Published online 4 May 2023

Nucleic Acids Research, 2023, Vol. 51, No. 11 ***5847–5863***
https://doi.org/10.1093/nar/gkad323

Position-dependent sequence motif preferences of SpCas9 are largely determined by scaffold-complementary spacer motifs

Krisztina Huszár[1,2,3], Zsombor Welker[1,4], Zoltán Györgypál[4,5], Eszter Tóth[1,3], Zoltán Ligeti[1,6,7], Péter István Kulcsár[1], János Dancsó[1,4], András Tálas[1], Sarah Laura Krausz[1,8], Éva Varga[1,6,7] and Ervin Welker[1,6,*]

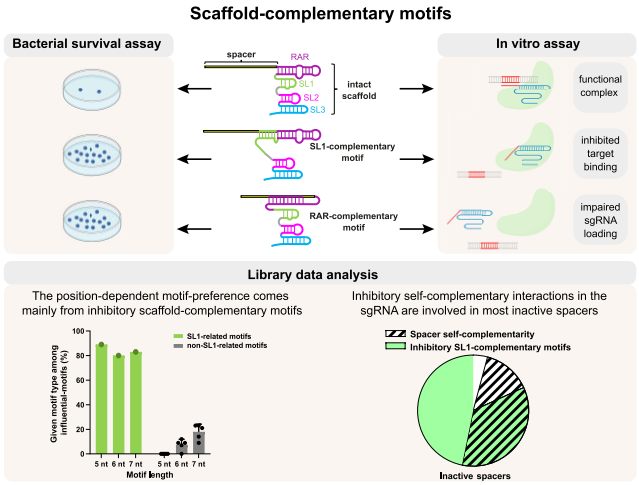
[1]Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary, [2]Department of Genetics, Doctoral School of Biology, Faculty of Science, Eötvös Loránd University, Budapest, H-1117, Hungary, [3]Gene Design Ltd, Szeged, Hungary, [4]Biospiral-2006 Ltd, Szeged, Hungary, [5]Institute of Biophysics, Biological Research Centre, Szeged, Hungary, [6]Institute of Biochemistry, Biological Research Centre, Szeged, Hungary, [7]Doctoral School of Multidisciplinary Medical Science, University of Szeged, Hungary and [8]School of Ph.D. Studies, Semmelweis University, Budapest, Hungary

Received July 23, 2022; Revised April 04, 2023; Editorial Decision April 04, 2023; Accepted May 02, 2023

ABSTRACT

Streptococcus pyogenes **Cas9 (SpCas9) nuclease exhibits considerable position-dependent sequence preferences. The reason behind these preferences is not well understood and is difficult to rationalise, since the protein establishes interactions with the target-spacer duplex in a sequence-independent manner. We revealed here that intramolecular interactions within the single guide RNA (sgRNA), between the spacer and the scaffold, cause most of these preferences. By using *in cellulo* and *in vitro* SpCas9 activity assays with systematically designed spacer and scaffold sequences and by analysing activity data from a large SpCas9 sequence library, we show that some long (>8 nucleotides) spacer motifs, that are complementary to the RAR unit of the scaffold, interfere with sgRNA loading, and that some motifs of more than 4 nucleotides, that are complementary to the SL1 unit, inhibit DNA binding and cleavage. Furthermore, we show that intramolecular interactions are present in the majority of the inactive sgRNA sequences of the library, suggesting that they are the most important intrinsic determinants of the activity of the SpCas9 ribonucleoprotein complex. We also found that in pegRNAs, sequences at the 3′ extension of the sgRNA that are complementary to the SL2 unit are also inhibitory to prime editing, but not to the nuclease activity of SpCas9.**

GRAPHICAL ABSTRACT



INTRODUCTION

Streptococcus pyogenes Cas9 (SpCas9) is the most frequently explored nuclease of the type II CRISPR adaptive immune system of bacteria and archaea for gene editing, genome remodelling and transcriptome modulating applications (1–4). It cleaves both strands of the target DNA by its two nuclease domains in a sequence specific manner (5–8). SpCas9 is active in a ribonucleoprotein (RNP) form in complex with two small RNA molecules, the crRNA and tracrRNA that are associated through complementary segments forming the repeat : anti-repeat (RAR) duplex. For applications, a single guide RNA (sgRNA) is used, which is formed by covalently connecting the crRNA and the

*To whom correspondence should be addressed. Tel: +361 382 6610; Email: welker.ervin@ttk.hu

tracrRNA via a tetraloop (5). The first twenty nucleotides at the 5′ end of the sgRNA make up the variable spacer sequence, which provides sequence specificity to the nuclease by the virtue of its complementarity to the target sequence, while the invariable, about 80-nt long 3′ section of the sgRNA (9), called the scaffold sequence, mediate the interactions with the SpCas9 protein. The scaffold sequence consists of four structural units; the repeat: anti-repeat duplex (RAR) and stem loop 1–3 (SL1, SL2 and SL3, Figure 1A), according to the nomenclature of Nishimasu *et al.* (6).

In theory, SpCas9 could be programmed to bind and cleave any possible 20-nt long sequence as long as they are complementary to its 20-nt long spacer sequence, however, in reality, its activity varies widely from target sequence to target sequence; some of them being completely defiant to cleavage. The sequence restrictions, that are accounted for by interactions with various factors of the complex cellular environment, resulting in the inhibition of target binding or decreased sgRNA expression levels, are typically not position-dependent (10–13). These types of interactions do not explain many of the observed sequence preferences, including spacer position-dependent effects and the known phenomenon of targets resisting cleavage even *in vitro* and in bacterial cells, in which settings far fewer factors are expected to interfere with nuclease activity. We refer the corresponding spacers as 'inactive'. Interestingly, there is a discrepancy in the identified position-dependent shorter and longer motifs across different data sets. Position-dependent mono- and dinucleotide preferences found by several studies employing different data sets vary greatly (14–29). However, preferences for some trinucleotide and longer motifs seem to be more consistent amongst different data sets, many of them are even shared between data sets from mammalian and bacterial libraries (16,22). Longer (>5-nt long) motifs can only be identified by exploring the largest target libraries (16). The most commonly discussed preference may be the GCC motif at the PAM-proximal positions (18-20, 17-19 and 16-18) of the spacer (16,18,30,31), which is associated with reduced cleavage activity of SpCas9 as well as some of the increased-fidelity variants, like eSpCas9 and SpCas9-HF1 (16,22). It has been proposed that this inhibitory effect of the GCC sequence could be the direct result of inaccessibility of the seed region of the spacers, inefficient loading, non-specific binding to off-targets or cofactor-dependent mechanistic problems (30). However, the underlying mechanisms behind the position-dependent motif preferences are not yet understood. The origins of these types of sequence preferences are even more unclear in light of the fact that the X-ray structure of the RNP complex reveals predominantly sequence-independent interactions between the protein and the RNA:DNA hybrid helix (6).

Although the motif preferences are not well-understood, several factors have been identified within the sgRNA itself that influence the activity of SpCas9. Interestingly, by exploring data from just 1841 spacers [generated by Doench *et al.* (14)], Wong *et al.* discovered several of these factors (21),—that were later confirmed using much larger libraries (16),—such as the free energy of the whole sgRNA, the stability of the spacer/target DNA duplex and the accessibility of nucleotides in positions 18–20 of the spacer and in 51–53 of the scaffold. The low free energy of the spacer indicating the presence of self-complementarity within the spacer, is also identified as one of the most impactful features (16). It has been also revealed that interactions between the spacer and the scaffold sequences could decrease the activity of SpCas9. Wong *et al.* suggested that the decreased accessibility of positions 18–20 of the spacer and in positions 51–53 of the scaffold actually derives from their base pairing that extends the RAR duplex and decreases the activity of SpCas9 (21). Konstantakos and colleagues pointed out that these nucleotides of the spacer need to be a CUU sequence to be complementary to the scaffold sequence in question (18), thus, this effect may also contribute to the inhibitory effects of the PAM-proximal UU nucleotides observed in a former study (30). Thyme *et al.* have experimentally proved that internal sgRNA interactions could reduce the activity of SpCas9 (10).

We proposed that the sequence preferences of SpCas9 may arise from the effect of interactions between the sequences of the spacer and the scaffold. Here, we systematically examined the impact of such self-complementarity by monitoring the effects of specifically designed self-complementary sgRNAs and by analysing cleavage data derived from a one million sequence library we published recently (16). We found that the activity of SpCas9 is largely unaffected by the presence of scaffold-complementary motifs in the spacer sequences that are shorter than 9-nt unless they are complementary to the SL1 unit. This latter effect is what explains the spacer position-dependent motif preferences. Furthermore, the tolerance level of SpCas9 to the inhibitory effects of self-complementary motifs within the spacer or between the spacer and the SL1 stem of the scaffold seems to be the main determinant of the inactive spacer sequences. We have also shown that scaffold-complementary motifs may also inhibit prime editing activity when they are located in the 3′ region of the prime editing guide RNA (pegRNA), even if they were only 5-nt in length.

MATERIALS AND METHODS

Materials

Restriction enzymes, T4 ligase, TranscriptAid T7 High Yield Transcription Kit, Dulbecco's modified Eagle's medium (DMEM), fetal bovine serum, Turbofect and Penicillin/Streptomycin were purchased from Thermo Fischer Scientific. DNA oligonucleotides and the GenElute HP Plasmid Miniprep kit used in plasmid purifications were acquired from Sigma-Aldrich. T4 Polynucleotide Kinase, Q5 High-Fidelity DNA Polymerase, NEB Stable Competent *E. coli*, HiFi Assembly Master Mix were from New England Biolabs Inc. NucleoSpin Gel and PCR Clean-up kit used to clean up DNA from agarose gels were purchased from Macherey-Nagel. ZymoPURE Plasmid Midiprep Kit and RNA Clean & Concentrator Kit were from Zymo Research. Ampicillin, kanamycin and chloramphenicol antibiotics were from Sigma-Aldrich, NaCl, yeast extract and tetracycline were from Molar chemicals, agarose and tryptone were from VWR.

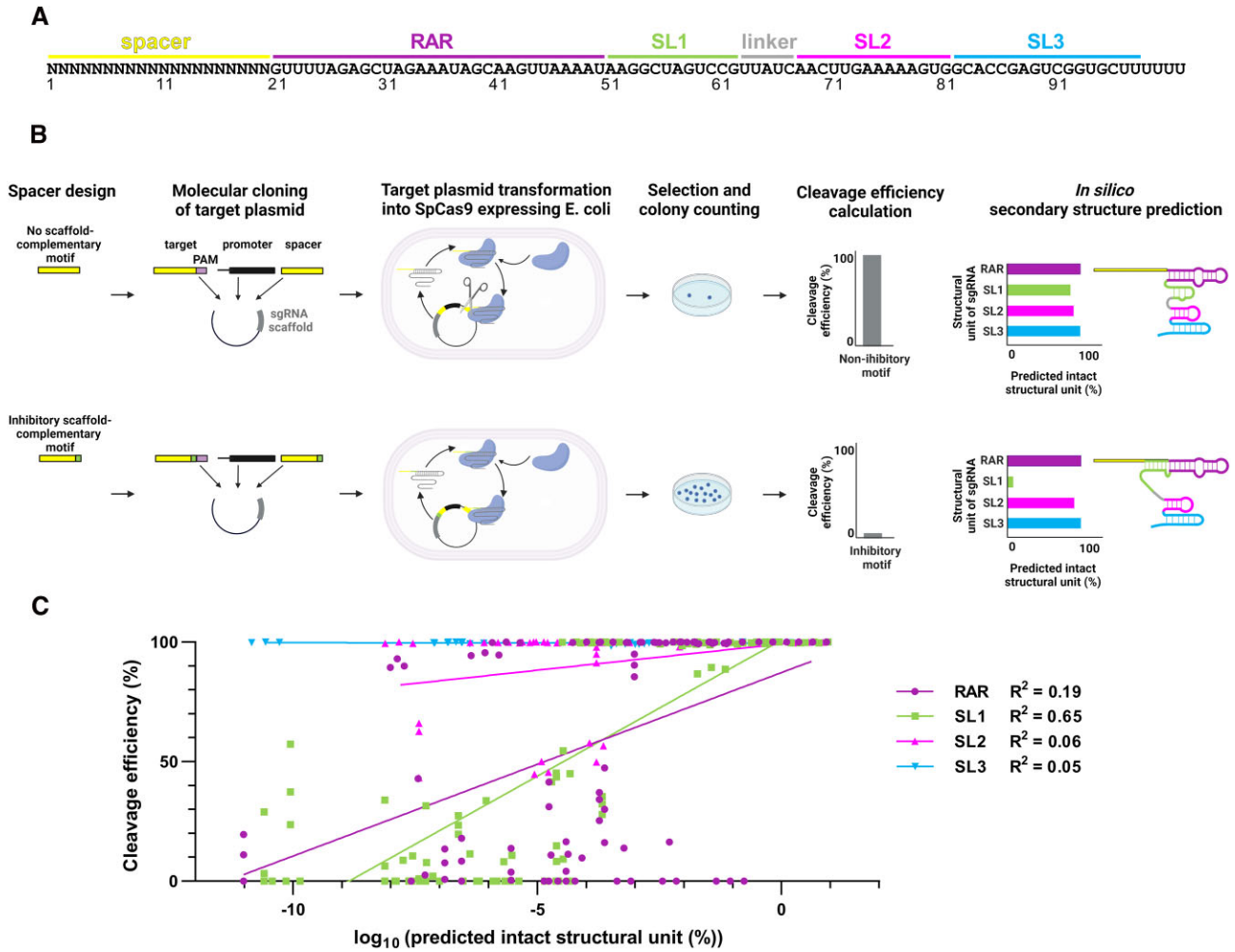


Figure 1. The effect of scaffold-complementary spacer motifs on the structure and function of sgRNAs. (**A**) The sequence of the scaffold of sgRNAs (9) that is used and referred to as wild type (WT) scaffold in this study. (**B**) The workflow of the bacterial survival assay, which allows the identification of the inhibitory effect of scaffold-complementary motifs designed into the spacers of the SpCas9 complex. Bacteria with plasmids encoding motifs that cause misfolding in the sgRNAs and thereby inhibit the SpCas9 complex survive the selection scheme. Cleavage efficiency is calculated based on the number of the surviving colonies. The presence of intact structural units in sgRNAs are predicted *in silico* to determine the extent of disruption to the secondary structure by the presence of scaffold-complementary motifs. Details are described in Materials and Methods. (**C**) Scatter plot shows the cleavage efficiency of SpCas9 with a given spacer and the corresponding predicted percentage of the sgRNA with an intact structural unit in the conformational ensemble of the lowest energy structures ($r = 8$), as determined by using an in-house software, RNAfold-wg. Data points are represented by dots and correspond to one of three replicates in the bacterial survival assay for spacers with RAR- ($n = 30$), SL1- ($n = 32$), SL2- ($n = 15$) and SL3- ($n = 10$) complementary motifs. Each spacer is assigned to the structural unit showing the lowest percentage value for the intact predicted secondary structure and this percentage value is shown. R square is the coefficient of determination for a linear fit. Further detailed data of the linear regressions are provided in Source Data Figure 1.

Plasmid construction

Vectors were constructed using standard molecular biology techniques. For detailed cloning, plasmid and oligo sequence information see Supplementary Information and Supplementary Table S1. A list of spacers containing self-complementary motifs, rescue- and extended-scaffold sequences used in the bacterial survival assay are available in Supplementary Table S2. The sequences of all plasmid constructs were confirmed by Sanger sequencing (Microsynth AG).

Plasmids acquired from the non-profit plasmid distribution service Addgene (http://www.addgene.org/) are the following: pdCas9-bacteria and pwtCas9-bacteria was a gift from Stanley Qi [Addgene plasmid # 44249; http://n2t.net/addgene:44249; RRID:Addgene_44249 and Addgene plasmid # 44250; http://n2t.net/addgene:44250; RRID:Addgene_44250 (32)], pET-dCas9-VP64-6xHis and pCMV-PE2 was a gift from David Liu [Addgene plasmid # 62935; http://n2t.net/addgene:62935; RRID:Addgene_62935 (33) and Addgene plasmid # 132775; http://n2t.net/addgene:132775; RRID:Addgene_132775 (34)]

Plasmids previously developed by our research group and deposited at Addgene are the following:

pX330-Flag-wtSpCas9 [#92353 (35)], pX330-Flag-dSpCas9 [#92113 (35)], pAT9624-BEAR-cloning [#162986 (36)], pDAS12069-U6-pegRNA-mCherry [#177180 (37)].

Bacterial survivor assay for testing the activity of sgRNAs

The bacterial selection system used here employs two-plasmids in *E. coli.* Plasmids expressing either active or inactive (dead or dSpCas9 containing the mutation D10A/H840A) SpCas9 containing p15A ori and co-expressing a chloramphenicol resistance gene were transformed into NEB Stable Competent cells. (The SpCas9 expressing competent cells were prepared using $CaCl_2$ method.) The target plasmids bearing an ampicillin resistance gene, the pUC ori and encoding the sgRNAs under a fdVIII promoter, also contain the corresponding target sequence with a CGG PAM. To perform the assay, the plasmids with ampicillin resistance were transformed into both the active and the inactive SpCas9-expressing competent bacteria, at a concentration of 1 ng plasmid per 20 μl competent cells. After transformation, bacteria were plated onto Luria-Bertani (LB)-agar plates containing ampicillin and chloramphenicol antibiotics at a final concentration of 50 μg/ml and 25 μg/ml, respectively. To observe the differences in the competence of the inactive and the active SpCas9-expressing cells, a control ampicillin resistance gene-bearing plasmid, which does not encode a target or sgRNA expression cassette, was also transformed into both cells. The plates were incubated at 37°C overnight and the colonies were counted after. Cleavage efficiency value was calculated by the following equation:

$$Cleavage\ efficiency = \left(1 - \frac{\mathrm{ca}}{\mathrm{db}}\right)$$

where *a* and *b*: average CFU of triplicates (*a*1, *a*2, *a*3, and *b*1, *b*2, *b*3, respectively) obtained by transforming the control ampicillin resistance plasmid into the bacteria expressing inactive and active SpCas9, respectively.

c and *d*: CFUs were obtained by transforming the plasmid under investigation into bacteria expressing inactive and active SpCas9, respectively. The experiment was performed three times and the average cleavage efficiencies were calculated.

If the calculated value of the cleavage efficiency was negative, it is shown as zero in the figure. For raw CFU data of each sgRNA sequence see Supplementary Table S3.

Figure 1C and Supplementary Figures S1–S3 show all three data points of the triplicates. To give equal weight to all spacer sequences, where more than three measurements were made, the three closest ones to the average are depicted (construct numbers: 1484; 10021; 10363; 10365; 10374; 10376; 10535; 10536; 10867; 10883).

Protein expression and purification

Active SpCas9 or inactive (D10A/H840A mutated) SpCas9 expression constructs were transformed into *E. coli* BL21 Rosetta 2 (DE3) cells, and a colony was grown in 50 ml Luria-Bertani (LB) medium at 37°C for 16 h. 1 l of growth media (12 g/l Tripton, 24 g/l Yeast, 10 g/l NaCl, 883 mg/l NaH_2PO_4 H_2O, 4.77 g/l Na_2HPO_4, pH 7.5) was inoculated with 10 ml from this culture (1:1000) and cells were grown at 37°C up to $OD_{600} = 0.6$, then they were chilled to 18°C, and after induction with 0.2 mM IPTG, proteins were expressed at 18°C for 16 h. The bacterial cells were centrifuged at 6000 rcf for 15 min at 4°C. The cells were resuspended in 30 ml of Lysis Buffer (40 mM Tris pH 8.0, 500 mM NaCl, 20 mM imidazole, 1 mM TCEP) supplemented with Protease Inhibitor Cocktail (1 tablet/30 ml; complete, EDTA-free, Roche) and were then sonicated on ice. Lysate was cleared by centrifugation at 48000 rcf for 40 min at 4°C. The chromatographic steps were conducted using NGC Scout Medium-Pressure Chromatography Systems (Bio-Rad). Clarified lysate was bound to a 5 ml Mini Nuvia IMAC Ni-Charged column (Bio-Rad). The resin was washed extensively with a solution of 40 mM Tris pH 8.0, 500 mM NaCl and 20 mM imidazole, and the bound protein was eluted by a solution of 40 mM Tris pH 8.0, 250 mM imidazole, 150 mM NaCl and 1 mM TCEP. 10% glycerol was added to the eluted sample and the His6x-MBP fusion protein was cleaved by TEV protease (3 h at 25°C) except for the inactive SpCas9, as it did not have an MBP-tag or a TEV site. The volume of the protein solution was supplemented with buffer (20 mM HEPES pH 7.5, 100 mM KCl, 1 mM DTT) to make up 100 ml. The cleaved protein was purified on a 5 ml HiTrap SP HP cation exchange column (GE Healthcare) and eluted with 1 M KCl, 20 mM HEPES pH 7.5 and 1 mM DTT. The protein was further purified by size exclusion chromatography on a Superdex 200 10/300 GL column (GE Healthcare) in 20 mM HEPES pH 7.5, 200 mM KCl, 1 mM DTT and 10% glycerol. The eluted protein was confirmed by SDS-PAGE and Coomassie brilliant blue R-250 staining. The protein was stored at −20°C.

SgRNA production and purification

SgRNAs were *in vitro* transcribed using TranscriptAid T7 High Yield Transcription Kit and PCR-generated double-stranded DNA templates carrying a T7 promoter sequence. Primers used for the preparation of the DNA templates are listed in Supplementary Table S1. SgRNAs were purified with the RNA Clean & Concentrator kit and then reannealed (95°C for 5 min, ramp to 4°C at 0.3°C/s). SgRNAs were quality checked using 10% denaturing polyacrylamide gels and ethidium bromide staining.

In vitro assays

The assays employ a 2658 bp long linearised plasmid DNA containing the appropriate target sequences (For detailed information about DNA target sequences and preparation see Supplementary Information-DNA targets for *in vitro* experiments), of which cleavage results in two fragments, 1424 and 1234 bp long. In the target DNA cleavage assay, active SpCas9 nuclease (155 nM final concentration) was first incubated with the sgRNA of interest (500 nM final concentration) in reaction buffer without $MgCl_2$ (20 mM HEPES, 200 mM KCl, 1mM TCEP, 2% glycerine, pH 7.5) at 25°C for 5 min. Then the target DNA (15.5 nM final concentration) was added to the mixture at 37°C. To trigger cleavage reaction, $MgCl_2$ solution (2.5 mM final concentration) was added to the mixture. Reactions were stopped by heating samples to 80°C for 5 min with EDTA solution (30 mM final concentration) at different time points (1 min, 30 min, 3 h). The resulting cleavage products were analysed with Bioanalyzer 2100 machine (using Agilent DNA 7500 kit).

In the target binding assay, inactive (D10A/H840A mutated) SpCas9 nuclease (465 nM final concentration) was first incubated with the sgRNA of interest (900 nM final concentration) in reaction buffer at 25°C for 10 min. Then target DNA (15.5 nM final concentration) was added into the mixture at 37°C for 15 min for allowing DNA binding, and then the preincubated active SpCas9-sgRNA (with rescue scaffold) complex was added, and the mixture was incubated for 30 min. Reactions were stopped and analysed. In the loading assay, active SpCas9 nuclease (350 nM final concentration) was first incubated with the sgRNA of interest (non-targeting the target site; 525 nM final concentration) in reaction buffer at 25°C for 10 min. Then the target DNA (18.75 nM final concentration) with the targeting sgRNA were added into the mixture at 37°C. Reactions were stopped at different time points (15, 60, 180 sec) and analysed.

EMSA (electrophoretic mobility shift assay)

Target binding assays were performed in reaction buffer (100 ng/μl heparin, 20 mM HEPES, 200 mM KCl, 1mM TCEP, 2% glycerine, 2.5 mM $MgCl_2$, pH 7.5) in a total volume of 20 μl. Inactive (D10A/H840A mutated) SpCas9 nuclease (1052 nM final concentration) was first incubated with the sgRNA of interest (three times the molar amount of protein) in reaction buffer at 25°C for 5 min. Then target DNA (33 nM final concentration) was incubated with the resulting dSpCas9–sgRNA complex (1052 nM final; or in the case of the serial dilution allow DNA binding. Samples were experiments with 1052; 526; 263 and 131,5 nM protein–sgRNA complex concentrations, respectively) at 37°C for 30 min to run on an 8% native polyacrylamide gel containing 0.5× TBE at 4°C. The gel was stained with ethidium bromide. (For detailed information about DNA target sequences and preparation see Supplementary Information-DNA targets for *in vitro* experiments).

Establishing a set of SL1-related and non-SL1-related motifs

The SL1-related motif set, containing motifs that are not perfectly matching but still complementary to the SL1 region was established by applying the following procedure. First, we generated all possible sequence variations of 4 to 7-nt long sequences, and then, used the RNAcofold program (ViennaRNA Package 2.0) to predict for each sequence the most stable dimer structure that could form with the SL1 unit. From these, we selected the SL1-related sequences by counting the number of nucleotides bound in each of the best dimer structures. To be designated as SL1-related motifs, sequences of 4, 5 or 6–7 nt in length required at least either 3, 4 or 5 bound nucleotides in the dimer structure respectively. We assigned a MotCutEff value [the average of the cutting efficiencies of targets containing the concerned motifs in the Tálas et al. 2021 data set (16)] for each of the SL1-related motifs and further restricted the group by excluding motifs with MotCutEff >= 0.95 or occurring in 10 or less spacers. For SL1-related motifs see Supplementary Table S4. For control, random sequences (rs1- ACACGACCACAC; rs2- GUACUCACCCUC; rs3- CGAACCCUCAAG; rs4- CCCAGCGAAAAC and rs5- GCACCAUGAAGC) with the same length and same GC-content but without sequence similarity to the SL1 segment were selected and used to generate five, non-SL1-related motif sets applying the same procedure described for the SL1-related motifs.

For assessing the contribution of scaffold-complementary motifs to the motif preferences of SpCas9, NucCom sequences were selected with a MotCutEff value lower than 0.66 [as calculated in (16)], and therefore comprising those motifs that SpCas9 preferentially refuse to cleave (influential-motifs), and for the second group, we put all NucCom sequences that had a cutting efficiency higher than 0.98, thus forming the control group with sequences that do not substantially contribute to the motif preferences of SpCas9 (neutral-motifs). The 4-nt long SL1-related motifs were not graphically represented due to their negligible effect on SpCas9 activity.

In silico analysis of the predictability of SpCas9 efficiency with spacers containing scaffold-complementary motifs using deep-learning-based algorithms

Data from the Wang library (22) was used to analyse the predictive power of two deep-learning based efficiency prediction tools, DeepSpCas9 (38) and DeepWT (22), on spacers containing long SL1-complementary motifs. At first, scaffold-complementary sgRNAs were selected into a group ($n = 21$) and for control 1000 reference groups were created by randomly selecting 21 non-scaffold-complementary sgRNAs with matching activity in each control group. Matching the activities was necessary, as we found that the accuracy of the prediction is highly dependent on the activity of the sgRNA; sgRNAs with high activity can be predicted more precisely than sgRNAs with low activity. In addition, spacers with long SL1 motifs tend to have lower than average cleavability. To deal with this problem, sgRNAs were organized into 10 groups based on their activity. When creating the control groups, it was ensured that each group maintained the same activity distribution as the scaffold-complementary sample group by selecting the same number of sgRNAs from each activity group as there was in the sample group. The 8-nt long scaffold motifs are AACGGACT, GGACTAGC, ACGGACTA, CGGACTAG, TAACGGAC, GACTAGCC that were used to generate the scaffold-complementary spacer group. Pearson correlation between measured and predicted cleavage activity of spacers was calculated for each group in case of both DeepSpCas9 and DeepWT predictions.

Calculation of conformational propensity of sgRNAs and pegRNAs

The percentages of sgRNA species with an intact secondary structural unit among the generated conformational ensemble were calculated by an in-house software, RNAfold-wg, which by exploring RNAsubopt and RNAfold (ViennaRNA Package 2.0 version 2.4.18) generates the conformational ensemble of the sgRNA species with free energy in the lowest energy range ($r = 8$) and count the correct conformations (https://github.com/welkergroup/rnafold-wg) (39–50). For values and constraints see Supplementary Table S5.

The percentages of pegRNA species with an intact secondary structural unit/module among the generated conformational ensemble (r = 8) were calculated by RNAsubfold.pl script. (https://github.com/welkergroup/rnafold-wg). For values and constraints see Source Data Figure 6.

For calculation of the minimal free energy of dimers between the sequences and either SL1 or 5 random 12-mer sequences RNAcofold (ViennaRNA Package 2.0 version 2.4.18) was used (39–41,46,50,51).

Tm calculation of scaffold-complementary motifs

Motif Tm was calculated by Bio.SeqUtils.MeltingTemp modul of Biopython v: 1.79. RNA_NN3: values were used (52,53).

Calculating the effects of the strongest short inhibitory motifs (CUU and GCC)

To examine the inhibitory effects of PAM proximal GCC (at spacer positions 16–18, 17–19 and 18–20) and CUU motifs (at spacer positions 18–20), we calculated their motif CutEffs using spacers from the one million spacer library that contained the given motif and calculated their average CutEff values. To ensure that the resulting values represent the specific effect of these motifs, we excluded spacers with a strong secondary structure (Folderg < −7) within the spacer and ones with longer continuous or interrupted SL1 motifs, that could also cause the inhibitory effect. For detailed information about the algorithms and values see Source Data Supplementary Figure S6.

Cell culturing

Cells employed in this study were HEK293T (ATCC, CRL-3216) and HEK-293.EGFP cells. The HEK-293.EGFP cell line contains a single integrated copy of an EGFP cassette driven by the *Prnp* promoter (35).

Cells were grown at 37°C in a humidified atmosphere of 5% CO_2 in high glucose Dulbecco's Modified Eagle medium (DMEM) supplemented with 10% heat inactivated fetal bovine serum, 4 mM L-glutamine (Gibco), 100 units/ml penicillin and 100 μg/ml streptomycin. Cells were passaged up to 20 times (washed with PBS, detached from the plate with 0.05% Trypsin-EDTA and replated). After 20 passages, cells were discarded. Cell lines were not authenticated as they were obtained directly from a certified repository or cloned from those cell lines. Cells were tested for mycoplasma contamination.

PEAR assay

The PEAR reporter was used as described previously (37). Briefly, HEK293T cells were plated on 48-well plates one day before transfection, at a density of 5×10^4 cells/well. Cells were co-transfected with 308 ng of PE2 coding plasmid, 49 ng of sgRNA-mCherry plasmid, 55 ng of PEAR-GFP target plasmid, 153 ng of pegRNA-mCherry plasmid (WT pegRNA-mCherry or rescue mutant pegRNA-mCherry) using 1 μl Turbofect reagent diluted in 50 μl serum-free DMEM. The mixture was incubated at RT for 30 min before adding it to the cells. All transfections were performed in triplicates. Cells were analysed by flow cytometry on the third day after transfection.

The number of GFP positive cells within the mCherry positive population was measured.

GFP disruption assay

HEK-293.EGFP cells were plated on 48-well plates one day before transfection, at a density of 3×10^4 cells/well. Cells were co-transfected with 175 ng of SpCas9 plasmid and 75 ng of pegRNA-mCherry plasmid (WT pegRNA or rescue mutant pegRNA coding plasmid) using with 1 μl Turbofect reagent diluted in 50 μl serum-free DMEM. The mixture was incubated at RT for 30 min before adding it to the cells. All transfections were performed in triplicates. Transfected cells were analysed by flow cytometry on the third and on the seventh day after transfection.

Transfection efficacy was calculated via the amount of mCherry-expressing cells measured on the third day post-transfection.

GFP disruption was calculated based on the EGFP loss on day 7 normalized to the average transfection efficacy. Background EGFP loss was determined using co-transfection of dead SpCas9 expression plasmid with one of the pegRNA coding plasmids (10569; 10570). EGFP disruption values were calculated as follows: the average EGFP background loss from dead SpCas9 control transfections was subtracted from the value of each individual sample in that experiment. The mean values and the standard deviation (SD) were calculated from three parallel transfections.

Flow cytometry

Attune NxT Acoustic Focusing Cytometer (Applied Biosystems by Life Technologies) was used for flow cytometry analysis. In all experiments, a minimum of 2000–10 000 viable single cells were acquired by gating based on the side and forward light-scatter parameters. The GFP and mCherry signal was detected using 488 and 561 nm diode laser for excitation, and 530/30 nm and 620/15 nm filter for emission, respectively. For data analysis Attune Cytometric Software v.4.2 was used.

Statistical analysis

Pearson correlation between measured and predicted cleavage activity of spacers was calculated using scipy.stats SciPy v1.8.1. Differences between WT and rescue pegRNA samples were tested by unpaired *t*-test with Holm-Sidac method. Data normality was tested by Shapiro–Wilk normality test. P values < 0.05 were considered statistically significant. Linear regression, data visualization and Pearson correlation between cleavage activity of spacers and motif length/Tm were performed using GraphPad Prism 9.1.2.

RESULTS

RAR- and SL1-complementary spacer motifs inhibit SpCas9 activity

We examined the effect of complementarity between the spacer and the scaffold sequence (Figure 1A) by

designing 88 spacer sequences that are complementary to different parts of the scaffold, using complementary motifs of six to twenty nucleotides in length (Supplementary Table S2). These spacer sequences were investigated in a bacterial survival assay, where the target plasmid contained not just the target sequence, but also the expression cassette of the corresponding sgRNA along with an antibiotic resistance gene. The target plasmid was transformed into competent bacteria harbouring a SpCas9 expressing plasmid. Therefore, if the sgRNA is functional, SpCas9 cleaves the target plasmid causing the bacteria to be unable to survive on solid media plates with appropriate antibiotics. As expected, bacteria survived when the SpCas9 was inactive or in the absence of a sgRNA and a target. The number of surviving bacterial clones were counted in relation to colonies on control plates where the SpCas9 was inactive and/or the target plasmid did not contain a targeting sgRNA (Figure 1B). To clarify whether survival was indeed due to the inhibition of SpCas9 cleavage by the intended spacer-scaffold complementarity, we designed mutations that could rescue the activity of the nuclease by disrupting the predicted spacer-scaffold structure. To avoid the uncertainty that is associated with altering the sequence of the spacer, we disrupted these dsRNA segments by introducing mutations to the corresponding parts of the scaffold (Supplementary Figure S1a and Supplementary Table S2). In control experiments, the rescue mutations of the scaffold were also examined by using an active spacer sequence, to ensure that the mutations did not interfere with the activity of SpCas9. Only scaffold variants that demonstrated uncompromised activity in this bacterial system were used for rescuing nuclease activity (Supplementary Figure S1a). Furthermore, to understand the extent of the disruption caused by the complementary motifs to the secondary structure of each of the four structural units of the sgRNA, *in silico* we determined the percentages of the species with an intact structural unit in the conformational ensemble of the sgRNAs (Figure 1B right panels and Supplementary Table S5).

Figure 1C (and Supplementary Figure S1) shows the results of the bacterial survival assays. 33 out of the examined 88 self-complementary spacer sequences showed little or no activity (average cleavage efficiency [ACE] <40%), 4 spacer sequences showed activity reduced to intermediate (40% < ACE < 80%) and the rest seemed unaffected (ACE > 80%). Out of the 37 spacers with altered activity, 36 could be rescued by using scaffold-variant sgRNAs (Supplementary Figure S1). We did not observe a relation between the activity of the sgRNAs and the length (Supplementary Figure S2a) or the Tm (Supplementary Figure S2b) of the dsRNA segment that could form between the spacer motif and the scaffold. However, we realised that most motifs that decreased the activity of SpCas9 were designed to be complementary to either the RAR or the SL1 scaffold units. To determine whether the motifs indeed disrupt the secondary structure they were designed to do so, we predicted the secondary structures of the sgRNAs. The sgRNAs were clustered and plotted according to which structural unit was most disrupted by spacer-scaffold interactions in the predictions. Figure 1C reveals that primarily the disruption of the RAR and the SL1 structural units lead to diminished SpCas9 activity. Interestingly, a much stronger predicted disruption was necessary for SL1-complementary motifs to inhibit activity than for RAR-complementary motifs. Several motifs spanned through two structural units of the scaffold, and in order to clarify the sensitivity of the individual units to scaffold-complementary motifs we excluded these motifs in Supplementary Figure S2c and d. These figures show that the activity of SpCas9 is fully protected against the effects of motifs that are complementary to SL2 or SL3. The minimum length of inhibitory motifs was much smaller for SL1 (>7-nt) than for RAR (>11-nt), while the minimal Tm that is necessary for cleavage inhibition was slightly higher for SL1 (>40°C) than for RAR (>37°C). SL1 is the shortest stem, and its secondary structure can be disrupted by shorter motifs *in silico*, however, the Tm for motifs to become disruptive *in silico* were found to be similar for RAR and SL1 (Supplementary Figure S2e, f). Altogether, these experiments revealed that only motifs that are complementary to the RAR and SL1 structural units of the scaffold inhibit SpCas9 activity.

Extension of the RAR, but not the SL1 stem, increases protection against the inhibitory effects of scaffold-complementary spacer motifs

Recently, an extended RAR stem has been reported to increase the stability of the sgRNA and to ensure higher SpCas9 activity (Supplementary Figure S3a) (54). We examined whether the longer RAR stem may be more stable against scaffold-complementary motifs by testing 10 spacer sequences containing motifs that are complementary to the extended RAR sequence. While only motifs with a T_m of at least 37°C complementary to the WT RAR sequence inhibited SpCas9 activity, in case of the extended RAR-complementary motifs, Tm of at least 60°C was necessary for inhibition (Supplementary Figure S3b, c). We applied a similar strategy to construct sgRNAs with an SL1 stem-modified scaffold to increase its stability. When designing the mutations, nucleotides with a known ability to diminish SpCas9 activity when modified were not altered (unpublished results K.H and Briner et al. 2014) (55). Eight SL1-modified variants were generated and each of them preserved its activity (Supplementary Figure S3d). We designed 14 spacers harbouring a motif complementary to the mutant scaffold containing the longest SL1 stem. Interestingly almost all of them diminished SpCas9 activity even with very low T_m (~17°C) (Supplementary Figure S3e, f), whilst motifs complementary to the WT SL1 stem inhibited SpCas9 activity just above a Tm of 40°C. This suggests a different inhibitory mechanism for motifs complementary to RAR or SL1.

RAR- and SL1-complementary spacer motifs interfere with sgRNA loading and target binding, respectively

In vitro DNA-cleavage assays were used to determine which step of the SpCas9 cleavage process, the loading, the target binding or the cleavage is inhibited by the complementary motifs of SL1 and RAR. For both stems, we chose a stronger and a weaker inhibitory motif, so that they have a similar disruptive effect on each stem *in silico*. The SL1-complementary motifs did not interfere with the loading

of the sgRNA into the SpCas9 protein, but they diminished DNA binding, and thus cleavage, in line with previous reports (10,30). In contrast, the RAR-complementary motifs already interfered with the sgRNA loading step, indicating distinct mechanisms of inhibition for RAR- and SL1-complementary motifs (Figure 2). EMSA experiments confirmed that scaffold-complementary spacer motifs in the sgRNA interfere with the target DNA binding of SpCas9 (Supplementary Figure S4).

A deeper understanding of the impact of scaffold-complementary motifs on SpCas9 activity in general requires analysis of many more spacer sequences. Several sequences vs. activity data collection are available that contain by chance scaffold-complementary spacer sequences as well. We wanted to include the longest possible motifs, the random occurrence of which is relatively rare. Therefore, for the analysis we used the largest sequence library currently available, which we have recently published, containing one million target sequences and which is by far the largest sequence vs. activity data collection available (16). The cleavage efficiency (CutEff value) of the SpCas9–sgRNA complex on a target falls in the range of zero to one, with the average CutEff value of the whole library is 0.98. This means that on average, approximately two out of a hundred spacer sequences are inactive. To analyse these data, first, we generated all possible motif sequences that are complementary to the scaffold sequence of the sgRNA which are between 3 and 13 nucleotides in length. Next, for each motif sequence we computed the average of the CutEff values of the spacers within the one million sequence library that contained the given motif (Supplementary Table S6). When the presence of a motif within the spacer substantially inhibited SpCas9 activity, the average CutEff value of the spacers containing this motif was proportionally smaller. Here, we used this average CutEff value of the motifs (hereafter referred to as MotCutEff) to characterise their impact on the activity of SpCas9. The number of spacer sequences containing each motif in any spacer position is given in Supplementary Tables S6 and S7. The occurrence of 3-nt long motifs ranged from 96464 to 388136, however, for 9-nt motifs it ranged between only 6 and 143. We only investigated motifs, that appeared in at least 30 spacer sequences. In Figure 3 (and Supplementary Figure S5), the MotCutEff of all motifs of up to 8 nt in length are shown demonstrating that, in practice, only SL1-complementary motifs inhibited SpCas9 activity substantially. Most inhibitory motifs fit the region between positions 51 and 65 of the scaffold, where the repressive effects of 5 to 8-nt long motifs could be observed. A slight decrease of the MotCutEffs were also apparent at positions 28–33 for motifs of at least 5 nt that were complementary to the RAR unit, and at positions 81–88 for motifs complementary to the SL2/3 units. Interestingly, the inhibitory motifs at the latter two positions showed high degree of similarity to SL1 motif sequences (Supplementary Figure S5a, b). These results confirm our findings from the bacterial survival experiments, that motifs complementary to the non-SL1 sections of the scaffold only have a significant inhibitory effect when they are longer, at least 12 nt in length, hence the reason why they are not apparent here. The random occurrence of such long motifs is so rare, that these motif preferences have no general relevance or practical consequences for the use of SpCas9 in genome editing.

SL1-complementary motifs cause the sequence preference of SpCas9

Furthermore, we examined whether SL1-complementary motifs could explain the spacer position-dependent motif preferences of SpCas9 reported previously with 1- to 7-nt motifs (16). In such case, the inhibitory scaffold-complementary motifs should show substantial spacer position-dependence. Supplementary Figure S5 shows the effect of all spacer position-dependent scaffold-complementary motifs of 3–7 nt in length. Many of the inhibitory motifs identified in Figure 3 showed considerably different effects according to their position within the spacer. Several motifs, especially ones complementary to the 5′ end of the SL1 stem had a substantially stronger effect, when the motif was located at the 3′ region of the spacer (Figure 4A). This effect could be accounted for by multiple factors. (i) These 3′ sequences of the spacer are involved in the initial interactions responsible for starting the R-loop formation, thus their accessibility may be more critical for DNA binding, than that of the 5′ spacer region (21). (ii) Within the SpCas9-sgRNA complex, the 3′ spacer section is located near the SL1 scaffold unit, while interactions with the 5′ spacer part may be spatially more hindered. (iii) Motifs located in the 3′ spacer positions that are complementary to the 5′ part of the SL1 unit create an extension to the RAR stem, which has been suggested to strengthen these spacer-scaffold interactions (10,21).

Considering the 3-nt long motif preferences of SpCas9, the motif with the strongest inhibitory effect that has been apparent in both mammalian and bacterial experiments is the GCC motif at positions 17–19 and 18–20 of the spacer (16,30) and with a weaker effect at positions 16–18 (16,22). Motifs ending with the GCCUU sequence are the ones that extend the RAR stem when hybridising to the SL1, and these motifs showed the strongest position-dependent effect amongst the SL1-complementary motifs (Supplementary Figure S5). Since spacers with GCC and GCCU motifs at their 3′ end can also extend the RAR stem by forming a 1 or 2-nt long bulge at the junctions, it seems to be plausible that the observed inhibitory effect of the GCC sequence near the 3′ end of the spacer originates from spacer-scaffold interactions. To confirm this interpretation, using the bacterial survival assay, we examined 8 spacer sequences containing a 3′ GCC in positions 17–19 or 18–20 using either wild type (WT) or scaffold-mutant sgRNAs with altered SL1 sequence which disrupts the complementarity with the GCC sequences. Data in Figure 4B revealed that 7 out of the 8 spacer sequences showed compromised activity, which were all rescued by scaffold mutations disrupting complementarity. These results verify that the presence of the 3′ GCC motifs at the PAM-proximal positions 18–20 and 17–19 contributes to the diminished activity of these spacers and revealed that the observed motif preference of SpCas9 against the 3′ GCC sequences is due to the interactions within the sgRNAs between the SL1-complementary spacer motif and the scaffold.

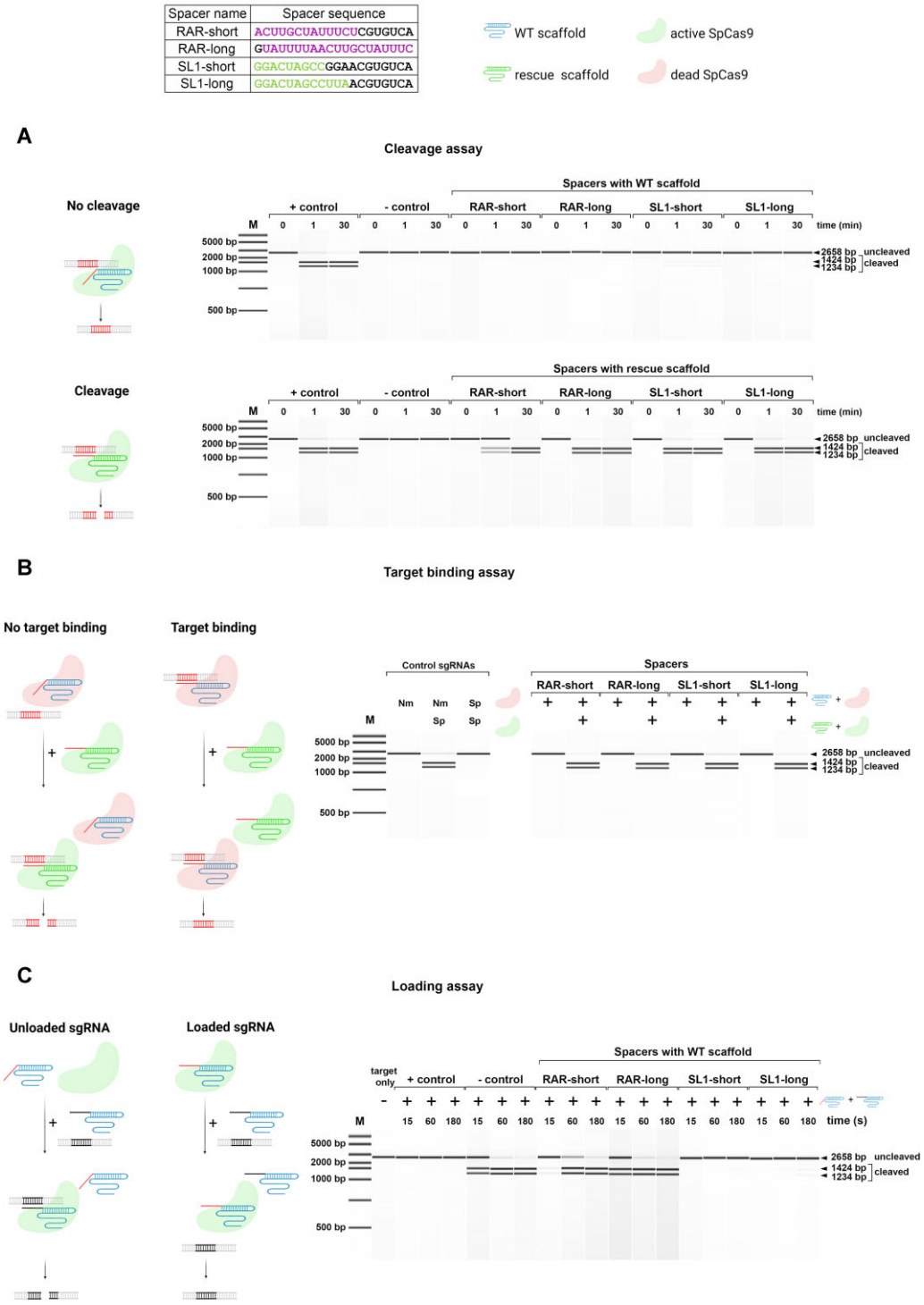


Figure 2. Motifs complementary to RAR or SL1 inhibit different steps of the cleavage process. (A–C) Plasmid-cleaving *in vitro* assays of SpCas9 with four spacers containing either RAR- or SL1-complementary motifs, for which RNAfold-wg predictions suggested similar levels of disruption effect to the concerned stem. Schematic diagrams depict the principle of the cleavage, binding and loading assays. Next to them, a corresponding representative electrophoretogram of three parallel experiments are shown. On the schematic figure SpCas9 is faint green, dead SpCas9 is faint red, inactive scaffold-complementary spacers and corresponding targets are red, active control spacer and target are black, wild type (WT) scaffold is blue and scaffolds with rescue mutations are green. (**A**) *In vitro* cleavage assay. Linearised plasmids are cleaved by SpCas9 harbouring a sgRNA with either WT or rescue scaffold, as indicated. Cleavage products are shown at different time points (0, 1 and 30 min). An active sgRNA and an NmCas9 sgRNA are shown as positive and negative controls, respectively. (**B**) Target binding assay. Dead SpCas9 sgRNA complexes (with spacers as indicated and WT scaffold) incubated with target DNA. Successful target binding protects the target from cleavage by the SpCas9 complex harbouring the same spacer and the corresponding rescue scaffold. An active sgRNA and an NmCas9 sgRNA are shown as positive and negative controls, respectively. (**C**) Loading assay. SpCas9 is incubated with sgRNAs as indicated. Successful loading blocks the loading of an active control sgRNA, and thus, the cleavage of its target. An active, non-targeting sgRNA and an NmCas9 sgRNA are shown as positive and negative controls, respectively.

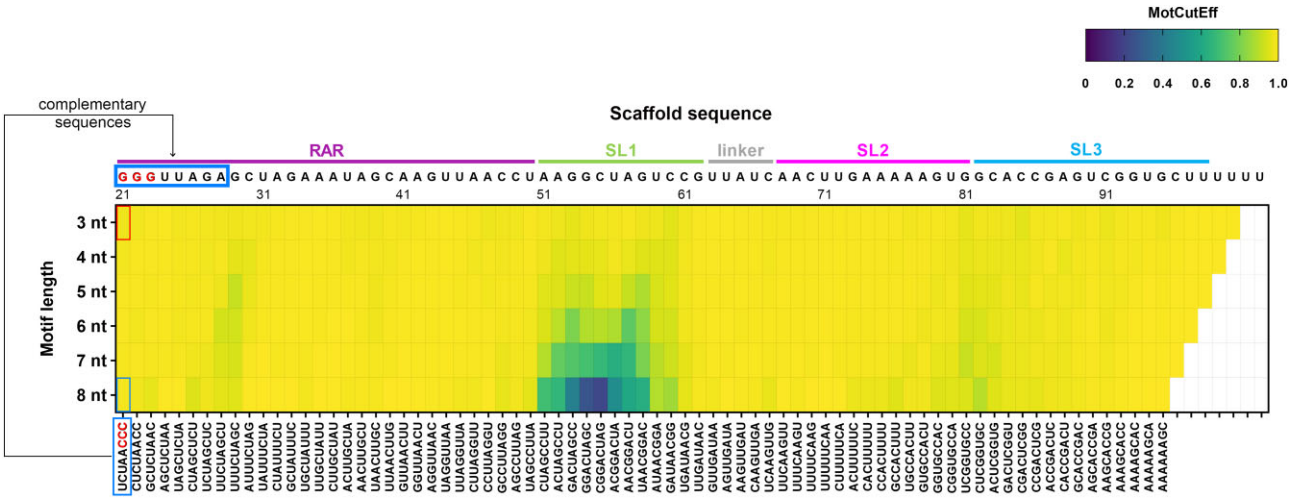


Figure 3. SL1-complementary motifs efficiently inhibit the activity of SpCas9. Heatmap shows the motif cutting efficiency (MotCutEff) values, i.e. the average of the cutting efficiency of spacers containing the given motif in the one million spacer library of Tálas et al. (16) for all 3 to 8-nt long scaffold-complementary motifs (MotCutEff values are calculated from at least 30 spacers). The MotCutEff value of a motif without inhibitory effect approach 0.98, which is the overall average of the cutting efficiency of all spacers in the library. The scaffold sequence used in the study of Tálas et al. (16) is shown above the heatmap, while the 8-nt long scaffold-complementary motifs are shown below. The first 3-nt long motif is coloured red and the 3- and 8-nt long motifs are framed in red and blue, respectively, alongside the corresponding scaffold sequence. Rectangles on the heatmap correspond to individual motifs and are placed at the 5′ position of their complementary scaffold sequences.

If it is solely the perfectly matching SL1-complementary motifs that inhibit SpCas9 activity, the overall significance of the effect is limited due to the small number of motifs concerned and thus few spacer sequences affected. Nucleic acid stems can form with G:U pairing, mismatches and bulges, although these are known to decrease the stability of secondary structures. To test whether non-perfectly matching SL1-complementary motifs in spacers can also interfere with the nuclease activity, we choose three 6-nt SL1-complementary motifs and systematically generated sequences with all possible substitutions and 1 nt insertions and deletions for each of the three motifs. Figure 4C shows that not only the motifs that are perfectly complementary to the SL1 unit are detrimental to the activity of SpCas9, but many motifs with substitutions, insertions and deletions as well. These greatly increase the number of potential motifs that decrease SpCas9 activity through SL1-complementary interactions.

Next, we examined the hypothesis that the position-dependent motif preferences of SpCas9 are related to scaffold-complementary motifs. We established 5, 6 and 7-nt long motif sets that included the sequences that are either perfectly or not perfectly complementary to the SL1 unit (SL1-related motifs). For control, we generated 5 similar random pools of motifs for each of the SL1-related motif sets containing non-SL1-related sequences as described in the Materials and Methods section. To see the extent to which the presence of scaffold-complementary motifs in the spacer are responsible for the observed motif preferences of SpCas9, we examined whether the motifs with the greatest impact on SpCas9 activity were indeed SL1-complementary. We selected those 5 to 7-nt long position-dependent sequence motifs of SpCas9 (hereafter: influential-motifs) that have the largest impact on SpCas9 activity, identified in the study of Tálas *et al.* (16), and determined their overlapping fraction with the SL1-related motifs. Influential-motifs are the motifs that SpCas9 preferentially refuses to cleave. For control, we selected motifs that do not substantially contribute to the motif preferences of SpCas9 (hereafter: neutral-motifs). Figure 5 shows the percentage of both SL1-related and non-SL1-related motifs within the group of influential-motifs (Figure 5A) or within the group of neutral-motifs (Figure 5B). About 80–89% of the influential-motifs proved to be SL1-related (Figure 5A). By contrast, within the neutral-motif group, the fraction of the SL1-related motifs was only 3–16%, depending on the length of the motif (Figure 5B), which is comparable to the fractions of the five non-SL1-related motif sets within both groups (Figure 5A, B). These data indicate that the spacer position-dependent motif preference of SpCas9 predominantly originates from spacer-scaffold complementary sequence interactions within the sgRNA.

Considering all the above results, it seems unlikely that short motifs alone, rather than as part of longer motifs, are sufficient to inhibit SpCas9 activity. To confirm this, we calculated the average cutting efficiency of spacers containing a PAM-proximal GCC or CUU motif, which are or are not part of a longer SL1-associated motif. Supplementary Figure S6 shows that in the absence of a longer SL1-related sequence even the strongest short inhibitory motifs, CUU and GCC, have little to no discernible impact on the activity of SpCas9. It is worth noting that the spacers in Figure 4B also contained longer than 3-nt SL1 motifs.

The primarily intrinsic factor inhibiting SpCas9 activity is intramolecular interactions within the sgRNAs

We also sought to determine the fraction of the inactive spacer sequences (CutEff $\leq$ 0.3) that were affected by SL1-complementary motifs. To do this, we identified the SL1-related motifs for each spacer sequence that contained one. When the spacer contained more than one

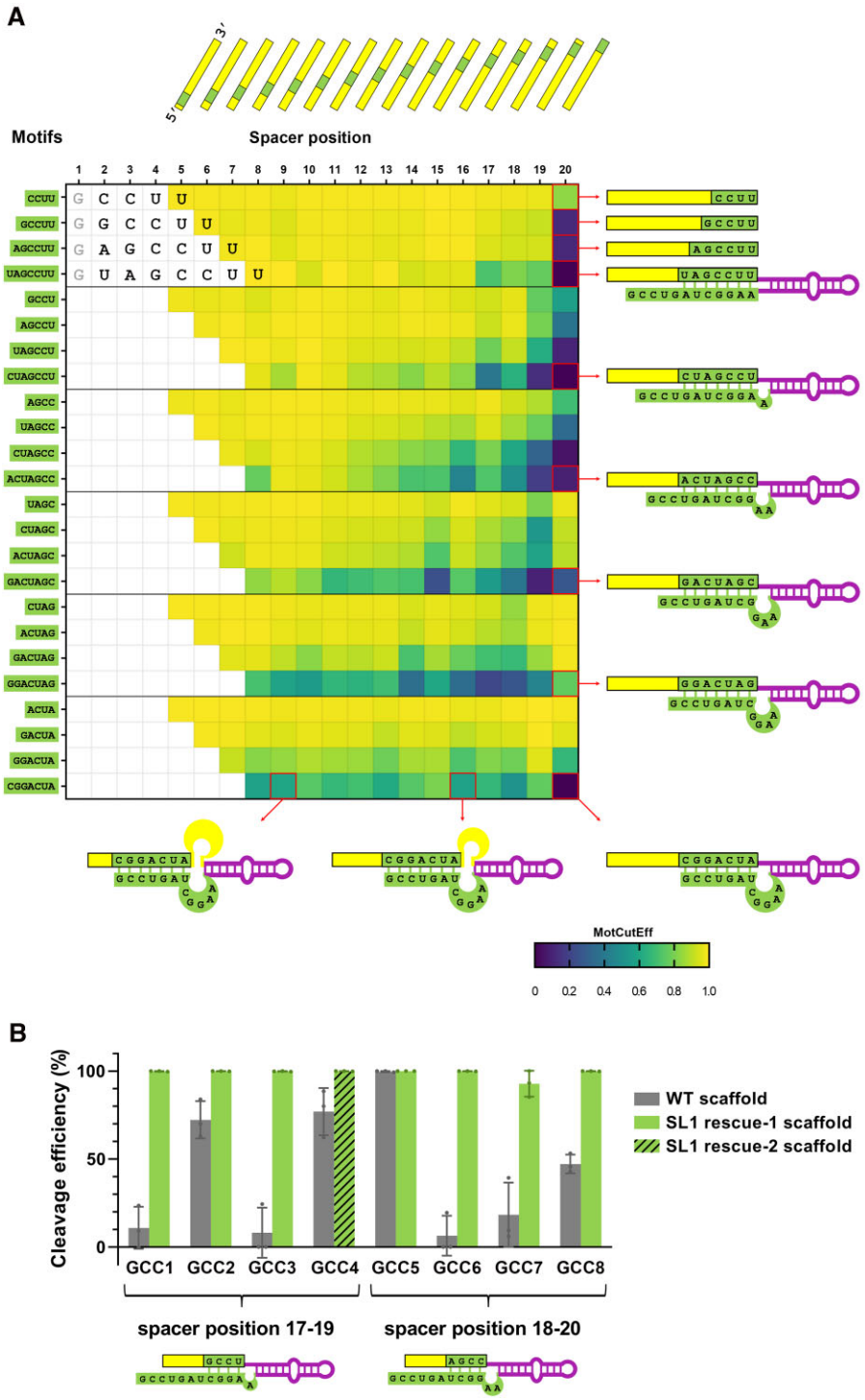


Figure 4. SL1-complementary motifs could explain the spacer position-dependent motif preferences of SpCas9. (**A**) The heatmap shows the MotCutEff values of the 4 to 7-nt long SL1-complementary motifs in different spacer positions. Each motif is represented by a rectangle at its 3′ end position. MotCutEff values were calculated from at least 10 spacers. When motifs are located in the 20th spacer position (the top four red framed motifs on the panel), their hybridisation to SL1 extends the RAR stem without forming a bulge. (**B**) Cleavage efficiency of spacers with 3′ GCC motifs in spacer positions 18–20 and 17–19 in sgRNAs with either the WT or a rescue scaffold in bacterial survival assay. Columns correspond to means ± SD of triplicates. (**C**) Heatmap illustrates the MotCutEff values of SpCas9 with sgRNAs that contain 6-nt long motifs in spacer position 20 perfectly or imperfectly matching with the SL1 sequence. MotCutEff values are calculated from at least 10 spacers. Mismatches are all possible 1 nt substitutions, insertions or deletions for each of the three motifs. Altered nucleotides are indicated by red letters, except for mismatches that result in G:U base pairing (blue letters). When systematic substitutions result in the same motif, each identical motif is indicated for completeness, but the second occurrence is marked in grey.

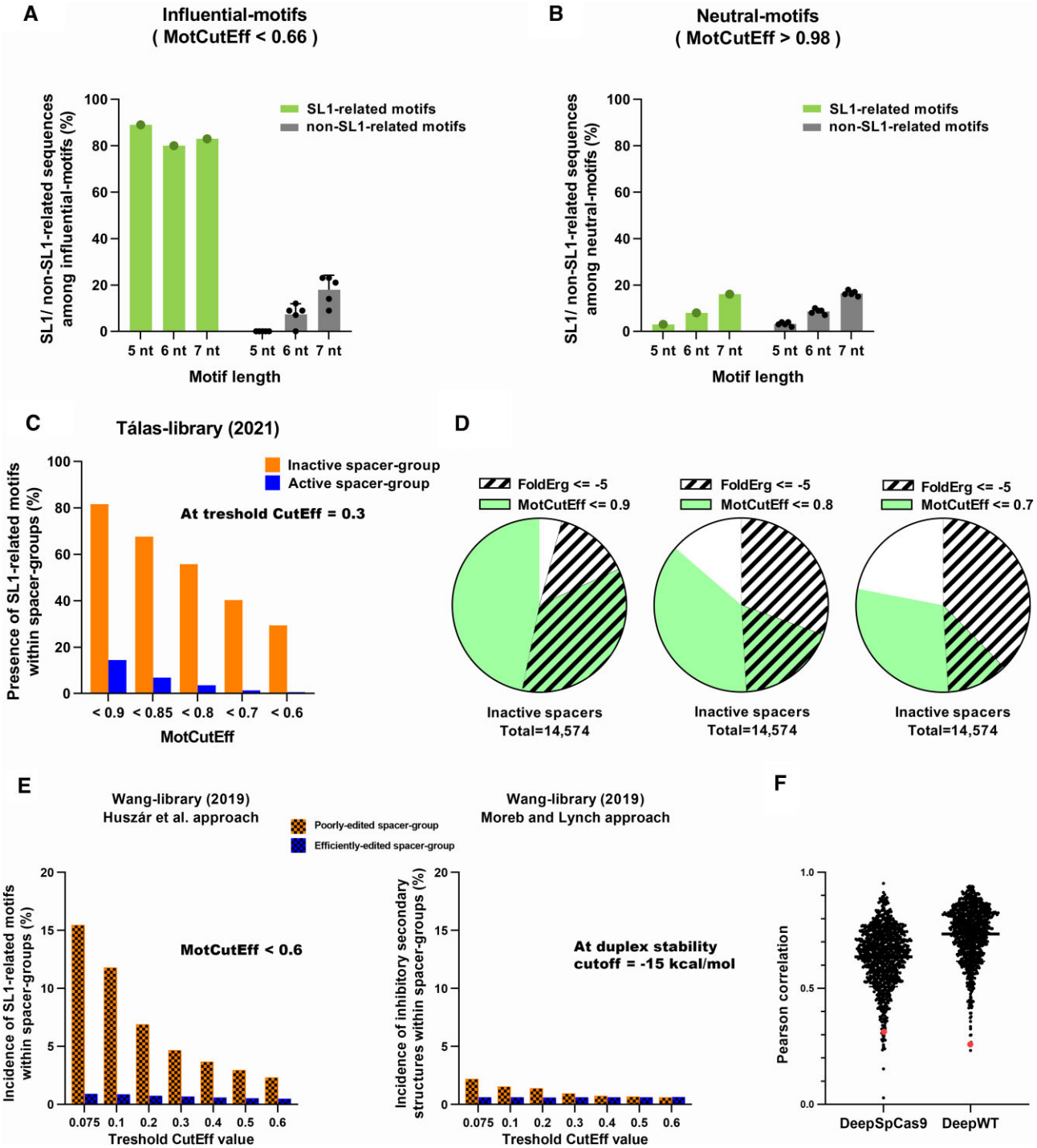


Figure 5. Scaffold-complementary spacer motifs are the ultimate cause of SpCas9 sequence preference and are present in the majority of inactive spacers. (A, B) The plots show the incidence of the SL1- and the non-SL1-related motifs of 5–7 nt in length among either (**A**) influential-motifs or (**B**) neutral-motifs as determined by Tálas *et al.* (16). (**C**) The incidence of SL1-related motifs (with various maximum MotCutEff values) among active (CutEff > 0.3) and inactive (CutEff ≤ 0.3) spacers in the one-million spacers library is shown. (**D**) Pie charts show the fraction of spacers containing self-complementary sequences or SL1-related scaffold-complementary motifs (with various maximum MotCutEff values) among the inactive spacers of the library of Tálas *et al*. The MFE values (FoldErg) were taken from Tálas *et al.* (16). (**E**) The plots show the incidence of SL1-related motifs (with MotCutEff < 0.6) on the left panel and the incidence of inhibitory secondary structures [with a duplex energy < -15 kcal/mol as defined in (57)] on the right panel among poorly-edited and efficiently-edited spacers of the Wang-library (22) at various threshold CutEff values separating the two classes. We chose the 0.6 MotCutEff value on the left panel, that gives comparable incidence of SL1-related motifs (left panel) and inhibitory secondary structures amongst the efficiently-edited spacers (false positive) by both approaches, to make the two predictions more easily comparable. (**F**) Pearson correlation was calculated between experimentally-determined (22) vs predicted [by DeepWT (22) and DeepSpCas9 (38)] sgRNA efficiencies for a group of sgRNAs containing 8-nt long SL1-complementary motifs alongside 1000 randomly generated control groups (see Materials and Methods). The sample groups (red dots) are significantly different from the control groups (Wilcoxon test, $P < 0.0001$ for both predictions).

SL1-related motifs, the most disruptive one (with the lowest MotCutEff value) was selected. Since the effect of the SL1-complementary motifs are not binary (inhibits or not) to get a more comprehensive picture we determined the overlapping fractions between the active/inactive spacers and spacers containing SL1-related motifs with different maximal MotCutEff (between 0.6 and 0.9). Figure 5C shows that up to 14.4% of all the active spacer sequences contained SL1-related motifs. By contrast, this percentage was 81.7% in the case of inactive spacers (at a maximal MotCutEff < 0.9). These data show that scaffolds inhibitory motifs exert their effect in concert with other factors and have a contribution in the vast majority of inactive spacers.

Figure 5D shows that most of the inactive spacer sequences that do not contain SL1-related motifs are self-complementary. Indeed, only 606 inactive sgRNAs (at MotCutEff < 0.9) out of the 14 574 were not affected by self-complementary sequences either between the spacer and the SL1 stem of the scaffold or within the spacer. These results indicate that self-complementary interactions within the sgRNAs are the major internal factors affecting the activity of SpCas9–sgRNA complex.

Next, we looked at whether the effect of SL1-related motifs is also readily recognisable in mammalian data. We examined the largest mammalian SpCas9 activity data set (22) with >50 000 targets, and we divided them into efficiently-edited and poorly-edited classes. The percentage of spacers with SL1-related motifs amongst the target sequences that are efficiently edited by SpCas9 was found to be approximately the same in the mammalian dataset as in the one million sequence library. However, they account for a smaller but still significant fraction of poorly edited spacers. The ratio ranges from about half to one-sixth when using the same set SL1-related motifs depending on the arbitrary threshold between poorly and efficiently edited target sequences (Figure 5E, left panel). These results are consistent with the concept that the inactive spacers constitute a smaller fraction in mammalian cells where many more cellular factors interfere with the activity of SpCas9. These data also suggest that the effect of scaffold-complementary spacer motifs should be considered in the efficiency prediction of SpCas9 activity. Thyme *et al.* (10) have demonstrated the existence of inhibitory scaffold-complementary interactions, and this finding has since been incorporated into the chop-chop website (56), where the instructions state that if a complementary stem longer than four nucleotides is formed between the spacer and the scaffold sequences, it will be scored regardless of which scaffold region is involved. Given our results, this may not provide optimal filtering of inhibitory sequences. We speculate that, due to the relatively small size of the sequence libraries used for training machine learning/neural network algorithms, the activity of SpCas9 is likely to be predicted less accurately on targets that contain longer SL1-complementary motifs. To test this, we chose the two algorithms; DeepSpCas9 (38) and DeepWT (22), that were trained on the largest data sets. Figure 5F shows that they indeed predict such targets significantly less accurately, highlighting the necessity of an algorithm that identifies the hindering effect of target sequences containing SL1-complementary motifs. We built a web tool (Scaffold-Complementary Motifs Finder, SCMF (scmf.welkergroup.hu)) that identifies, when an SL1-related motif is present in spacer sequences and indicates the extent of its effect.

A recent study by Moreb and Lynch (57), published during the preparation of this paper, also analysed the data of Tálas et al. They suggested that some scaffold-complementary motifs decrease the activity of SpCas9, when the duplex stability of the dimer between the spacer and the nucleotides in positions 51–67 of the scaffold, containing the SL1 with the neighbouring unstructured linker region, is less than −15 kcal/mol. They formulated their approach without considering the effect of the position of the motifs within the spacer. Furthermore, they also considered motifs that are complementary to the linker region to be effective, a region for which we found no inhibitory effect (Figure 3, Supplementary Figure S5). Thus, we proposed that they could not capture the magnitude of this effect correctly. Figure 5E shows that it is the case. Our approach identified the effect of scaffold-complementary motifs in spacers, substantially decreasing the activity of SpCas9, about 6-fold more specifically than the Moreb and Lynch approach.

Scaffold-complementary motifs interfere with prime editing activity of SpCas9

One of the most interesting results of our study is that amongst scaffold-complementary motifs, the ones extending the RAR duplex of the sgRNA are the most detrimental. We hypothesized, that such interactions may also occur during prime editing, a recently developed versatile genome modification approach exploiting also SpCas9 (34), and it may be responsible for the low activities of some of the pegRNAs. PegRNAs contain a canonical sgRNA with a 3′ extension, the sequence of which varies from target to target as it is partially determined by the DNA sequence downstream of the target. When the 5′ end of the extension sequence immediately downstream of the canonical sgRNA is complementary to the SL2 stem, their hybridization extends the SL3 stem of the sgRNA (Figure 6A). To test whether this effect diminishes prime editing we exploited PEAR, a fluorescence assay, we recently developed for the monitoring of prime editing activity (37). This assay allows the testing of any user-defined target sequence and pegRNA (Figure 6B). We designed sequences that contained SL2-complementary motifs with increasing length and tested their activity in mammalian cells using pegRNAs with either WT or SL2-altered, rescue scaffold variant. As planned, increasing the motif length increasingly disrupts the SL2 stem of pegRNAs *in silico* with the WT scaffold, but not with the sequence-altered rescue scaffold. (Figure 6C). Accordingly, prime editing activity decreased with the increasing length of the SL2-complementary motifs in the case of the WT, but not the rescue scaffold (Figure 6C). By contrast, the nuclease activity of the prime editor complex was not inhibited by the presence of these SL2-complementary motifs in the reverse transcription template (RTT) section of the pegRNA as shown in a disruption assay (Figure 6C). These results demonstrate that the presence of short scaffold-complementary motifs in the RTT section of the pegRNA can decrease prime editing activity and call

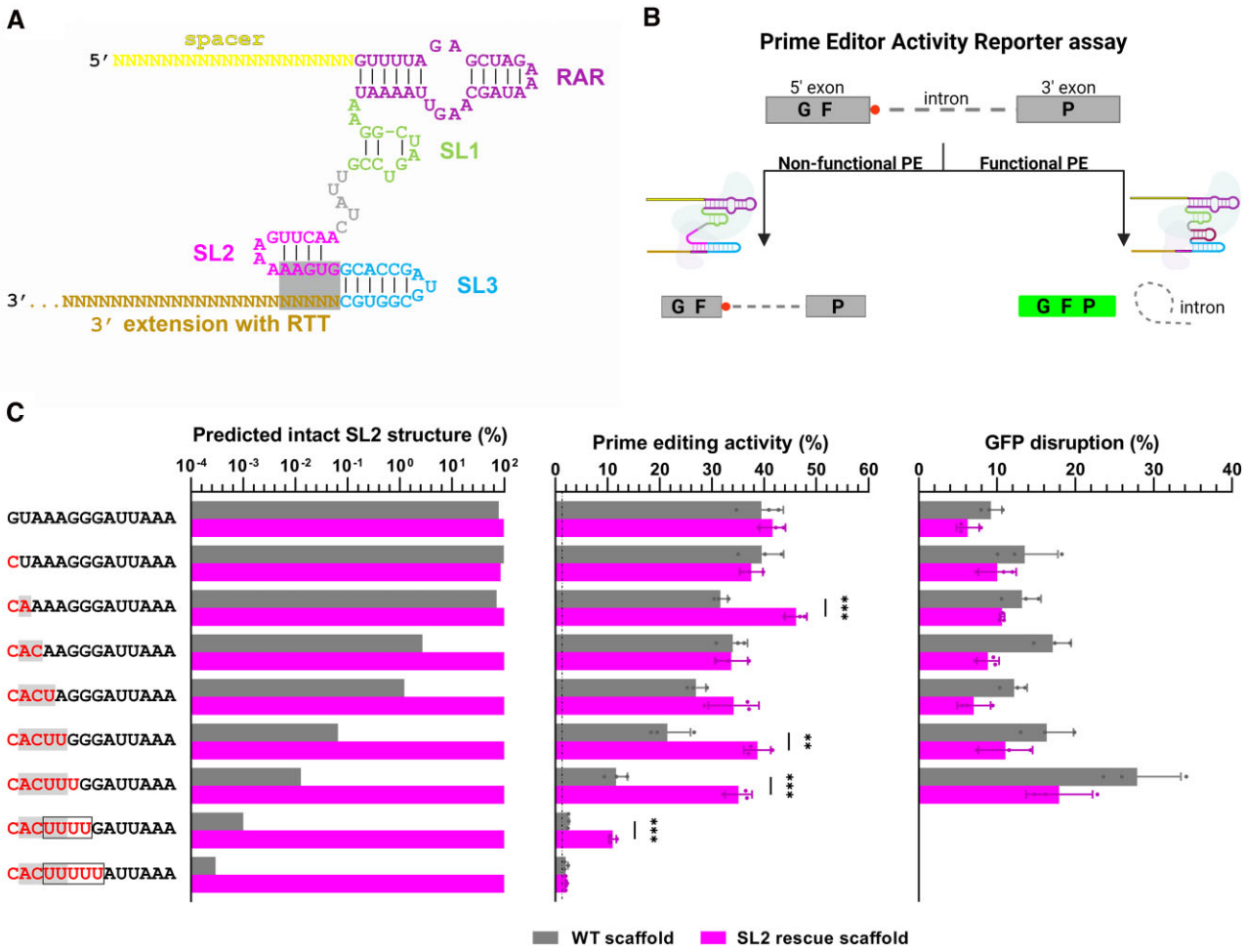


Figure 6. SL2-complementary motifs in the RTT region of pegRNAs diminish prime editing. (**A**) Schematic figure of secondary structure of a pegRNA showing that SL2-complementary motifs could potentially extend the SL3 stem structure. (**B**) Schematic illustration of Prime Editor Activity Reporter (PEAR) assay (37). The mutated and therefore inactive splice site (red dot) of an intron-separated GFP (grey box) can be corrected by a functional prime editor resulting in the restoration of GFP expression (green box). (**C**) Reverse transcriptase template (RTT) sequences of the pegRNA were designed to contain SL2-complementary motifs (red letters) with increasing lengths with either WT (grey) or rescue scaffold (magenta). Grey boxes indicate the positions where rescue scaffold mutations disrupt complementarity. Predicted intact SL2 percentages, prime editing (in a PEAR assay) and SpCas9 nuclease cleavage efficiency (in a GFP disruption assay) are presented for each pegRNA. Columns for the latter two represent the means, error bars represent the SD of triplicates. Differences between the samples of spacers with WT and rescue pegRNAs in a PEAR assay were tested using unpaired t-test. Only statistically significant differences are shown. $^{**}P < 0.01$, $^{***}P < 0.001$.

for more detailed analyses, similar to the one we did above with the scaffold-complementary motifs within the spacers.

DISCUSSION

In summary, our experiments revealed the origin of the spacer position-dependent motif preferences of SpCas9, shed light on the differences in the roles of the individual stems of the sgRNA, unveiled that spacer-scaffold interactions are the predominant factor affecting the intrinsic activity of the SpCas9 RNP complex and highlighted the effect of scaffold-complementary motifs in the RTT section of the pegRNA on the efficiency of prime editing.

It has been difficult to understand the spacer position-dependent motif preferences of SpCas9 for two reasons. First of all, in practice, the protein establishes interactions with the sugar-phosphate backbone of the hybrid RNA-DNA helix in a sequence independent manner and secondly, the motif preferences of SpCas9 of one or two nucleotides vary tremendously amongst studies and data sets (14–17,21–29). Here, we demonstrated how spacer-scaffold interactions give rise to the observed motif preferences of SpCas9. The analysis of the one million sequence library (16) was instrumental in the realization that the motif preferences come mainly from the SL1-complementary sequences. From solely the sequence analysis, however, it would not be possible to distinguish whether SL1-complementary sequences were detrimental to nuclease activity by virtue of their complementarity, or whether the sequence of SL1 just happens to be complementary to sequences with which the SpCas9 RNP has reduced activity. This question could only be answered experimentally, which we did with the bacterial survival assay, using rescue sgRNAs whose scaffold mutations abolish complementarity but retain activity. Spacer-scaffold interactions are inherently unsuitable for direct establishment of one- and

two-nucleotide long sequence preferences; there is a minimum motif length/Tm that is required for this to occur. We have shown how shorter sequence preferences can actually be derived just from longer preferences using the example of the PAM-proximal CUU and GCC motifs (16,18,21,30) (Supplementary Figure S6). In fact, the shortest, one-, two- and three-nucleotide long, preferences (14–17,21–29,31) are likely just the resultant of the longer motif preferences. This also explains the apparent differences in the one- and two-nucleotide long motif preferences of SpCas9 across different data sets (14–17,21–29). The size of the libraries under study (14,16,22,38,58) is nowhere near the possible 4^{20} sequence variations of the spacers. Thus, the presence and distribution of different motifs in the spacers varies depending on the actual sequences of the various spacers present in the data sets. This randomly different distribution of longer motifs can lead to the emergence of the apparent one- and two-nucleotide long preferences that vary between data sets.

The results indicate that the presence of an intact RAR stem is required in the initial step of sgRNA loading, while the integrity of the other stems is not (Figure 2). This is interesting, since both RAR and SL1 stem are essential for sgRNA functioning (2,4–6,55,59–62). With motifs complementary to SL2 and SL3, SpCas9 appears to be functional, despite the apparent capability of the motifs to disrupt these stems. This is consistent with previous observations that the nuclease remains active in the absence of SL2/SL3 stems (5,6,9). Nevertheless, it is not likely that the nuclease is active while the SL2/SL3-spacer interaction is maintained, since if the spacer-scaffold interactions disrupt these stems, that implies that the function of the spacer is also inhibited. Therefore, we speculate that interactions with the protein may promote the formation of these stems in the sgRNA when it is already associated with the SpCas9 via the intact RAR stem. These interactions may stabilize and protect the nascent SL2 and SL3 stems from rebinding the spacer motif, thus releasing the spacer sequence for DNA binding and cleavage. Although the x-ray structures of SpCas9 reveal few protein interactions with the SL2 and SL3 stems, the fact that these stems are necessary for robust activation *in vivo* and that they promote SpCas9–sgRNA active complex formation supports this interpretation (2,6,9,59,62–66). This picture is further nuanced by the inhibition of prime editing by SL2-complementary motifs (Figure 6), suggesting that the SL2 stem is accessible for complementary motifs when they are in the RTT, but not in the spacer. In the case of SL1 the situation seems to be different from RAR and SL2, as the motif sequence in the spacer is able to form a bond with the SL1 sequence in the structure, which is in turn inhibitory to DNA binding. Thus, although the SL1 stem gets disrupted by shorter complementary motifs than the other scaffold stems, potentially due to its smaller size/less stability, this is not the only reason for its higher sensitivity. Motifs complementary to the linker region, which does not have any secondary structure, do not reduce SpCas9 activity. Thus, the higher sensitivity of the SL1 stem, compared to the SL2 and SL3 stems, is presumably due to the easier accessibility of the stem to the spacer in the formed protein-sgRNA complex.

The above view is further supported by the observation that RAR-complementary motifs can inhibit SpCas9 activity with a lower predicted disruption potential than SL1-complementary motifs (Figure 1C), suggesting that protein interactions grant extra stability to the SL1 stem. The extension of the stem reduces the resistance of the SL1 structural unit to the inhibitory effects of SL1-complementary motifs, rather than increasing it. This is probably due to the extended SL1 section not being stabilised by the protein, and/or because it disrupts the existing interactions between the protein and the SL1 unit without reducing the activity of SpCas9. In contrast, extending the stem increases the Tm, which is needed for inhibition by RAR-complementary motifs. This is consistent with the idea that inhibitory motifs may prevent the formation of the RAR stem before its recognition by the protein.

We made a web tool (SCMF) for the identification of spacer sequences that likely have diminished activity due to spacer-scaffold interactions. The identified spacer sequences, however, could still be used in a sgRNA containing a rescue scaffold.

Another important result of this work is the discovery that sequences that are self-complementary within the spacer and/or with the SL1 scaffold unit are present in the vast majority of spacers of inactive SpCas9 RNP complexes (Figure 5C). Interestingly, the proportion of inactive spacers due to spacer-scaffold complementary interactions is significantly lower amongst the poorly-edited mammalian targets, indicating that factors other than spacers inactivity are often the primary reasons for the inability of SpCas9 to cleave efficiently in the more complex mammalian cellular environment (13,62). Nonetheless, filtering out these inactive spacer sequences is an important task, which is only met to a limited extent by programs developed using machine learning as it has small representation in the datasets used to train the deep learning algorithms (Figure 5F) (22,38).

We also find it very interesting that the complementary interactions between SL2 and RTT present in pegRNAs, in contrast to the spacer-SL2 complementary interactions, are inhibitory, even with short matches. This interaction does not appear to affect nuclease activity, but it appears to inhibit RTT sequence function. This result suggests that the SL2 stem is less protected from the effects of complementary motifs in the SpCas9 protein structure when they are located in the RTT, 3′ from the scaffold, rather than in the spacer in the 5′ direction. More comprehensive approaches (67–69) may provide a more complete understanding of how complementary motifs in the RTT influence prime editing activity.

DATA AVAILABILITY

The codes of in-house software, RNAfold-wg and RNAsubfold.pl script are deposited at Github: https://github.com/welkergroup/rnafold-wg and Zenodo, https://doi.org/10.5281/zenodo.7826599.

The Scaffold-Complementary Motifs Finder (SCMF) web tool are available from scmf.welkergroup.hu.

The source data is available for each Figure in the Source Data files.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ildikó Szűcsné Pulinka, Judit Szűcs, Vivien R. Karl, Lilla Burkus, Diána Szeregnyei, Ferencné Zájer, Dávid Fetter, Bernadett Cserkutiné Czene, Luca X. Pájer-Turgyán, Fanni Mráz, Orsolya Oravecz, Gábor Erdős, Viktória Faragó, Veronika Szabó-Csonka, Judit Kálmán, Balázs Bohár for their excellent laboratory assistance, Nóra Weinhardt, Dorottya A. Simon, Antal Nyeste, Elfrieda Fodor, István Vida and Vanessza L. Végi for their valuable help.
Cartoons in Graphical Abstract and in Figures 1b, 2, 6b were created with BioRender.com.

FUNDING

Ministry of National Economy [GINOP-2.1.2-8.1.4-16-2018-00414]; National Research, Development and Innovation Office [K128188, K134968, K142322 to E.W, PD134858 to P.I.K]; János Bolyai Research Scholarship of the Hungarian Academy of Sciences [BO/764/20 to P.I.K]. Funding for open access charge: National Research, Development and Innovation Office [K134968].
Conflict of interest statement. None declared.

REFERENCES

1. Anzalone,A.V., Koblan,L.W. and Liu,D.R. (2020) Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.*, **38**, 824–844.
2. Hsu,P.D., Lander,E.S. and Zhang,F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.
3. Pickar-Oliver,A. and Gersbach,C.A. (2019) The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.*, **20**, 490–507.
4. Jiang,F. and Doudna,J.A. (2017) CRISPR–Cas9 structures and mechanisms. *Annu. Rev. Biophys.*, **46**, 505–529.
5. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
6. Nishimasu,H., Ran,F.A., Hsu,P.D., Konermann,S., Shehata,S.I., Dohmae,N., Ishitani,R., Zhang,F. and Nureki,O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**, 935–949.
7. Anders,C., Niewoehner,O., Duerst,A. and Jinek,M. (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*, **513**, 569–573.
8. Jinek,M., Jiang,F., Taylor,D.W., Sternberg,S.H., Kaya,E., Ma,E., Anders,C., Hauer,M., Zhou,K., Lin,S. *et al.* (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*, **343**, 1247997.
9. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
10. Thyme,S.B., Akhmetova,L., Montague,T.G., Valen,E. and Schier,A.F. (2016) Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.*, **7**, 11750.
11. Orioli,A., Pascali,C., Quartararo,J., Diebel,K.W., Praz,V., Romascano,D., Percudani,R., van Dyk,L.F., Hernandez,N., Teichmann,M. *et al.* (2011) Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. *Nucleic Acids Res.*, **39**, 5499–5512.
12. Nielsen,S., Yuzenkova,Y. and Zenkin,N. (2013) Mechanism of eukaryotic RNA polymerase III transcription termination. *Science*, **340**, 1577–1580.
13. Verkuijl,S.A.N. and Rots,M.G. (2019) The influence of eukaryotic chromatin state on CRISPR–Cas9 editing efficiencies. *Curr. Opin. Biotechnol.*, **55**, 68–73.
14. Doench,J.G., Hartenian,E., Graham,D.B., Tothova,Z., Hegde,M., Smith,I., Sullender,M., Ebert,B.L., Xavier,R.J. and Root,D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.
15. Moreno-Mateos,M.A., Vejnar,C.E., Beaudoin,J.D., Fernandez,J.P., Mis,E.K., Khokha,M.K. and Giraldez,A.J. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods*, **12**, 982–988.
16. Tálas,A., Huszar,K., Kulcsar,P.I., Varga,J.K., Varga,E., Toth,E., Welker,Z., Erdos,G., Pach,P.F., Welker,A. *et al.* (2021) A method for characterizing Cas9 variants via a one-million target sequence library of self-targeting sgRNAs. *Nucleic Acids Res.*, **49**, e31.
17. Xu,H., Xiao,T., Chen,C.H., Li,W., Meyer,C.A., Wu,Q., Wu,D., Cong,L., Zhang,F., Liu,J.S. *et al.* (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Res.*, **25**, 1147–1157.
18. Konstantakos,V., Nentidis,A., Krithara,A. and Paliouras,G. (2022) CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Res.*, **50**, 3616–3637.
19. Cui,Y., Xu,J., Cheng,M., Liao,X. and Peng,S. (2018) Review of CRISPR/Cas9 sgRNA design tools. *Interdiscipl. Sci. Comput. Life Sci.*, **10**, 455–465.
20. Chuai,G.-h., Wang,Q.-L. and Liu,Q. (2017) In silico meets in vivo: towards computational CRISPR-based sgRNA design. *Trends Biotechnol.*, **35**, 12–21.
21. Wong,N., Liu,W. and Wang,X. (2015) WU-CRISPR: characteristics of functional guide rnas for the CRISPR/Cas9 system. *Genome Biol.*, **16**, 218.
22. Wang,D., Zhang,C., Wang,B., Li,B., Wang,Q., Liu,D., Wang,H., Zhou,Y., Shi,L., Lan,F. *et al.* (2019) Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.*, **10**, 4284.
23. Gagnon,J.A., Valen,E., Thyme,S.B., Huang,P., Akhmetova,L., Pauli,A., Montague,T.G., Zimmerman,S., Richter,C. and Schier,A.F. (2014) Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide rnas. *PLoS One*, **9**, e98186.
24. Labuhn,M., Adams,F.F., Ng,M., Knoess,S., Schambach,A., Charpentier,E.M., Schwarzer,A., Mateo,J.L., Klusmann,J.H. and Heckl,D. (2018) Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Res.*, **46**, 1375–1385.
25. Chari,R., Mali,P., Moosburner,M. and Church,G.M. (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods*, **12**, 823–826.
26. Guo,J., Wang,T., Guan,C., Liu,B., Luo,C., Xie,Z., Zhang,C. and Xing,X.H. (2018) Improved sgRNA design in bacteria via genome-wide activity profiling. *Nucleic Acids Res.*, **46**, 7052–7069.
27. Kim,N., Kim,H.K., Lee,S., Seo,J.H., Choi,J.W., Park,J., Min,S., Yoon,S., Cho,S.-R. and Kim,H.H. (2020) Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.*, **38**, 1328–1336.
28. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62–67.
29. Liu,X., Homma,A., Sayadi,J., Yang,S., Ohashi,J. and Takumi,T. (2016) Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system. *Sci. Rep.*, **6**, 19675.
30. Graf,R., Li,X., Chu,V.T. and Rajewsky,K. (2019) sgRNA sequence motifs blocking efficient CRISPR/Cas9-mediated gene editing. *Cell Rep.*, **26**, 1098–1103.
31. Moreb,E.A. and Lynch,M.D. (2021) Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. *Nat. Commun.*, **12**, 5034.
32. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an

RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.
33. Zuris,J.A., Thompson,D.B., Shu,Y., Guilinger,J.P., Bessen,J.L., Hu,J.H., Maeder,M.L., Joung,J.K., Chen,Z.Y. and Liu,D.R. (2015) Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nat. Biotechnol.*, **33**, 73–80.
34. Anzalone,A.V., Randolph,P.B., Davis,J.R., Sousa,A.A., Koblan,L.W., Levy,J.M., Chen,P.J., Wilson,C., Newby,G.A., Raguram,A. *et al.* (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, **576**, 149–157.
35. Kulcsár,P.I., Tálas,A., Huszár,K., Ligeti,Z., Tóth,E., Weinhardt,N., Fodor,E. and Welker,E. (2017) Crossing enhanced and high fidelity SpCas9 nucleases to optimize specificity and cleavage. *Genome Biol.*, **18**, 190.
36. Tálas,A., Simon,D.A., Kulcsár,P.I., Varga,É., Krausz,S.L. and Welker,E. (2021) BEAR reveals that increased fidelity variants can successfully reduce the mismatch tolerance of adenine but not cytosine base editors. *Nat. Commun.*, **12**, 6353.
37. Simon,D.A., Talas,A., Kulcsar,P.I., Biczok,Z., Krausz,S.L., Varady,G. and Welker,E. (2022) PEAR, a flexible fluorescent reporter for the identification and enrichment of successfully prime edited cells. *Elife*, **11**, e69504.
38. Kim,H.K., Kim,Y., Lee,S., Min,S., Bae,J.Y., Choi,J.W., Park,J., Jung,D., Yoon,S. and Kim,H.H. (2019) SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance. *Sci. Adv.*, **5**, eaax9249.
39. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA package 2.0. *Algorith. Mol. Biol.*, **6**, 26.
40. Lorenz,R., Hofacker,I.L. and Stadler,P.F. (2016) RNA folding with hard and soft constraints. *Algorith. Mol. Biol.*, **11**, 8.
41. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem. /Chemical Monthly*, **125**, 167–188.
42. Hofacker,I.L. and Stadler,P.F. (2006) Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*, **22**, 1172–1176.
43. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
44. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
45. Bompfünewerer,A.F., Backofen,R., Bernhart,S.H., Hertel,J., Hofacker,I.L., Stadler,P.F. and Will,S. (2008) Variations on RNA folding and alignment: lessons from Benasque. *J. Math. Biol.*, **56**, 129–144.
46. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 7287–7292.
47. Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
48. Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
49. Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.
50. Turner,D.H. and Mathews,D.H. (2009) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38**, D280–D282.
51. Bernhart,S.H., Tafer,H., Mückstein,U., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Molecular Biology : AMB*, **1**, 3.
52. Chen,J.L., Dishler,A.L., Kennedy,S.D., Yildirim,I., Liu,B., Turner,D.H. and Serra,M.J. (2012) Testing the nearest neighbor model for canonical RNA base pairs: revision of GU parameters. *Biochemistry*, **51**, 3508–3522.
53. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
54. Chen,B., Gilbert,L.A., Cimini,B. A., Schnitzbauer,J., Zhang,W., Li,G.-W., Park,J., Blackburn,E.H., Weissman,J.S., Qi,L.S *et al.* (2013) Dynamic imaging of genomic loci in living Human cells by an optimized CRISPR/cas system. *Cell*, **155**, 1479–1491.
55. Briner,A.E., Donohoue,P.D., Gomaa,A.A., Selle,K., Slorach,E.M., Nye,C.H., Haurwitz,R.E., Beisel,C.L., May,A.P. and Barrangou,R. (2014) Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell*, **56**, 333–339.
56. Labun,K., Montague,T.G., Gagnon,J.A., Thyme,S.B. and Valen,E. (2016) CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.*, **44**, W272–W276.
57. Moreb,E.A. and Lynch,M.D. (2022) A meta-analysis of gRNA library screens enables an improved understanding of the impact of gRNA folding and structural stability on CRISPR-Cas9 activity. *CRISPR J*, **5**, 146–154.
58. Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.
59. Hu,L.F., Li,Y.X., Wang,J.Z., Zhao,Y.T. and Wang,Y. (2023) Controlling CRISPR-Cas9 by guide RNA engineering. *Wiley Interdiscipl. Rev. RNA*, **14**, e1731.
60. Jiang,M., Ye,Y. and Li,J. (2021) Core hairpin structure of SpCas9 sgRNA functions in a sequence- and spatial conformation-dependent manner. *SLAS Technol.*, **26**, 92–102.
61. Jiang,F., Zhou,K., Ma,L., Gressel,S. and Doudna,J.A. (2015) A Cas9–guide RNA complex preorganized for target DNA recognition. *Science*, **348**, 1477–1481.
62. Javaid,N. and Choi,S. (2021) CRISPR/Cas system and factors affecting its precision and efficiency. *Front. Cell Dev. Biol.*, **9**, 761709.
63. Jinek,M., East,A., Cheng,A., Lin,S., Ma,E. and Doudna,J. (2013) RNA-programmed genome editing in human cells. *Elife*, **2**, e00471.
64. Jiang,F., Taylor,D.W., Chen,J.S., Kornfeld,J.E., Zhou,K., Thompson,A.J., Nogales,E. and Doudna,J.A. (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, **351**, 867–871.
65. Mekler,V., Minakhin,L., Semenova,E., Kuznedelov,K. and Severinov,K. (2016) Kinetics of the CRISPR-Cas9 effector complex assembly and the role of 3′-terminal segment of guide RNA. *Nucleic Acids Res.*, **44**, 2837–2845.
66. Wright,A.V., Sternberg,S.H., Taylor,D.W., Staahl,B.T., Bardales,J.A., Kornfeld,J.E. and Doudna,J.A. (2015) Rational design of a split-Cas9 enzyme complex. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 2984–2989.
67. Mathis,N., Allam,A., Kissling,L., Marquart,K.F., Schmidheini,L., Solari,C., Balázs,Z., Krauthammer,M. and Schwank,G. (2023) Predicting prime editing efficiency and product purity by deep learning. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-022-01613-7.
68. Kim,H.K., Yu,G., Park,J., Min,S., Lee,S., Yoon,S. and Kim,H.H. (2021) Predicting the efficiency of prime editing guide rnas in human cells. *Nat. Biotechnol.*, **39**, 198–206.
69. Koeppel,J., Weller,J., Peets,E.M., Pallaseni,A., Kuzmin,I., Raudvere,U., Peterson,H., Liberante,F.G. and Parts,L. (2023) Prediction of prime editing insertion efficiencies using sequence features and DNA repair determinants. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-023-01678-y.