0\cdots \\ 1-p_{2,3} & -1 & p_{2,3} & 0 & 0\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots & 1-p_{i,i+1} & -1 & p_{i,i+1} & 0\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots & 0 & 1-p_{N-2,N-1} & -1 & p_{N-2,N-1} \\ 0\cdots & 0 & 0 & 1-p_{N-1,N} & -1 \end{pmatrix} \begin{pmatrix} t_{1,N} \\ t_{2,N} \\ \vdots \\ t_{i,N} \\ \vdots \\ t_{N-2,N} \\ t_{N-1,N} \end{pmatrix} = \begin{pmatrix} -t_{1,k\neq 1} \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

(12)

and

$$\begin{pmatrix} -1 & p_{2,3} & 0 & 0 & 0\cdots \\ 1-p_{3,4} & -1 & p_{3,4} & 0 & 0\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots & 1-p_{i,i+1} & -1 & p_{i,i+1} & 0\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots & 0 & 1-p_{N-2,N-1} & -1 & p_{N-1,N} \\ 0\cdots & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} t_{2,1} \\ t_{3,1} \\ \vdots \\ t_{i,1} \\ \vdots \\ t_{N-1,1} \\ t_{N,1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ -t_{N,k\neq N} \end{pmatrix},$$

(13)

where we have used the assumption that the time spent in the nonspecific testing states is negligible, i.e. that $t_{i,k\neq i} = 0$ for all $i \in [2,N\text{-}1]$, and note that $t_{1,N} = 1/k_a[R]$, $t_{N,1} = 1/k_d$, $t_{1,k\neq 1} = 1/k_{on,max}[R]$ and $t_{N,k\neq N} = 1/k_{off,\mu}$.We now consider how changes in one specific $p_{i,i+1}$ parameter will change the k_a and k_d obtained from solving Eqs. 12 and 13, while keeping the rest of the model parameters constant. Since the time spent in any of the testing states is assumed to be 0, effectively the model is always in state 1 (free protein) or state N (bound protein). Hence, the passage times for reaching state i from state 1 must be exponentially distributed, with an average passage time of $1/k_{on,max}P_{1,i}$, where $P_{1,i}$ is the probability that the model will reach state i given that the model has just left state 1, before returning to state 1 again. Similarly, the passage times required to go from state N to state i must also be exponentially distributed, with an average passage time of $1/k_{off,\mu}P_{N,i}$, where $P_{N,i}$ is the probability that the model will reach state i given that the model has just left state N, before returning to state N again. If we then view states {1, 2,...i-1} as one state, state i as one state, and states {i+1,i+2,...N} as one state, this new model describes a three-state continuous time Markov chain, which we have shown to have k_a and k_d coupled by Eq. 1. In this case, when considering the effect of changing one $p_{i,i+1}$ parameter, we can rewrite Eq. 1 as

$$k_a = k_{on,max} P_{1,i} - \frac{k_{on,max} P_{1,i}}{k_{off,\mu} P_{N,i}} k_d, \tag{14}$$

where $P_{1,i}$ and $P_{N,i}$ are functions of $p_{j,k}$, where $j \neq i$ and $k \neq i+1$. With $k_{on,max,i-1} = k_{on,max} P_{1,i}$ and $k_{off,\mu,i-1} = k_{off,\mu} P_{N,i}$,we obtain the same notation as used in Fig. 3, where $k_{off,\mu,i-1}$ is the effective rate of transitioning to state *i* from state *N*. Using this notation in Eq. 14 gives an equation of the familiar form

$$k_a = k_{on,max,i-1} - \frac{k_{on,max,i-1}}{k_{off,\mu,i-1}} k_d. \tag{15}$$

The model described by Eqs. 12 and 13 can be used to show effects of mutations in certain sequence regions (certain $p_{i,i+1}$) on association and dissociation, which can be compared with experimental rates to deduce which nucleotides the protein contacts first, second, and so on in the recognition process of binding. To demonstrate this, we have presented such a model (Fig. S5A) with high-throughput data of association and dissociation available for dCas9 binding to off-target, mismatch mutants ((*14*), Fig. 3C). dCas9 is guided by an RNA (gRNA) when it binds DNA, with a reaction coordinate for the testing of recognition that is already well-established (*15–17*). Cas9 first detects a protospacer adjacent motif (PAM, NGG sequence), which is followed by DNA melting and gRNA-DNA hybridization , were the gRNA binds the DNA by base pairing in a sequential manner starting from a seed sequence, and then continuing hybridisation at base pairs more distal from the seed. If this sequential binding is completely memoryless, DNA binding sites with mismatches corresponding to gRNA region *j* should have the same microscopic dissociation rate ($k_{off,\mu,j}$) for the transition from a completely hybridized and specifically bound gRNA, to a melted region *j*. Just as our theory predicts according to Eq.

15, DNA binding sites with mutations in the same DNA region have association rates that are anti-correlated to the dissociation rates (Fig. 3C), where the k_d-intercept of each k_a versus k_d line is the effective microscopic dissociation rate $k_{off,\mu,j}$ associated with each DNA region. When mismatches are present in the seed sequence, the slope of the k_a versus k_d line line is steep (low $k_{off,\mu,j}$ and mostly k_a modulation). The further away from the seed sequence the mismatches are, the flatter the slope becomes (higher $k_{off,\mu,j}$ and more k_d modulation).

We fit a model with eight-step sequential recognition to the data, that is with eight regions in the on-target DNA and one $p_{i,i+1}$ parameter fitted for each unique sequence region, by solving Eq. 12 and 13 for $1/t_{1,N}$ and $1/t_{N,1}$ and choosing the parameters ($k_{on,max}$,$k_{off,\mu,n}$, $p_{2,3}$...) so that the summed squared deviation between the model predicted (k_a,k_d) and the experimentally obtained (k_a,k_d) is minimized. In total the model has 199 parameters, which are fitted to 2586 data values. With this model we then show the predicted effect of mutations for the individual DNA regions (colored lines in Fig. 3C, varying one $p_{i,i+1}$ for each line). The model fit captures the large-scale changes and sequence-dependent coupling between k_a (Fig. S5B) and k_d (Fig. S5C) for single-and double-mismatch mutants, also when the model is trained with only half of the measured k_a and k_d values, while being tested on the other half of the dataset (Fig. S5D). We note that the model is simplistic since it assumes that the testing of recognition is infinitely fast. Since dCas9 is observed to bind more off-targets than Cas9 can cleave with high efficiency (*16*), a more realistic model would be one where the time that Cas9 actually spends in the testing state is considered (*25*). This discrepancy is a likely reason why our simple model is bad at predicting rates for triple mutants when trained on single and double mutants (Fig. S5E).

Measurements of LacI operator binding in relation to previous work

In one of our previous papers, we performed salt titrations with LacI labeled with bifunctional rhodamine (LacI-R) and detected the binding of O_1 operators via single-molecule FRET (*8*). From the titration end points of these measurements we obtained K_d, $k_{a,obs}$ and $k_{d,obs}$ estimates of 0.0974 ± 0.0005 nM, 0.0033 ± 0.0003 $s^{-1}nM^{-1}$ and 0.0062 ± 0.0005 s^{-1} at 1 mM supplemented NaCl, or of 3.4 ± 0.6 nM, 0.0009 ± 0.0001 $s^{-1}nM^{-1}$ and 0.0089 ± 0.0007 s^{-1} at 80 mM supplemented NaCl. We note that these previous measurements, when compared with the measurements in this current work, were performed in a baseline imaging buffer with substantially higher ionic strength (10% glucose, 10% glycerol, 1mM NaCl, 0.05mM EDTA, 0.01% Tween 20, 0.1mg/ml BSA, 1mM 2-Mercaptoethanol, 2 mM Trolox, and 100mM K_2HPO4:KH_2PO_4 pH 7.4 in the previous work versus 20mM K_2HPO_4:KH_2PO_4 pH 7.4 in this current work), which makes it most suitable to compare the results in this current work with measurements from the previous work that were obtained with ~100 mM lower concentrations of supplemented NaCl. The earlier estimates above should therefore be compared with the K_d, k_a, and k_d values for LacI-Far-2 at 100 mM or 200 mM supplemented NaCl in this work (0.025 ± 0.028 nM, 0.088 ± 0.029 $s^{-1}nM^{-1}$ and 0.0014 ± 0.0017 s^{-1} or 0.13 ± 0.07 nM, 0.023 ± 0.0096 $s^{-1}nM^{-1}$, and 0.0023 ± 0.0004 s^{-1}, respectively). Furthermore, the previous work uses a LacI construct labeled at a different location, with which rates were measured using a different method. Most likely, the discrepancy between estimates from the current and earlier work is predominantly caused by the difference in how association events are detected. For the single-molecule FRET assay, more stringent selection occurred in the detection of association events, since the associating protein and the fluorophore must adopt a specific conformation capable of producing an interpretable FRET signal. In our single-molecule colocalization measurements, the

only requirement for the detection of an association event is that the protein must contain an intact fluorophore label. Effectively, when normalizing the association rate to the concentration of labeled protein, single-molecule FRET is expected to yield lower apparent association rate constants compared to the single-molecule colocalization measurements, just as observed when comparing the measurements for LacI-R with those for LacI-Far-2. We note that our previous work focused on measuring intramolecular properties of target search when LacI scans the DNA via 1D diffusion, and that the absolute value of the apparent intermolecular association rate constant $k_{a,obs}$ reported there does not influence any of the conclusions. Prior to our work, the K_d for the full length LacI tetramer binding to a 80bp DNA fragment with the O_1 operator has been estimated via nitrocellulose filter binding to be 0.16 nM at 50 mM KCl, and 0.43 nM at 100 mM KCl (*26*). Taken all together, we believe that the discrepancy in K_d estimates is in line with what can be expected from measurements with different protein constructs and batches that were carried out in different buffers and with different experimental methods.

If we assume that the observed $k_{d,obs}$ at 1 mM NaCl in the previous work is predominantly due to photobleaching (k_d= 0 and k_a= $k_{on,max}$ for this salt concentration), we can obtain estimates of k_d for the other salt concentrations in the single-molecule FRET measurements by fitting a linear equation (Eq. 1) to the six salt concentration data points, and by subtracting the k_d value corresponding to $k_a = k_{on,max}$ from all the data points. The k_d-intercept of this photobleaching-corrected line is then $k_{off,\mu}$. When we perform this analysis, we obtain $k_{off,\mu} = 0.0031 \pm 0.0002$ s^{-1}, which is identical to the $k_{off,\mu}$ value for LacI-Far-2 binding to O_1 as obtained in our single-molecule colocalization measurements in Fig 1H. For LacI-R, error bars in this section are standard errors obtained by propagating and resampling the experimental errors, while assuming that the errors reported in (*8*) are normally distributed, and the K_d estimates are reported after

correcting for photobleaching as described above. For LacI-Far-2, error bars are standard errors with $n = 2$ independent experiments.

Energy landscape for reaching the specific bound state via the putative reaction coordinate

To estimate and draw the energy landscapes *in vivo* for the different operators in Fig. 2D, we first consider the free energy difference between the free (state 1) and bound (state 3) state. Since the time spent in the testing state is much shorter than the time spent in the free and bound states, this free energy difference is directly given by the measured K_d values according to

$$\Delta G_{3\to1} = -log(K_d), \tag{16}$$

with the energy difference given in k_bT units. Next, we consider the free energy difference between the free state and the testing state, which is the same for all operators. This free energy difference is defined as

$$\Delta G_{1\to2} = -log(\frac{P_{testing}}{P_{free}}), \tag{17}$$

where $P_{testing}$ is the probability that the protein is testing for recognition within one sliding length from the operator, and P_{free} is the probability that the protein is searching somewhere else in the cell at any given time point. This probability can be calculated according to

$$P_{free} = P_{free,1D} + P_{free,3D}, \tag{18}$$

where $P_{free,1D}$ is the probability that the protein is bound nonspecifically to DNA somewhere else in the genome of the cell, and $P_{free,3D}$ is the probability that the protein is dissociated from DNA

and is searching in the cytoplasm. The fraction of time *f* that the protein spends nonspecifically bound to DNA when searching is thus defined as

$$f = \frac{P_{free,1D}}{P_{free,1D} + P_{free,3D}}. \tag{19}$$

After combining Eqs. 17-19 we obtain

$$\Delta G_{1\to 2} = log(\frac{P_{free,1D}}{P_{testing}} + \frac{1-f}{f}\frac{P_{free,1D}}{P_{testing}}) \tag{20}$$

where the ratio $P_{\text{free,1D}}$ /P_{testing} can be calculated from the size of the genome N_{genome} and the average sliding length N_{sliding} as

$$\frac{P_{free,1D}}{P_{testing}} = \frac{N_{genome}}{N_{sliding}}. \tag{21}$$

With $N_{\text{genome}} = 5\times10^6$ base pairs, $N_{\text{sliding}} = 45$ base pairs (*7*), and $f = 0.9$ (*27*) we obtain the free energy difference shown in Fig. 2D.

The relative difference in activation energy on the transition path between the testing state (state 2) and bound state (state 3) for the different operators can be calculated from the measured $k_{\text{off},\mu}$ values. To achieve this, we model the transition state that the protein has to go through to reach the bound state from the testing state as a distinct species. This modeling scheme is similar to what is used in transition state theory (TST), but here we model the transition state as a true equilibrated Markovian state, instead of as a quasi-equilibrated state as is done in TST. The model is thus

$$S_2 \underset{1-p_{\ddagger,3}}{\overset{k_{2,\ddagger}}{\rightleftharpoons}} S_{\ddagger} \underset{k_{3,\ddagger}}{\overset{p_{\ddagger,3}}{\rightleftharpoons}} S_3, \tag{22}$$

where S_2 is the testing state, $S_{\ddagger}$ is the transition state, and S_3 is the bound state. As we have shown, Eq. 7 describes how the effective transition rate from the last state to the first state in this type of model can be calculated. This effective rate is now $k_{\text{off},\mu}$, and with the notations used in Eq. 22, Eq. 7 is for this model

$$k_{off,\mu} = \frac{1 - p_{\ddagger,3}}{\frac{1}{k_{3,\ddagger}} + t^{\ddagger}}, \tag{23}$$

where $t^{\ddagger}$ is the average time that the model spends in the transition state. Furthermore, the free energy difference between the bound state and the transition state is defined as

$$\Delta G_{3\to\ddagger} = -log(\frac{k_{3,\ddagger}}{k_{\ddagger,3}}) = -log(\frac{k_{3,\ddagger}}{p_{\ddagger,3}\frac{1}{t^{\ddagger}}}). \tag{24}$$

Solving for $k_{3,\ddagger}$ in Eq. 23 and putting this into Eq. 24 gives

$$\Delta G_{3\to\ddagger} = -log(k_{off,\mu}) + log(1 - p_{\ddagger,3} - k_{off,\mu}t^{\ddagger}) + log(\frac{p_{\ddagger,3}}{t^{\ddagger}}). \tag{25}$$

When $1/k_{\text{off},\mu} >> t^{\ddagger}$, i.e. when the time spent in the transition state is very small, Eq. 25 can be written as

$$\Delta G_{3\to\ddagger} = -log(k_{off,\mu}) + c, \tag{26}$$

where

$$c = log(1 - p_{\ddagger,3}) + log(\frac{p_{\ddagger,3}}{t^{\ddagger}}). \tag{27}$$

If we now assume that $t^{\ddagger}$ and $p_{\ddagger,3}$ are the same for all operators, the difference in $\Delta G_{3\to\ddagger}$ for different operators, i.e. the difference in energy barrier between the bound state and the transition state for different operators in Fig. 2D, is given by the difference in -log($k_{off,\mu}$) for the different operators, where a fast $k_{off,\mu}$ gives a low energy barrier, and a slow $k_{off,\mu}$ gives a high energy barrier.

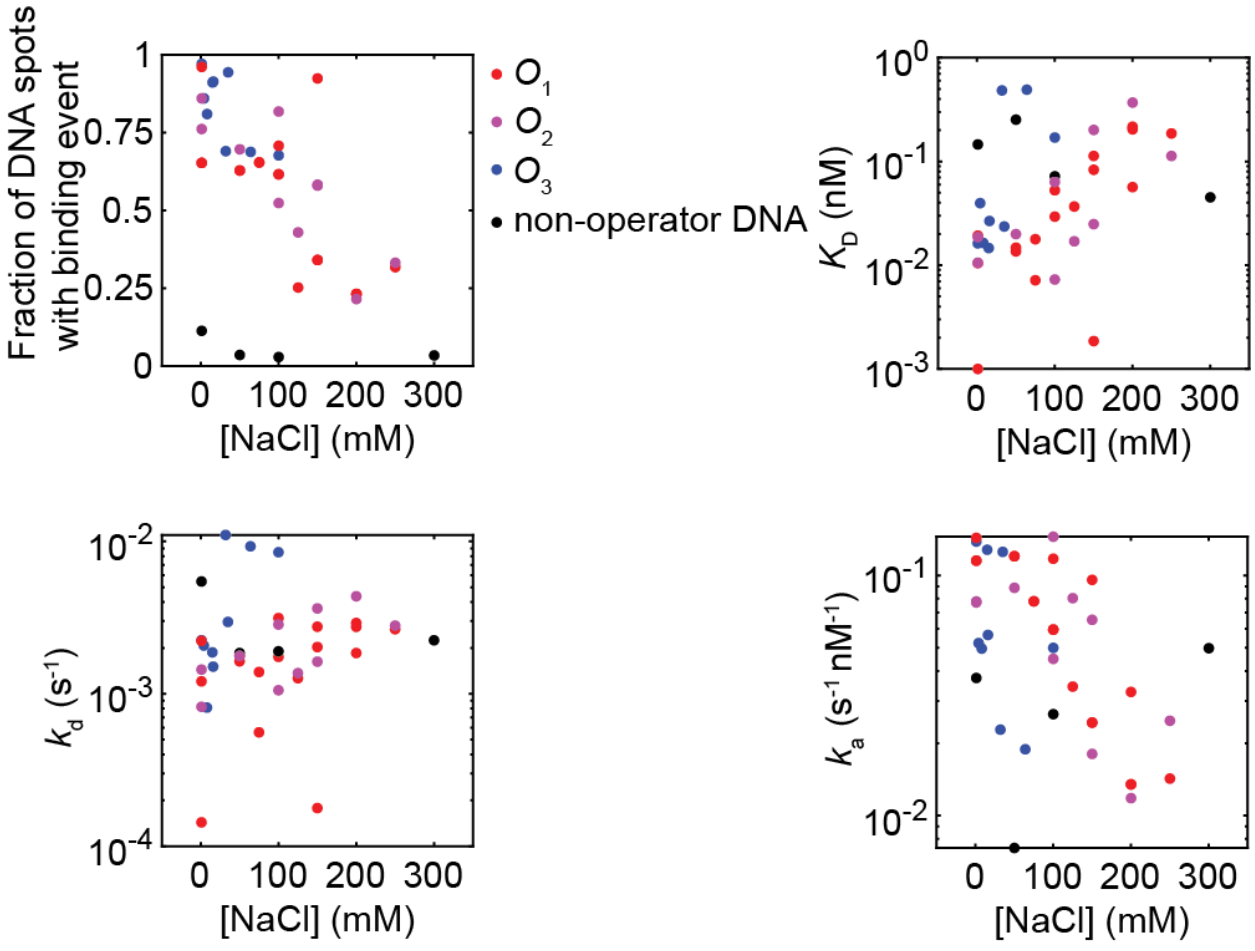

Fig. S1. Binding and specificity in LacI-DNA colocalization measurements. The fraction of DNA spots that had at least one LacI binding event, as a function of the salt concentration of the experiment, for different operators and one non-operator DN (top left). A binding event was detected when a trace had a fluorescence count higher than 50,000 in a 12-s moving-average window. K_d (top right), k_d (bottom left) and k_a (bottom right) for the different DNA constructs as a function of the salt concentration of the experiment.

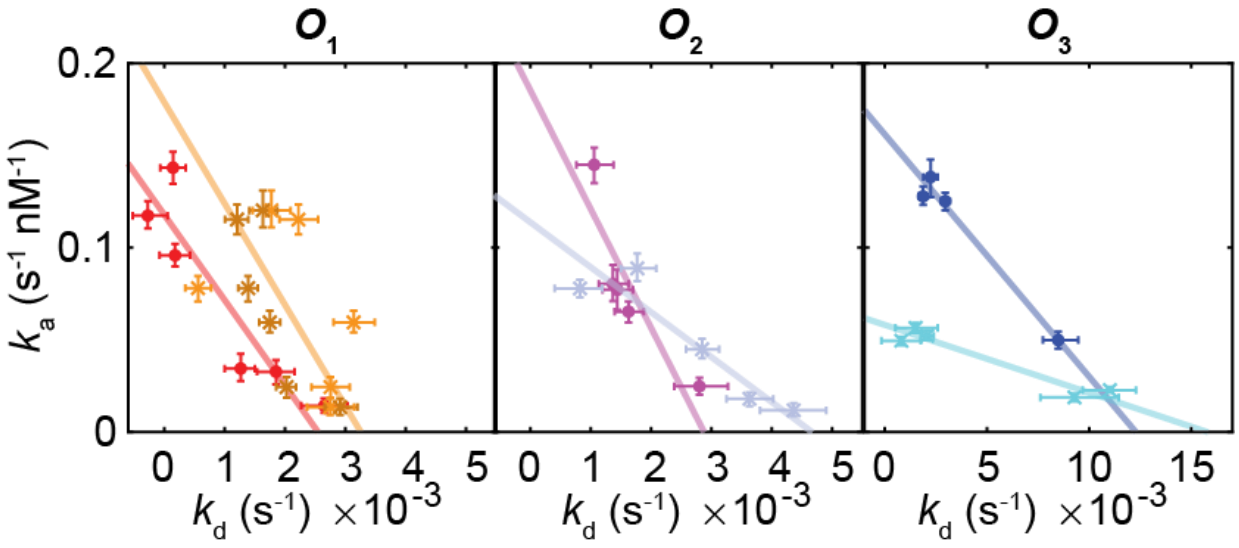

Fig. S2. Single-molecule colocalization measurements for salt titration repeats. Measured k_a and k_d values from two salt titration repeats (dots and crosses) for the *lac* operators, and fits to Eq. 1 for each repeat (colored lines). For one of the O_1 salt titrations, k_d was estimated with both 1/6 Hz (orange crosses) and 1/12 Hz (brown crosses) frame rate, while the corresponding k_a was estimated once with a 0.5 Hz frame rate. See Methods for additional experimental conditions. Error bars are 68 % confidence intervals obtained by bootstrapping the fluorescence traces.

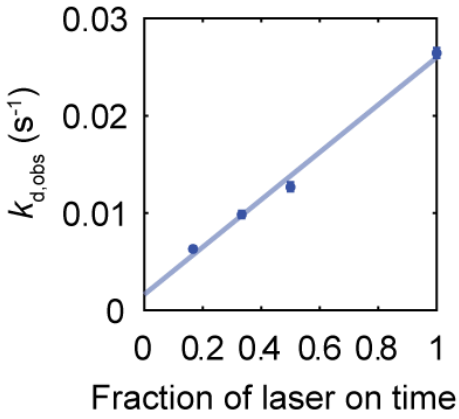

Fig. S3. Bleaching calibration for k_d estimate. Observed dissociation rates $k_{d,obs}$, obtained as the initial slopes of the dissociation curves, plotted against the fractional exposure time f used in the experiment. $k_{d,obs}$ was measured for O_3 dissociation at 1 mM NaCl, with a laser exposure time of 1 s, and frame rates of 1,0.5, 1/3 and 1/6 Hz (blue points). The blue line is the best fit to the equation $k_{d,obs} = k_d + f\,k_{bleach}$, so that $f\,k_{bleach}$ from the fit can be subtracted from $k_{d,obs}$ to obtain k_d for each measurement. Error bars are 68 % confidence intervals obtained by bootstrapping the fluorescence traces in each experiment.

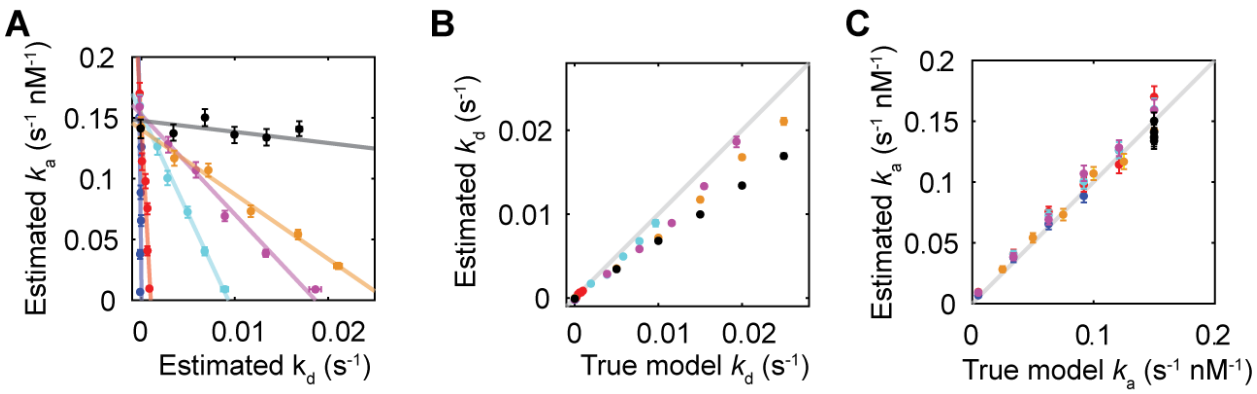

Fig. S4. Stochastic simulations and simulated microscopy. (**A**) Estimated k_a and k_d for the simulated data. (**B**) True and estimated k_d values. (**C**) True and estimated k_a values. See Table S2 for true and estimated $k_{off,\mu}$ values for the simulated data. All error bars are 68 % confidence intervals obtained by bootstrapping the simulated fluorescence traces.

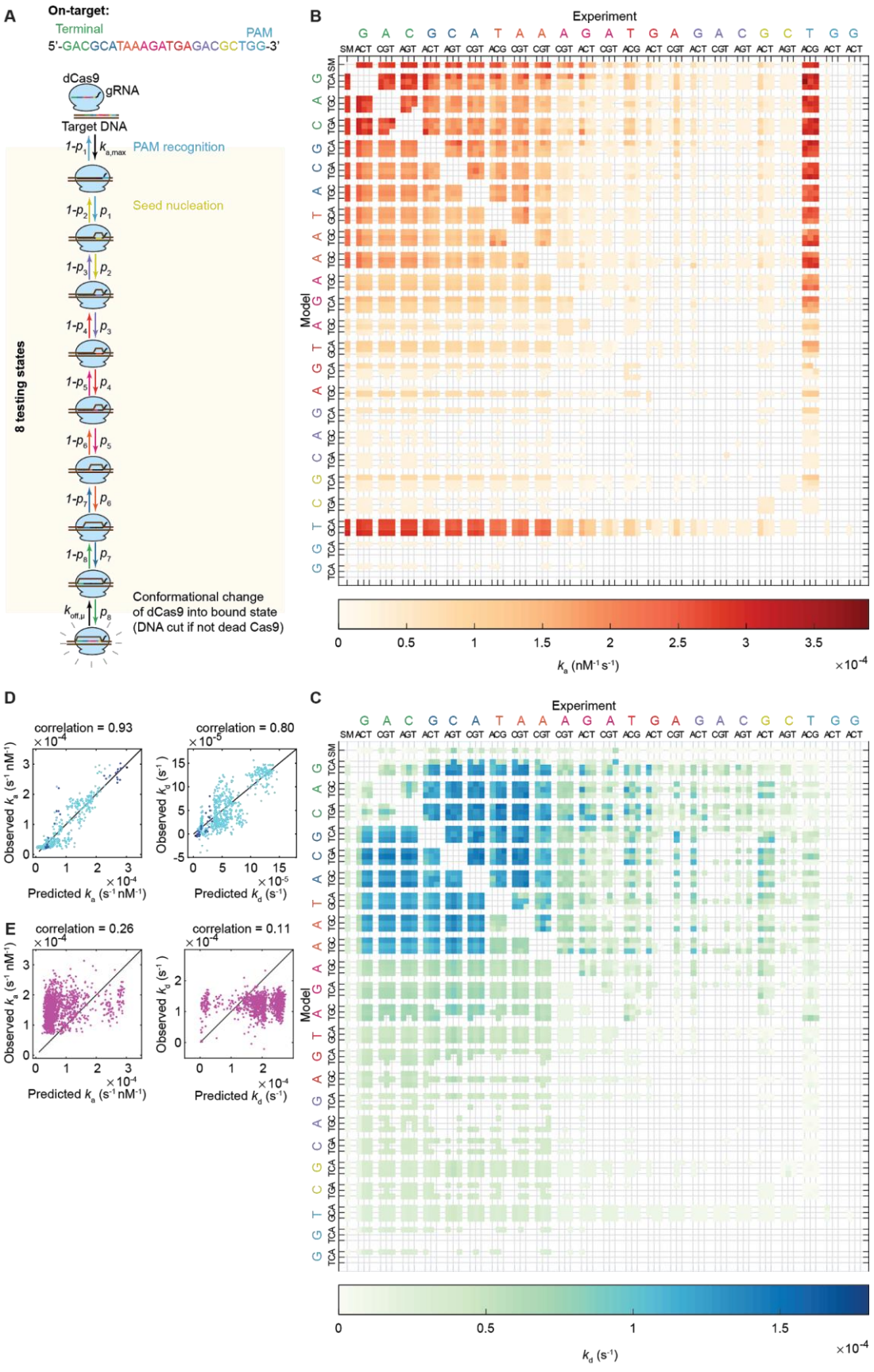

Fig. S5. Model fit of dCas9 off-target binding and unbinding data. (**A**), Cartoon model diagram. Measured (above diagonal) and model-predicted (below diagonal) association (**B**) and dissociation (**C**) rates for single and double off-target mutants, when simultaneously fitting association and dissociation rates. The on-target, all single mutants, and all double mutants except the ones containing a mutation in the degenerate PAM (T in TGG) were used in the training data set. (**D**) Correlation between the model-predicted and observed association (left) and dissociation (right) rates when half of the data set (random sample) in (**B**,**C**) was used for training, and the other half of the data set was used for testing (points in the plots), for the on-target (black), single (blue) and double mutants (cyan). (**E**), Correlation between the predicted and observed association (left) and dissociation (right) for triple mutants when the same data set as in (**B**,**C**) is used for training the model.

Table S1. Sequence of LacI and DNA constructs for single-molecule colocalization measurements. For LacI-Far-2, the cysteine introduced for labeling is marked in red.
For the DNA constructs, modifications are reported using Integrated DNA Technologies (IDT) nomenclature. LacI operator sites are highlighted in orange.

name	Sequence
LacI-Far-2	MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAEL NYIPNRVAQQLAGKQSLLIGVATSSLALHAPSQIVAAIKSRADQLGAS VVVSMVERSGVEAAKAAVHNLLAQRVSGLIINYPLDDQDAIAVEAAA TNVPALFLDVSDQTPINSIIFSHEDGTRLGVEHLVALGHQQIALLAGPL

	SSVSARLRLAGWHKYLTRNQIQPIAEREGDWSAMSGFQQTMQMLNE GIVPTAMLVANDQMALGAMRAITESGLRVGADISVVGYDDTEDSSCY IPPLTTIKQDFRLLGQTSVDRLLQLSQGQCVKGNQLLPVSLVKRKTTL APNTQTHHHHHH
O_1 construct	Top strand: 5'- /5BioTinTEG/TCGTACTTCAAGTTTTGGGCGTGTCAAGTCCAAGGATT GC TCTGTATACTTAAAAACGACGTGGCAGTAAAGGGAACGCAAGACT CTCAATCGCAATTGTTATCCGCTCACAATTCCGAAAGCCT-3' Bottom strand: 5'- AGGCT/iCy5/TCGGAATTGTGAGCGGATAACAATTGCGAATGAGAGT CT TGCGTTCCCTTTACTGCCACGTCGTTTTTAAGTATACAGAGCAATCC TTGGACTTGACACGCCCAAAACTTGAAGTACGA-3'

O_2 construct	Top strand: 5'- /5BioTinTEG/TCGTACTTCAAGTTTTGGGCGTGTCAAGTCCAAGGATTGC TCTGTATACTTAAAAACGACGTGGCAGTAAAGGGAACGCAAGACT CTCAATCGCGGTTGTTACTCGCTCACATTTCCGAAAGCCT-3' Bottom strand: 5'- AGGCT/iCy5/TCGGAAATGTGAGCGAGTAACAACCGCGAATGAGAG TCT TGCGTTCCCTTTACTGCCACGTCGTTTTTAAGTATACAGAGCAATCC TTGGACTTGACACGCCCAAAACTTGAAGTACGA-3'

O_3 construct	Top strand: 5'- /5BioTinTEG/TCGTACTTCAAGTTTTGGGCGTGTCAAGTCCAAGGATTGC TCTGTATACTTAAAAACGACGTGGCAGTAAAGGGAACGCAAGACT CTCAATCGCGGCAGTGAGCGCAACGCAATTCCGAAAGCCT-3' Bottom strand: 5'- AGGCT/iCy5/TCGGAATTGCGTTGCGCTCACTGCCGCGAATGAGAGT CT TGCGTTCCCTTTACTGCCACGTCGTTTTTAAGTATACAGAGCAATCC TTGGACTTGACACGCCCAAAACTTGAAGTACGA-3'

Non-operator construct	Top strand: 5'- /5BioTinTEG/TCGTACTTCAAGTTTTGGGCGTGTCAAGTCCAAGGATTGC TCTGTATACTTAAAAACGACGTGGCAGTAAAGGGAACGCAAGACTCTCA /iCy5/TCGCGATTGCAGCTCGAAGCAGCATCCGAAAGCC-3' Bottom strand: 5'- GGCTTTCGGATGCTGCTTCGAGCTGCAATCGCGAATGAGAGTCTTGCG TTCCCTTTACTGCCACGTCGTTTTTAAGTATACAGAGCAATCCTTGGACTTGACACGCCCAAAACTTGAAGTACGA-3'

Table S2. Stochastic simulations and simulated microscopy. True values of $k_{off,\mu}$ for k_a and k_d values and lines put into stochastic simulations, and estimated $k_{off,\mu}$ values obtained after running simulated microscopy and the analysis pipeline. Estimations of $k_{off,\mu}$ are given as 68% confidence

intervals obtained by bootstrapping the simulated fluorescence traces. See Fig. S4 for true and estimated k_a and k_d values for these $k_{off,\mu}$ lines.

True $k_{off,\mu}$ (s^{-1})	$1\cdot 10^{-4}$	$1\cdot 10^{-3}$	$1\cdot 10^{-2}$	$2\cdot 10^{-2}$	$3\cdot 10^{-2}$	100
Estimated $k_{off,\mu}$ (s^{-1})	[-3.6,1.1] $\cdot 10^{-4}$	[0.84,1.2] $\cdot 10^{-3}$	[0.89,0.96] $\cdot 10^{-2}$	[1.8,1.9] $\cdot 10^{-2}$	[2.5,2.7] $\cdot 10^{-2}$	[8.7,∞] $\cdot 10^{-2}$